

# **SCHOLARLY PUBLICATIONS**

*A CURRENT AWARENESS BULLETIN  
OF RESEARCH OUTPUT*

**@DTU**

(129<sup>th</sup> Edition)

**SEPTEMBER 2023**

**BY: CENTRAL LIBRARY**

**DELHI TECHNOLOGICAL UNIVERSITY**

(FORMERLY *DELHI COLLEGE OF ENGINEERING*)

GOVT. OF N.C.T. OF DELHI

SHAHBAD DAULATPUR, MAIN BAWANA ROAD

DELHI 110042



# **PREFACE**

This is the **One Hundred Twenty Ninth** Issue of Current Awareness Bulletin started by Delhi Technological University, Central Library. The aim of the bulletin is to compile, preserve and disseminate information published by the faculty, students and alumni for mutual benefits. The bulletin also aims to propagate the intellectual contribution of Delhi Technological University (DTU) as a whole to the academia.

The bulletin contains information resources available in the internet in the form of articles, reports, presentations published in international journals, websites, etc. by the faculty and students of DTU. The publications of faculty and student which are not covered in this bulletin may be because of the reason that the full text either was not accessible or could not be searched by the search engine used by the library for this purpose.

The learned faculty and students are requested to provide their uncovered publications to the library either through email or in CD, etc. to make the bulletin more comprehensive.

This issue contains the information published during **September, 2023**. The arrangement of the contents is alphabetical. The full text of the article which is either subscribed by the university or available in the web is provided in this bulletin.

**Central Library**

# CONTENTS

1. A Dynamic and Efficient Active Power Support Scheme Using VSC-HVDC System for Quick Frequency Restoration, **6.Ashima Taneja**, **3.Radheshyam Saha** and **3.Madhusudan Singh**, Electrical, DTU
2. A High PFC Integrated AC-DC Circuit With Inherent Lossless LCD Snubber, **3.Vinod Kumar Yadav**, Electrical, DTU
3. A multilevel authentication-based blockchain powered medicine anti-counterfeiting for reliable IoT supply chain management, **6.Neetu Sharma** and **3.Rajesh Rohilla**, Electronics, DTU
4. A Single MOS-Memristor Emulator Circuit, **6.Rahul Kumar Gupta**, **3.Mahipal Singh Choudhry**, Varun Saxena and **3.Sachin Taran**, Electronics, DTU
5. A Spectrum of Solutions: Unveiling Non-Pharmacological Approaches to Manage Autism Spectrum Disorder, Arunima Mondal, **8.Rashi Sharma**, Umme Abiha, Faizan Ahmad, Anik Karan, Richard L. Jayaraj and Vaishnavi Sundar, Biotechnology, DTU
6. Analysis of the Properties of Recycled Aggregates Concrete with Lime and Metakaolin, Manvendra Verma, Arti Chouksey, Rahul Kumar Meena and **6.Indrajeet Singh**, Civil, DTU
7. Analyzing a higher order  $q(t)$  model and its implications in the late evolution of the Universe using recent observational datasets, **7.Madhur khurana**, **7.Himanshu Chaudhary**, Saadia Mumtaz, S. K. J. Pacif, and G. Mustafa, Applied Physics and Applied Mathematics, DTU
8. Android Malware Analysis using Coefficient of Multiple Correlation, **8.Shresth Jain**, **8.Sarthak Kapoor**, **3.Anshul Arora** and **8.Sachin Kumar Sharma**, Applied Mathematics, DTU

9. Blood pressure estimation and classification using a reference signal-less photoplethysmography signal: a deep learning framework, Pankaj, Ashish Kumar, Rama Komaragiri and **3.Manjeet Kumar**, Electronics, DTU
10. Capabilities and Features Offered by SDN on the Cloud Network Infrastructure, Khwaja Hedayetulla Sidiqi, Zalmai Zormatai and **6.Samiullah Mehraban**, CSE, DTU
11. CFD Analysis of 2x2 Rod Bundles at Supercritical Flow Condition, **6.Gaurav Kumar** and **3.Raj Kumar Singh**, Mechanical, DTU
12. Characterization and Steady State Analysis of Multiport Switched Boost Converter, **8.Rishabh Bansal**, **8.Vaibhav Tokas**, **8.Rushiv Bansal** and **3.Mayank Kumar**, Electrical, DTU
13. Circular Polarization Division Multiplexing in Visible Light Communication System by Incorporating QPSK and Distortion Compensation Enabled DSP/IDSP, Hunny Pahuja, **3.Monika Verma**, Shippu Sachdeva, Simarpreet Kaur, Manoj Sindhwani and Manoj Kumar Shukla, Electrical, DTU
14. Classification of Toxic Comments Unified Through Diverse Internet Forums, **3.Gull Kaur**, **8.Aakash Kumar**, **8.Aarjav Chauhan** and **8.Abhishek Babbar**, CSE, DTU
15. Collective Behavior based Slime Mold Optimization with application to UAV Energy conversion device analysis, **3.Monika Verma**, **3.Mini Sreejeth**, and **3.Madhusudan Singh**, Electrical, DTU
16. Comparative Study on Forecasting of Schedule Generation in Delhi Region for the Resilient Power Grid Using Machine Learning, Lakshmi D, Ravi Shekhar Tiwari, **6.1.Neelu Nagpal**, Neelam Kassarwani, Vishnuvarthanan G and Abhishek Srivastava, Electrical, DTU
17. Comparative Analysis of ResNet and DenseNet for Differential Cryptanalysis of SPECK 32/64 Lightweight Block Cipher, **8.Ayan Sajwan** and Girish Mishra, Physics, DTU

18. Computational Fluid Dynamics (CFD) Simulations of Domestic Hybrid Solar Dryer Under Varying Mass Flow Rates, Mukul Sharma, Deepali Atheaya, **3.Anil Kumar** and Pawan Mishra, Mechanical, DTU
19. Constant Current Fault-Tolerant Buck Type Interleaved DC-DC Converter for Battery Charging Applications, **7.Abhishek Chawla** and **3.Mayank Kumar**, Electrical, DTU
20. CORRDroid - Android Malware Detection using Association amongst Permissions, **8.Ankita Jain**, **8.Lakshit Rustagi**, **8.Mayank Aggarwal** and **3.Anshul Arora**, Applied Mathematics, DTU
21. Deep Convolutional Neural Network With Attention Module for Seismic Impedance Inversion, Vineela Chandra Dodda , Member, IEEE, Lakshmi Kuruguntla, **3.Anup Kumar Mandpura**, Karthikeyan Elumalai, and Mrinal K. Sen, Electrical, DTU
22. Deep Fake Detection using Transfer Learning, **3.Rahul Thakur**, Amit kumar samanta, Amrit and Daksh Garg, Electronics, DTU
23. Denoising using a Hybrid Filter Comprised of GGIF, WLS, and 2D Bilateral Filtering, **3.Abhilasha Sharma**, **8.Aryaman Dosajh** and **8.Moin Ahmad Chalkoo**, Software Engineering, DTU
24. Design and Analysis of LLC Resonant Converter for Electric Vehicle Battery Charging, **8.Shreyas** and **3.Mayank Kumar**, Electrical, DTU
25. Design of 50 kW Two Stage off-Board EV Charger using CC-CV Algorithm, **7.Kushank Singh** and **3.Vanjari Venkata Ramana**, Electrical, DTU
26. Design of a novel robust recurrent neural network for the identification of complex nonlinear dynamical systems, **6.R. Shobana**, **3.Bhavnesht Jain** and Rajesh Kumar, Electrical, DTU
27. Detection of Sleep Apnea and its Intensity in Adults, **8.Bhupinder Singh Saini**, **8.Chirag Kaushik**, **8.Ayussh Vashishth** and **3.Lavi Tanwar**, Electronics, DTU

28. Determinants of sustainable frugal innovation in higher education: a massive open online courses perspective, **3.Dr. Shikha N. Khera & 6.Himanshu Pawar**, DSM, DTU
29. Device modelling of lead free  $(\text{CH}_3\text{NH}_3)_2\text{CuX}_4$  based perovskite solar cells using SCAPS simulation, **6.Rahul Kundara** and **3.Sarita Baghel**, Applied Physics, DTU
30. DVRGNet: an efficient network for extracting obscenity from multimedia content, **6.Kamakshi Rautela**, **6.Dhruv Sharma**, **6.Vijay Kumar** and **3.Dinesh Kumar**, Electronics, DTU
31. Effect of partial shading on photovoltaic systems performance and its mitigation techniques-a review, **6.Nikhil Kushwaha**, **3.Vinod Kumar Yadav** and **3.Radheshyam Saha**, Electrical, DTU
32. Effects of Surface Modified Recycled Coarse Aggregates on Concrete's Mechanical Characteristics, **6.Harish Panghal** and **3.Awadhesh Kumar**, Civil, DTU
33. End-to-End Historical Handwritten Ethiopic Text Recognition using Deep Learning, **3.Ruchika Malhotra** and **6.Maru Tesfaye Addis**, CSE, DTU
34. Energy Analysis of the Multi-Stage Vapour Compression Refrigeration System Using Eight Low GWP Refrigerants, **6.Manjit Singh & 3.Akhilesh Arora**, Mechanical, DTU
35. Enhancement of Power Quality of Three-Phase GC Solar Photovoltaics, **6.Sukhbir Singh** and **3.J N Rai**, Electrical, DTU
36. Environomical Analysis of Green Building Having Various Window-to-Wall Ratio, Asim Ahmad, **8.Om Prakash**, **8.Pranav Nayan**, Anil Kumar, **3.Bharath Bhushan** and Rajeshwari Chatterjee, Mechanical, DTU
37. Environomical Analysis of Sensible Heat Storage-Based Greenhouse Dryer, Asim Ahmad, Om Prakash, **3.Anil Kumar** & Md Shahnawaz Hussain, Mechanical, DTU
38. Experimental and Numerical Analysis of Residual Stresses in Similar and Dissimilar Welds of T91 and Super304H Steel Tubes, Ranjeet Kumar, Prahlad Halder, Murugaiyan Amrithalingam, **3.N. Yuvraj**, Anand Varma, Y. Ravi Kumar, Suresh Neelakantan and Jayant Jain, Mechanical, DTU

39. Federated learning inspired privacy sensitive emotion recognition based on multi-modal physiological sensors, **6.Neha Gahlan** and **3.Divyashikha Sethia**, Software Engineering, DTU
40. Fin field-effect-transistor engineered sensor for detection of MDA-MB-231 breast cancer cells: A switching-ratio-based sensitivity analysis, **6.Bhavya Kumar** and **3.Rishu Chaujar**, Applied Physics, DTU
41. From methods to datasets: A survey on Image-Caption Generators, **6.Lakshita Agarwal** and **3.Bindu Verma**, IT, DTU
42. Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities, **7.Anjum** and **3.Rahul Katarya**, CSE, DTU
43. Impact of Feature Selection Algorithms on Network Intrusion Detection, **8.Samyak Jain**, **8.Siddharth Bihani**, **8.Satyam Jaiswal** and **3.Anshul Arora**, A Mathematics, DTU
44. Investigation on the impact of elevated temperature on sustainable geopolymer composite, Manvendra Verma, Rahul Kumar Meena, **7.Indrajeet Singh**, Nakul Gupta, Kuldeep K Saxena, M Madhusudhan Reddy, Karrar Hazim Salem and Ummal Salmaan, Civil, DTU
45. Kinetic treatment of lower hybrid waves excitation in a magnetized dusty plasma by electron beam, **6.Anshu**, **3.S C Sharma** and J Sharma, Applied Physics, DTU
46. Liver-Type Fatty Acid Binding Protein (FABP1) Has Exceptional Affinity for Minor Cannabinoids, Dr. Fred Shahbazi, Sanam Mohammadzadeh, Dr. Daniel Meister, Valentyna Tararinaa, **8.Vagisha Aggarwal** & Dr. John F. Trant, A Chemistry, DTU
47. Localization of broken instruments inside the root canal using pulse-echo mode of ultrasonic signal, **8.Rahul Kumar** and **3.Rajiv Kapoor**, ECE, DTU
48. Low-Frequency Waves in a Strongly Correlated Collisional Magnetized Dusty Plasma Cylinder, **6.Harender Mor**, Kavita Rani Segwal and **3.Suresh C. Sharma**, Applied Physics, DTU

49. Miniaturized Quad-Port Conformal Multi-Band (QPC-MB) MIMO Antenna for On-Body Wireless Systems in Microwave-Millimeter Bands, **7.1.Manish Sharma**, Prabhakara Rao Kapula, Shailaja Alagrama, Kanhaiya Sharma, **7.1.Ganga Prasad Pandey**, Dinesh Kumar Singh, Milind Mahajan and Anupma Gupta, ECE, DTU
50. Mixing Enhancement in Vortex Serpentine Micromixer Having Two and Four Non-Aligned Inlets, **7.Deepak Kumar**, **7.Abhyuday Singh Latwal**, **3.Mohammad Zunaid** and **Samsher**, Mechanical, DTU
51. Modeling and analysis for enhanced hydrogen production in process simulation of methanol reforming, **6.Neeraj Budhraja**, **3.Amit Pal** & **3.R.S. Mishra**, Mechanical, DTU
52. Modeling of water surface profile in non-prismatic compound channels, **7.Vijay Kaushik**, **3.Munendra Kumar**, Bandita Naik and Abbas Parsaie, Civil, DTU
53. Modified High Gain Non-Isolated Boost DC-DC Converter for Electric Vehicles, **8.Mohd Adib** and **3.Saurabh Mishra**, Electrical, DTU
54. Monoclinic to cubic structural transformation, local electronic structure, and luminescence properties of Eu-doped HfO<sub>2</sub>, **3.Rajesh Kumar**, Jitender Kumar, Ramesh Kumar, Akshay Kumar, Aditya Sharma, S. O. Won, K. H. Chae, **3.Mukhtiyar Singh** and Ankush Vij, Applied Physics, DTU
55. MTSO: Multi-Target Search Optimisation based on Probability Map, **8.Aman Virmani**, **8.Vayam Jain**, **8.Nilesh Aggarwal**, **8.Arjun Gupta**, **8.Anunay** and **3.Dr. Anup Kumar Mandpura**, Electrical, Applied Physics, Electronics and Mechanical, DTU
56. Multimodal Sarcasm Recognition by Fusing Textual, Visual and Acoustic content via MultiHeaded Attention for Video Dataset, **8.Sajal Aggarwal**, **8.Ananya Pandey** and **3.Dinesh Kumar Vishwakarma**, IT, DTU
57. Multi-objective optimization of mechanical properties of chemically treated bio-based composites using response surface methodology, Ankit Manral, Rakesh Singh, **6.Furkan Ahmad**, Partha Pratim Das, Vijay Chaudhary, Rahul Joshi and Pulkit Srivastava, Mechanical, DTU

58. Novel Band-Subtraction Technique to Differentiate Screws for Microwave Cavity Filter Tuning, Even Sekhri, Mart Tamre, **3.Rajiv Kapoor** and Dhanushka Chamara Liyanage, Electronics, DTU
59. Optimal Harmonic Current Extractor using Digital Warped Filter for a Single – Phase PV Integrated Grid-Tied System with 5-Level DSTATCOM, **6.Praveen Bansal** and **3.Alka Singh**, Electrical, DTU
60. Performance Evaluation of Machine Learning Methods for Detecting Credit Card Fraud, **8.Anuj Yadav**, **8.Arpaajit Adhikary**, **8.Aryan Kainth** and **3.Rohit Kumar**, Electronics, DTU
61. Physics Based Numerical Model of a Nanoscale Dielectric Modulated Step Graded Germanium Source Biotube FET Sensor: Modelling and Simulation, Amit Das, **3.Sonam Rewari**, Binod Kumar Kanaujia, S.S. Deswal and R.S. Gupta, ECE, DTU
62. Polymer nanocomposite film based piezoelectric nanogenerator for biomechanical energy harvesting and motion monitoring, **6.Shilpa Rana** and Bharti Singh, Applied Physics, DTU
63. Potato Peel Waste as an Economic Feedstock for PHA Production by Bacillus circulans, **6.Sonika Kag**, **3.Pravir Kumar** and **6.Rashmi Kataria**, Biotechnology, DTU
64. Power Flow Management of Solar PV fed Switched Boost Inverter, **8.Srishti Singh**, **8.Vansh Aggarwal** and **3.Mayank Kumar**, Electrical, DTU
65. Pressure Induced Surface States and Wannier Charge Centers in Ytterbium Monoarsenide, **6.Ramesh Kumar**, **3.Rajesh Kumar**, **3.Sangeeta** and **3.Mukhtiyar Singh**, Applied Physics, DTU
66. Rainfall Assessment and Water Harvesting Potential in an Urban area for Artificial Groundwater Recharge with Land Use and Land Cover Approach, **6.Ali Reza Noori** and **3.S.K. Singh**, Environmental, DTU
67. Role of Surface-Chemistry in Colloidal Processing of Ceramics: A Review, **6.Megha Bansal**, **3.Deenan Santhiya** and S. Subramanian, Biotechnology and Applied Chemistry, DTU



68. Socialization of the Importance of Building English Skills for Elementary School Children at SD Negeri 122395 Pematang Siantar, Vera Elisabet Siahaan, Angel Nerin Patricia Panggabean, Hanji Agustina Sitohang, Theresi Theresi, Santa Veronika Situmorang, Fitrianti Manurung, Herman Herman, Yanti Kristina Sinaga, Irene Adryani Nababan and **6.Prakash Puhka**, DTU
69. Solar light and ultrasound-assisted rapid Fenton's oxidation of 2,4,6-trichlorophenol: comparison, optimisation, and mineralization, **6.Shivani Yadav**, Sunil Kumar and **3.Anil Kumar Haritash**, Environmental, DTU
70. Speech Dereverberation with Frequency Domain Autoregressive Modeling, Anurenjan Purushothaman, Debottam Dutta, **8.1.Rohit Kumar** and Sriram Ganapathy, Electronics, DTU
71. Strategies in Design of Self-Propelling Hybrid Micro/Nanobots for Bioengineering Applications, Saurabh Shivalkar, Anwesha Roy, **7.Shrutika Chaudhary**, Sintu Kumar Samanta, Pallabi Chowdhary and Amaresh Kumar Sahoo, Biotechnology, DTU
72. Synergistic action of nano silica and w/b ratio on accelerated durability performance of concrete, Satish Kumar Chaudhary, **6.1.Ajay Kumar Sinha** and Praveen Anand, DTU
73. Synthesis and Wear Behaviour Analysis of SiC- and Rice Husk Ash-Based Aluminium Metal Matrix Composites, Sameen Mustafa, Julfikar Haider, Paolo Matteis and **3.Qasim Murtaza**, Mechanical, DTU
74. Topological phase transition and tunable surface states in YBi, **6.Ramesh Kumar** and **3.Mukhtiyar Singh**, Applied Physics, DTU
75. Traffic Prediction Model Using Machine Learning in Intelligent Transportation Systems, **3.Abhilasha Sharma** and **7.Prabhat Ranjan**, Software Engineering, DTU
76. TransNet: a comparative study on breast carcinoma diagnosis with classical machine learning and transfer learning paradigm, **6.Gunjan Chugh**, **3.Shailender Kumar** and Nanhay Singh, CSE, DTU

77. Unraveling the intricate relationship: Influence of microbiome on the host immune system in carcinogenesis, **8.Saksham Garg**, **8.Nikita Sharma**, **8.Bharmjeet** and **3.Asmiita Das**, Biotechnology, DTU
78. Unravelling the Ultralow Thermal Conductivity of Ternary Antimonide Zintl Phase RbGaSb<sub>2</sub>: A First-principles Study, **3.Sangeeta**, **3.Rajesh Kumar**, **6.Ramesh Kumar**, **6.Kulwinder Kumar** and **3.Mukhtiyar Singh**, Applied Physics, DTU
79. Unsupervised sentiment analysis of Hindi reviews using MCDM and game model optimization techniques, **6.NEHA PUNETHA** and **3.GOONJAN JAIN**, Applied Mathematics, DTU
80. Viscous fluid dynamics with decaying vacuum energy density, **3.C. P. Singh** and **6.Vinita Khatri**, Applied Mathematics, DTU
81. Vision Transformer Based Devanagari Character Recognition, **Shailendra Kumar**, **Abhinav Chopra**, **Sambhav Jain** and **Sarthak Arora**, DTU
82. Water quality management by enhancing assimilation capacity with flow augmentation: a case study for the Yamuna River, Delhi, **6.Nibedita Verma**, **3.Geeta Singh** and Naved Ahsan, Environmental, DTU

1. Vice Chancellor

1.1. Ex Vice chancellor

2. Pro Vice Chancellor

2.1. Ex Pro Vice Chancellor

3. Faculty

3.1. Ex Faculty

4. Teaching-cum-Research Fellow

4.1. Alumni

5. Asst. Librarian

5.1 Others

6. Research Scholar

6.1. Ex Research Scholar

7. PG Scholar

7.1. Ex PG Scholar

8. Undergraduate Student

8.1. Ex Undergraduate Student



# A Dynamic and Efficient Active Power Support Scheme Using VSC-HVDC System for Quick Frequency Restoration

Ashima Taneja<sup>1</sup> · Radheshyam Saha<sup>1</sup> · Madhusudan Singh<sup>1</sup>

Received: 25 January 2023 / Accepted: 13 July 2023  
© King Fahd University of Petroleum & Minerals 2023

## Abstract

For achieving carbon neutrality, increased renewable power generation is an emerging trend. Renewable power, often transmitted via VSC-HVDC, is variable, intermittent and could not contribute to grid inertia directly. Thus, quick restoration of system frequency should be prioritized, for which encapsulation of frequency regulators inside converter controllers is often suggested. Most of the reported frequency regulation schemes are only suited for converters carrying power below rated limits. The novelty of present work lies in its ability to regulate frequency by delivering power support above rated limits. In fact, by exploiting primary frequency reserves available at the other end and by efficiently utilizing converter's entire permissible operating region, frequency can be restored quickly within acceptable range. This can have significant implication for extracting frequency support from offshore wind farms coupled via VSC-HVDC. The proposed technique, based on dynamic converter current modulation, renders significant active power support depending upon severity of encountered frequency excursion. Its performance is justified by testing it by introducing disturbances of different magnitudes in an IEEE 14-bus system. Interestingly, the simulation results obtained using MATLAB/Simulink have confirmed that by employing proposed scheme, for a 22% load/generation unbalance, VSC can arrest frequency nadir within one second by delivering nearly 30% above rated converter power. In fact, fast frequency regulation is achieved without compromising AC voltage stability of the disturbed grid.

**Keywords** Fast frequency regulation · Load–frequency dynamics · VSC's permissible operating range · Converter current modulation · VSC-HVDC overload capability · VSC-HVDC operation beyond rated limits · Power–frequency control

## 1 Introduction

High voltage direct current (HVDC) technology has gained popularity in last few decades owing to its ability to transfer high quantum of power generated by various renewable energy sources (RES) located many miles away from congested load centers. With its wide variety of features like flexible power control, black starting, faster and effortless power reversal, smaller station footprints, etc. Voltage source converter (VSC)-based HVDC technology is a convenient choice made by power engineers for enabling green grid electrification. Although HVDC technology makes ties between asynchronous grids feasible, simultaneously it makes it difficult to extend frequency support (FS) from one grid to another

grid directly. Therefore, a frequency regulation feature is often incorporated inside converter control so that interregional frequency support can be rendered [1–3].

It is noted from literature survey [1–22] that various control schemes have been suggested for facilitating FS incorporating VSC-HVDC system. For a point-to-point (P2P) VSC-HVDC system, [4] introduces an active power–frequency ( $P$ – $f$ ) droop in its converter control which increases coupling between the grids causing HVDC system to no longer behave as a firewall, thus, resulting in FS provision from one grid to another. Likewise, FS is proposed by using offshore frequency-DC voltage ( $f_{\text{off}}-V_{\text{dc}}$ ) droop at rectifier controller and  $P$ – $f$  droop at inverter controller [5, 14]. The same is demonstrated in [6]; however, converter power carrying capability is under-utilized. It is shown in [7, 24, 25] that IGBT-based modular multilevel converter (MMC)-HVDC can be overloaded to up to 27.5% of its rated capacity by controlled injection of circulating currents while still functioning within safe junction temperature

✉ Ashima Taneja  
tanejaashima.1988@gmail.com

<sup>1</sup> Department of Electrical Engineering, Delhi Technological University, Delhi, India



of the used power electronic devices. However, resultant impact on converter's reactive power support and AC voltage of the disturbed grid during the overload period has not been accounted for. A multivariate random forest regression (MRFR) algorithm for FS provision in interconnected grids of a P2P-VSC-HVDC system has been proposed in [8] which estimates the power required for the caused load/generation unbalance and accordingly updates power order for converter stations and alternators. However, effect of time delay involved in parameter estimation as well as updating of the power orders on system's performance is not touched well. Inertia emulation-based control schemes are also proposed. [9, 10] use electrostatic energy stored in converter station capacitors for extracting frequency support without disturbing the healthy grid at the other end. However, super-sized capacitors are required instead for provision of substantial power support. In [11], a communication-less bidirectional frequency support scheme is proposed for P2P-VSC-HVDC system in which the grid having more rate of change of frequency (RoCoF) is supported from the grid at the other end. But, along with pre-requisite of accurate RoCoF measurement, power support from healthy grid to disturbed grid is not ensured always as no clear demarcation to differentiate between supporting and disturbed grid is suggested. Virtual synchronous generator (VSG)-based frequency support schemes are proposed in [12, 13] in which VSC is artificially

made to behave as an alternator to participate in frequency stability enhancement. However, such techniques lack of inherent converter current limiting controls which has to be additionally integrated. Table 1 enlists and compares features of various frequency regulation schemes listed in literature.

It is also noted that most of these FS strategies are demonstrated with VSC-HVDC carrying power significantly lesser than its rated power. Except [7], none of these works consider possibility of FS provision particularly when the converter is already carrying rated power and is expected to deliver power support above its rated limits. Also, all of these strategies are demonstrated only for case of nominal disturbances and are not tested for any credible contingencies. In fact, system response is indifferent to frequency excursions of different magnitudes. In addition, effect of load/generation unbalances on frequency dynamics of the interconnected grids is not quantized. It is also noted that additional support power is released steadily causing significant nadir and hence fast frequency restoration is not achieved. In fact, none of these demonstrates the effect of frequency mitigation technique on AC voltage stability of the disturbed grid.

The proposed work presents a dynamic frequency regulation technique using VSC-HVDC system that enables enhanced active power support for enabling the quick restoration of system frequency. The proposed frequency regulation

**Table 1** List of pros, cons and limitations of significantly reported frequency regulation strategies

S. no.	Existing articles	Pros	Cons	Limitations
1	Droop-based [4, 7, 11, 14]	Able to offer instantaneous response. Does not rely on accurate RoCoF measurement.	Offer limited power support capability. Slow frequency restoration speed.	Not suitable for HVDC system already operating at rated capacity. Unable to utilize efficiently the entire permissible VSC operating region.
2	Inertia emulation based [9, 10]	Instantaneous response is offered. Firewall advantage of HVDC is retained.	DC voltage stability is at risk. Frequency restoration speed depends upon emulated inertia constant.	Large investment cost for super capacitors at the converter station. Unable to utilize efficiently the entire permissible VSC operating region.
3	Direct estimation of unbalanced power [8]	Exact amount of deficient/surplus power can be compensated.	Requires accurate RoCoF measurement. Slow frequency restoration speed.	Needs dedicated communication system for parameter estimation and updating power order. Requires accurate estimation of unbalanced power.
4	VSG-based [12, 13]	Able to provide both inertial and primary frequency support.	Slow frequency recovery. Need accurate RoCoF measurement. Adopted power synchronization control is difficult to implement.	Overcurrent limitation control for valve protection needs to be additionally integrated. Not suitable for HVDC system already operating at rated capacity.



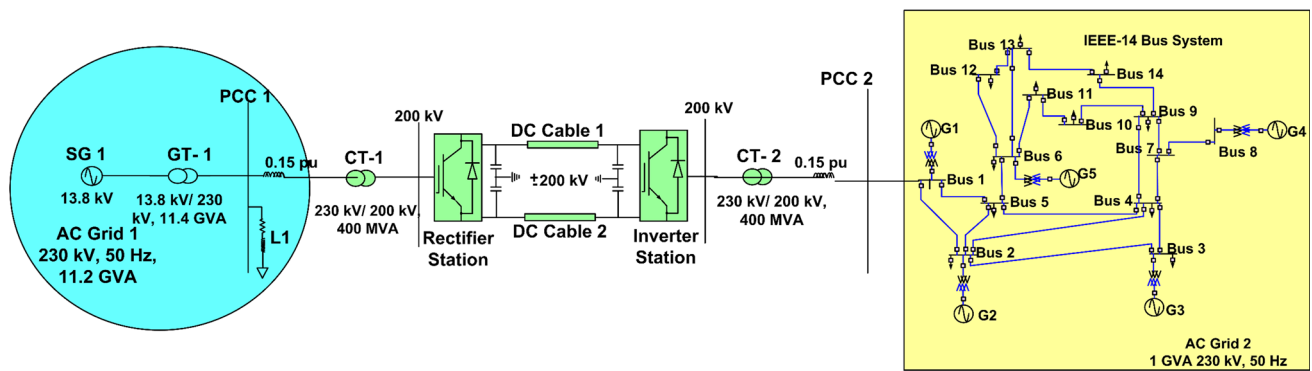


Fig. 1 VSC-HVDC transmission system coupling two AC grids

technique is based on a three-layer converter control mechanism in which significant power support is rendered depending upon the magnitude of encountered frequency excursion. In present paper, it is shown that appreciable power support (nearly 30% above rated) even for credible frequency disturbances (for load/generation unbalances up to 22%) can be imparted. This can result in significant implications for extracting frequency support from offshore wind farms so that additional power support above rated converter capacity can be transmitted via VSC-HVDC controlled via proposed scheme. In the proposed work, with increase in severity of frequency contingency, the entire permissible operating region of VSC-HVDC has been exploited for provision of frequency support while deliberately maintaining AC voltage stability of the disturbed network. Although, few research works mention similar type of approach [15, 16], but are silent over its resultant impacts on the AC/DC system and are unable to justify better frequency regulation that can be achieved with the aforesaid concept.

The major contributions of this paper may be mentioned as:

1. Modeling and verification of load frequency dynamics of disturbed and supporting grids at the two ends of VSC-HVDC transmission system.
2. Demonstration of utilization of converter's complete permissible operating range so as to offer maximum frequency support to the disturbed grid.
3. Performing stability analysis of the considered AC/DC system by deriving a dynamic state space model of it.
4. Verifying proposed scheme against frequency excursions of various magnitudes.

With Introduction as Sect. 1, this paper has six more sections. The description for the considered VSC-HVDC system coupling two different AC grids is presented in Sect. 2.

Development of mathematical model for frequency deviations in a contingency prone AC network of the integrated HVAC-HVDC grid is done in Sect. 3. The efficient utilization of VSC's permissible operating region for obtaining enhanced active power support and thus, for mitigating frequency excursion of different severities is proposed in Sect. 4. The stability analysis of the system is carried out in Sect. 5. Simulation results that justify the enhanced FS capability of the proposed control scheme are demonstrated in Sect. 6. In the end, Sect. 7 draws out conclusions.

## 2 System Configuration

A P2P-VSC-HVDC system interconnecting two AC grids, AC Grid 1 (ACG 1) and AC Grid 2 (ACG 2) is shown in Fig. 1. Rectifier controller maintains its local DC voltage to constant value while inverter controls its active power injection into ACG 2. The parameters for this system are given in Table 2. ACG 1 is represented by a single generator equivalent, SG-1, while ACG 2 is considered as IEEE 14-bus system having five alternators, G1–G5. With several alternators, loads and buses in it and to optimize simulation times as well, thus, IEEE-14 bus system is a better selection so that actual power system behavior can be visualized. The VSC-HVDC system is coupled via PCC-2 which is connected to Bus-1 of the IEEE bus system by a reactive impedance. With system voltage of 230 kV, short circuit capacity of the Bus-1 is fairly strong and determined to be 2000 MVA. This can be sufficient to absorb the inverter's power infeed, operating even above its rated limits. A two-level circuit with symmetrical monopolar configuration is chosen for VSC-HVDC system. For alternators, excitation system of Type 1 [26] and IEEE turbine model with PID controller for its governor system [27] are adopted.



**Table 2** AC-DC system parameters

S. no.	System components	Ratings
<i>AC grid 1 (ACG 1)</i>		
	System voltage and frequency	230 kV, 50 Hz
	Synchronous generator equivalent (SG 1):	
	Rated power	11.2 GVA
	Active power generation	0.75 pu
	Governor droop coefficient	0.05
	Inertia time constant	3.2 s
	Generator transformer equivalent (GT-1)	
	Voltage	13.8 kV/230 kV
	Rated MVA	11.4 GVA
<i>AC grid 2 (ACG 2)</i>		
	System voltage and frequency	230 kV, 50 Hz
	Alternators (G1, G2, G3, G4, G5)	
	Rated voltage	13.8 kV
	Rated MVA	200 MVA
	Active power generation	0.6 pu
	Governor droop coefficient	0.05
	Inertia time constant	3.5 s
	Generator transformers	
	Voltage	13.8 kV/230 kV
	Rated MVA	210 MVA
<i>VSC-HVDC system parameters</i>		
	Rated power capacity	400 MW
	Rated DC voltage	$\pm 200$ kV
	Converter transformers (CT-1 and CT-2)	230 kV/200 kV, 400 MVA, 50 Hz
	Length of DC cable 1 and 2	75 km
	Parameters of DC cable 1 and 2	$r = 0.139$ m $\Omega$ /km, $l = 15.9$ mH/km, $c = 23.1$ $\mu$ F/km
	DC capacitance	70 $\mu$ F
<i>VSC-HVDC controller parameters</i>		
a	<i>P-f</i> controller at inverter station	
	Active power setpoint	$-1$ pu
	Proportional and integral gain	0, 20
b	<i>Q</i> -controller at rectifier and inverter station	
	Reactive power setpoint	$-0.2$ pu
	Proportional and integral gain	0, 20
c	<i>V<sub>DC</sub></i> controller at rectifier station	
	DC voltage setpoint	1 pu

**Table 2** (continued)

S. no.	System components	Ratings
	Proportional and integral gain	2, 40
d	Inner Current Controller	
	Proportional and integral gain	0.6

### 3 Modeling Frequency Deviation Dynamics in AC Grids Interconnected to VSC-HVDC System

The VSC-HVDC system interconnecting two AC grids, as shown in Fig. 1, is considered. Assuming that ACG 1 and ACG 2 are having net active power generation of  $P_{SG1.net}$  and  $P_{SG2.net}$ , respectively.

It is considered that all the five alternators of ACG 2 are loaded equally and  $P_G$  is the power generated by each of them. With ' $n_g$ ' equal to total number of alternators, i.e., five here,

$$P_{SG2.net} = n_g \cdot P_G. \quad (1)$$

Let  $P_{L1}$  and  $P_{L2}$  are total power drawn by the loads in ACG 1 and ACG 2, respectively. Assuming  $P_{rect}$  and  $P_{inv}$  are active power flowing via rectifier and inverter stations, respectively. Let  $f_2^{rated}$  and  $f_2$  be rated and actual frequency of ACG 2 and  $P_{inv}^{rated}$  is rated value of inverter power. Let  $K_f$  is power-frequency coefficient (PFC) of inverter, such that

$$K_f = -\frac{\Delta P_{inv} / P_{inv}^{rated}}{\Delta f_2 / f_2^{rated}}, \quad (2)$$

where *PFC* is defined as ratio of per unit increase made in inverter power in response to per unit decrease in frequency of the connected grid.

Considering a load/generation unbalance in ACG 2,

$$\Delta P_{inv} + n_g \cdot \Delta P_G = \Delta P_{L2}. \quad (3)$$

Using (2),

$$\Delta P_{inv} = -P_{inv}^{rated} \cdot K_f \cdot \frac{\Delta f_2}{f_2^{rated}}. \quad (4)$$

Neglecting any switching and line losses,

$$\Delta P_{rec} = \Delta P_{inv}. \quad (5)$$

With  $P_G^{\text{rated}}$  as rated power and  $\rho_{AC,2}$  as governor droop coefficient of each of the alternators in ACG 2,

$$\Delta P_G = -\frac{P_G^{\text{rated}}}{\rho_{AC,2}} \cdot \frac{\Delta f_2}{f_2^{\text{rated}}} \quad (6)$$

### 3.1 Frequency Deviation in Disturbed Grid, ACG 2

Using (4) and (6) in (3),

$$\Delta f_2 = -\frac{\Delta P_{L2}}{\left( \frac{P_{\text{inv}}^{\text{rated}}}{K_f} + \frac{n_g \cdot P_G^{\text{rated}}}{\rho_{AC2}} \right)} f_2^{\text{rated}} \quad (7)$$

Thus, lesser frequency deviation occurs, when higher value of PFC ( $K_f$ ) is considered at inverter controller.

### 3.2 Frequency Deviation in Supporting Grid, ACG 1

Assuming no frequency disturbance occurring in ACG 1 at the same time of the load/generation unbalance event in ACG 2 and neglecting any losses,

$$\Delta P_{SG1.net} = \Delta P_{\text{rect}} \quad (8)$$

Using governor droop constant,  $\rho_{AC1}$  for SG 1,

$$\Delta P_{SG1.net} = -\frac{P_{SG1}^{\text{rated}}}{\rho_{AC1}} \cdot \frac{\Delta f_1}{f_1^{\text{rated}}} \quad (9)$$

where  $\Delta f_1$  is the frequency deviation caused in ACG 1 while mitigating frequency deviation in ACG 2.  $P_{SG1}^{\text{rated}}$  and  $f_1^{\text{rated}}$  are the rated values of active power of SG 1 and frequency of ACG 1, respectively. From (5) and (8),

$$\Delta P_{SG1.net} = \Delta P_{\text{inv}} \quad (10)$$

Using (9), (4), (7) in (10),

$$\Delta f_1 = -\frac{\Delta P_{L2}}{\frac{P_{SG1}^{\text{rated}}}{\rho_{AC1}} \left( 1 + \frac{n_g \cdot P_G^{\text{rated}}}{K_f \rho_{AC2}} \right)} f_1^{\text{rated}} \quad (11)$$

It can be noted from (7) and (11) that the supporting grid (ACG 1) undergoes frequency deviation of same nature as that of the disturbed grid (ACG 2).

## 4 Proposed Dynamic and Efficient Active Power Support Scheme for Quick Frequency Restoration

In conventional vector current control used for VSC-HVDC transmission system, so as to protect IGBT valves from overcurrent, reference values of direct and quadrature components of converter currents ( $i_d^*$  and  $i_q^*$ ) are limited to ceiling values of  $i_{d,upper}$  and  $i_{q,upper}$ , respectively [9, 17]. Clearly, the total permissible current that can be carried by converter valves, i.e.,  $i_{VSC,upper}$  is vector sum of  $i_{d,upper}$  and  $i_{q,upper}$ . Figure 2 shows the permissible region of operation for the inverter of the VSC-HVDC transmission system (Fig. 1). The left half of the  $i_q^*$  axis denotes inverter operation while the right half denotes the rectifier operation. Also, positive value of  $i_q^*$  signifies supply of reactive power to the connected grid from the VSC while negative value of  $i_q^*$  signifies absorption of excess reactive power from the grid by the VSC. At a time, only one of the two reactive power control objectives is selected. Due to the imposed converter current limits, VSC conventionally operates inside rectangular region,  $ABCD$ .

For any general power flow scenario, power and voltage controllers of VSC-HVDC are given set points according to which  $i_d^*$  and  $i_q^*$  are generated inside converter controller which may not be necessarily equal to  $i_{d,upper}$  and  $i_{q,upper}$ . Assuming that rated values of active and reactive power are exchanged by the inverter with ACG 2 and accordingly, rated values of direct and quadrature components of converter current are considered as  $i_{d,rated}$  and  $i_{q,rated}$  respectively. Thus, inverter operates inside the operating region defined by  $PQRS$  in Fig. 2. Assuming that the inverter is injecting reactive power into ACG 2, then point  $P$  is its operating point having coordinates  $(-i_{d,rated}, i_{q,rated})$ .

It is assumed that maximum current carrying capability of interconnecting cables/lines with the converter is at least equal to  $i_{VSC,upper}$  and sufficient power support is available from the grid connected with rectifier end of VSC-HVDC system, i.e., ACG 1 of Fig. 1. In case of a contingency in grid

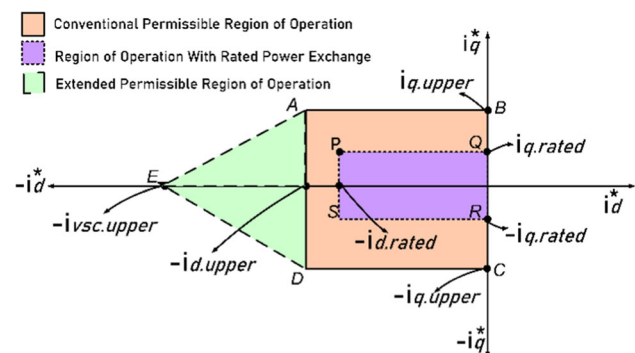


Fig. 2 Permissible operation region for inverter of VSC-HVDC system





connected with inverter end, i.e., inside ACG 2, it is possible to extend maximum amount of frequency support from ACG 1 to ACG 2, by allowing  $i_d^*$  at the converter controller to take value beyond  $i_{d,upper}$  and nearly as high as  $i_{vsc,upper}$ . Thus, the region of operation for VSC-HVDC now gets extended additionally beyond the rectangular region, ABCD to triangular region ADE, which is also shown in Fig. 2.

Assuming inverter is already carrying rated power, using (4) and from [9],  $i_d^*$  is generated inside the inverter controller as:

$$i_d^* = \frac{1 + K_f(f_2^{\text{rated}} - f_2)/f_2^{\text{rated}}}{1.5V_{d2}}, \quad (1)$$

where  $V_{d2}$  being direct-axis component of AC voltage at PCC-2 in Fig. 1. For a frequency excursion in ACG 2,

$$\Delta i_d^* = -\frac{K_f}{1.5 * V_{d2}} \frac{\Delta f_2}{f_2^{\text{rated}}}. \quad (2)$$

Thus, change in  $i_d^*$  made by the inverter controller is proportional to the magnitude of frequency deviation,  $\Delta f_2$  in ACG 2.

Three scenarios are possible, depending upon the change in  $i_d^*$  in response to magnitude of the encountered frequency deviation. These are listed as follows:

- Scenario 1:  $i_d^*$  is limited to  $i_{d,upper}$

In this scenario, the range of encountered frequency deviations are such that magnitude of  $i_d^*$  can be increased from its pre-disturbance value of  $i_{d,rated}$  to up to  $i_{d,upper}$ . Hence, the converter's operating region, PQRS, gets extended in comparison to earlier and is depicted in Fig. 3a. Also, the operating point, P, now takes coordinates in between  $(-i_{d,rated}, i_{q,rated})$  to  $(-i_{d,upper}, i_{q,rated})$ , depending upon the value of  $\Delta f_2$ .

Usually,  $i_{d,upper}$  is 10% more than  $i_{d,rated}$  [18]. Since  $i_d^*$  increases only up to 10% from its rated value, thus only smaller scale frequency deviations can be regulated in this scenario. This case can, thus, be referred as 'Small Frequency Excursion'. Let  $i_{d,upper} = m * i_{d,rated}$ , constant 'm' is 1.1. Thus, using (13), the frequency deviation gets limited to

$$\begin{aligned} -\frac{\Delta f_{2,small}}{f_2^{\text{rated}}} &= \frac{1.5V_{d2}}{K_f} * (i_{d,upper} - i_{d,rated}) \\ &= \frac{1.5V_{d2}}{K_f} * (m - 1) * i_{d,rated}. \end{aligned} \quad (14)$$

In addition, same  $i_q^*$  is maintained, i.e., equal to  $i_{q,rated}$ ; thus, reactive power support offered by the inverter remains unaltered in this scenario. By the converter current modulation performed in this scenario, it is taken care that frequency

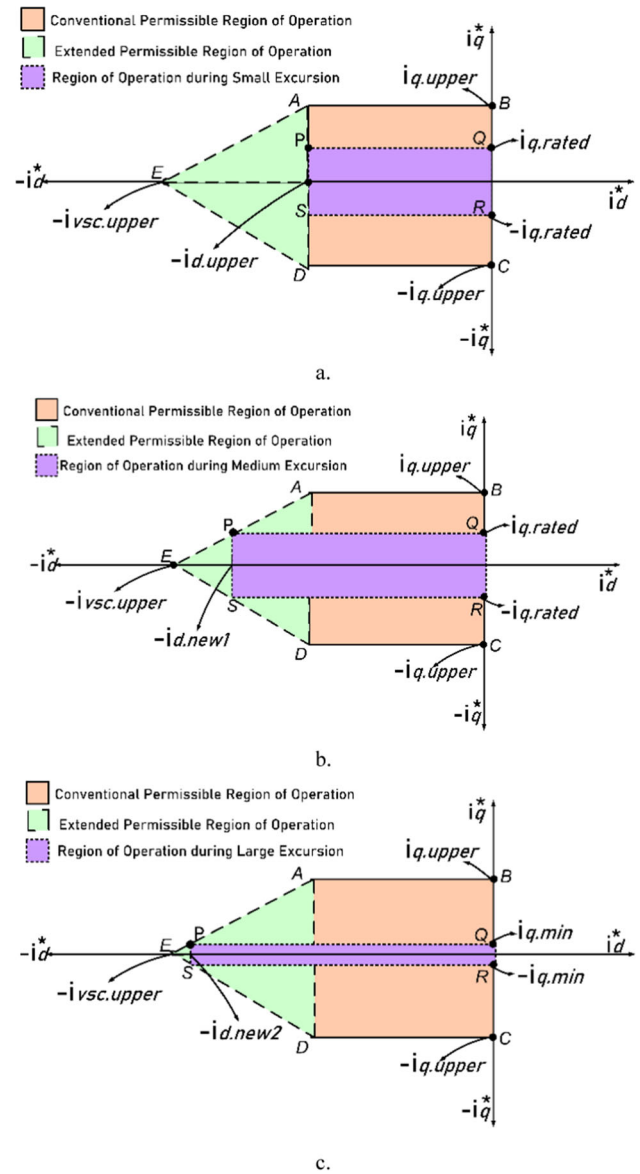


Fig. 3 Inverter's operating region during a small, b medium and c large frequency excursion

deviation caused by any load/generation unbalance should not exceed beyond value defined by (14). It can be observed from Fig. 3a that in this scenario, inverter still operates inside the operating region that has been defined conventionally.

- Scenario 2:  $i_d^*$  is limited between  $i_{d,upper}$  and  $i_{vsc,upper}$

This case is applicable when encountered frequency deviations are expected to be more than the previous case. So, it is termed as 'Medium frequency excursion'. It is proposed to increase  $i_d^*$  from  $i_{d,rated}$  to a value beyond  $i_{d,upper}$ . Since same pre-disturbance reactive power support from VSC should be maintained and its current should not exceed



**Table 3** VSC current modulation for possible frequency excursion scenarios

Scenario↓	Frequency deviation	$i_d^*$	$i_q^*$	Net VSC current, $i_{vsc}$
Steady state	$\Delta f_2 = 0$	$i_d^* = i_{d.rated}$	$i_q^* = i_{q.rated}$	$i_{vsc} = \sqrt{(i_d^*)^2 + (i_q^*)^2}$
Small excursion	$0 > -\Delta f_2 \geq -\Delta f_{2.small}$	$-i_{d.upper} \leq -i_d^* \leq -i_{d.rated}$		$i_{vsc} \leq i_{vsc.upper}$
Medium excursion	$-\Delta f_{2.small} > -\Delta f_2 \geq -\Delta f_{2.medium}$	$-i_{d.new.1} \leq -i_d^* \leq -i_{d.upper}$		
Large excursion	$-\Delta f_{2.medium} > -\Delta f_2 \geq -\Delta f_{2.large}$ or $-\Delta f_{2.large} > -\Delta f_2$	$-i_{d.new.2} \leq -i_d^* \leq -i_{d.new.1}$	$i_q^* = i_{q.min}$	

$i_{vsc.upper}$ , thus  $i_d^*$  should be limited to  $i_{d.new.1}$ , where  $i_{d.new.1} = \sqrt{i_{vsc.upper}^2 - i_{q.rated}^2}$ . Also, the operating point  $P$ , which was originally at  $(-i_{d.rated}, i_{q.rated})$ , can now further be advanced up to  $(-i_{d.new.1}, i_{q.rated})$ ; as shown in Fig. 3b.

Usually,  $i_{vsc.upper}$  is 30% to 50% more than  $i_{d.rated}$  [18]. Also,  $i_{q.rated}$  is lesser than  $i_{d.rated}$  [17]. So, assuming  $i_{vsc.upper} = k * i_{d.rated}$  and  $i_{q.rated} = p * i_{d.rated}$  where ‘ $k$ ’ and ‘ $p$ ’ are constants such that  $1.3 < k < 1.5$  and  $0 < p < 1$ . Using  $i_{d.new.1}$  in (13), frequency deviations can be limited up to

$$-\frac{\Delta f_{2.medium}}{f_2^{rated}} = \frac{1.5V_{d2}}{K_f} \cdot (i_{d.new.1} - i_{d.rated})$$

$$= \frac{1.5V_{d2}}{K_f} \cdot \left( \sqrt{k^2 - p^2} - 1 \right) i_{d.rated}. \quad (15)$$

Thus, in this scenario, converter current is modulated in such a way that reactive power support offered by the inverter to ACG 2 remains unaltered while frequency deviations are not allowed to exceed beyond the value defined by (15).

- Scenario 3:  $i_d^*$  is limited to  $i_{vsc.upper}$

So as to mitigate frequency excursions which are even larger than previous two scenarios, it becomes necessary to prioritize inverter’s active power support over its reactive power support. In addition, so as to maintain frequency stability along with AC voltage stability of the disturbed grid (ACG 2), it is proposed to reduce  $i_q^*$  to a minimum value,  $i_{q.min}$ . Therefore,  $i_d^*$  can now further be enhanced from  $i_{d.new.1}$  to a value equal to  $i_{d.new.2}$ , where  $i_{d.new.2} = \sqrt{i_{vsc.upper}^2 - i_{q.min}^2}$ . This scenario is referred as large frequency excursion. This is shown in Fig. 3c and after performing this current modulation, the operating point  $P$  is now at  $(-i_{d.new.2}, i_{q.min})$ .

Let  $i_{q.min} = \frac{i_{q.rated}}{d}$ , where  $d$  is constant such that  $d > 1$ , using (13),

$$-\frac{\Delta f_{2.large}}{f_2^{rated}} = \frac{1.5V_{d2}}{K_f} \cdot (i_{d.new.2} - i_{d.rated})$$

$$= \frac{1.5V_{d2}}{K_f} \cdot \left( \sqrt{k^2 - \frac{p^2}{d^2}} - 1 \right) i_{d.rated}. \quad (16)$$

Thus, frequency deviations encountered in this scenario can be limited to value defined by (16). The proposed control scheme is dynamic because the amount of power support released depends upon the magnitude of frequency excursion. Secondly, it is efficient because entire permissible operation range of VSC is utilized to derive maximum amount of power support while simultaneously taking care of AC voltage stability. And since the active power support is rendered immediately, as a result excursion gets mitigated quickly. For above scenarios, modulation of the converter currents is summarized in Table 3.

#### 4.1 Control Structure for Proposed Scheme

The structure of proposed scheme is depicted in Fig. 4. The control structure has three layers:

- Outer control (OC) loop for generation of  $i_d^*$  and  $i_q^*$ .
- Dynamic current modulation (DCM) layer for performing modulation of converter currents according to frequency deviation in the interconnected grid (ACG 2).
- Inner current control (ICC) loop for generation of reference converter voltage,  $V_{c,abc}^*$ .

For regulation of grid frequency, inverter’s PFC,  $K_f$  is introduced in the outer loop.  $\Delta P$  is required modification in set point ( $P^*$ ) of active power controller in the OC loop of inverter in response to the encountered frequency deviation.  $Q_{inv}^{ref}$  and  $Q_{inv}$  are desired and actual reactive power



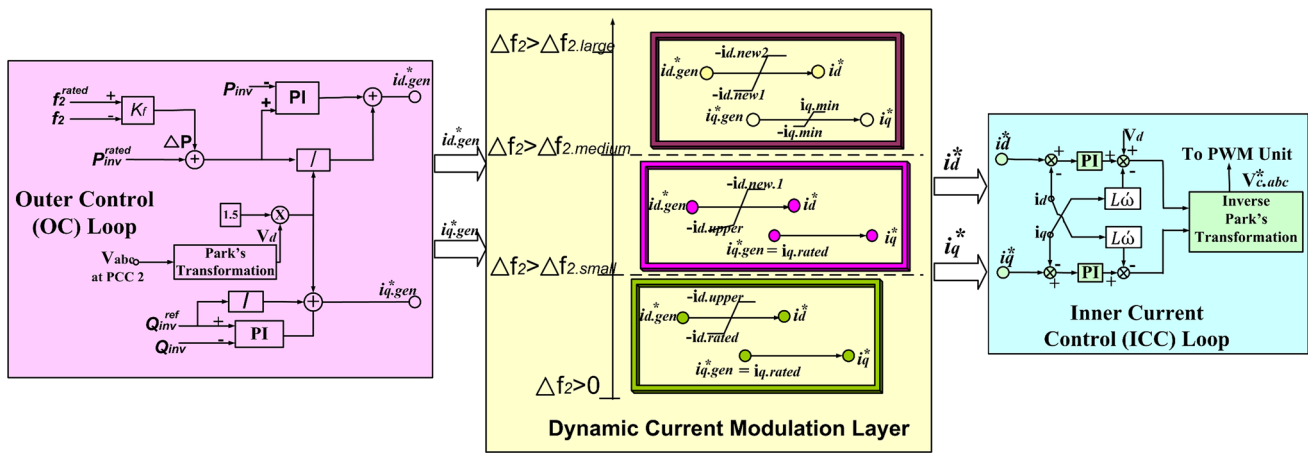


Fig. 4 Proposed three-layer control scheme for VSC-HVDC transmission system

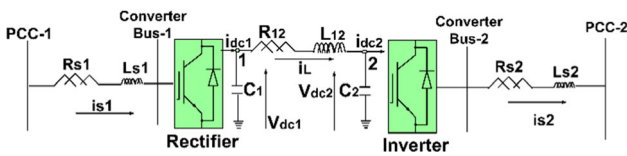


Fig. 5 Simplified AC/DC model

output of the inverter. It should be noted that unlike conventional control, converter current limitation is not performed in the OC loop. Rather, in the proposed scheme, depending upon the magnitude of frequency excursion encountered, converter currents are modulated inside the DCM layer which is buffered in between the outer and inner loop of the converter controller. This layer takes generated references for  $dq$  components of VSC currents as  $i_{d,gen}^*$  and  $i_{q,gen}^*$  from the OC loop and outputs final values as  $i_d^*$  and  $i_q^*$  to the inner loop after performing the current modulation as described earlier.

## 5 Stability Analysis of the Considered VSC-HVDC System

To study the stability analysis of the point-to-point VSC-HVDC system as shown in Fig. 1, its dynamic model is derived. A simple representation of the AC/DC system is shown in Fig. 5. The state space model of this system is derived by neglecting the converters switching losses and resistive losses of the grids. The stability analysis is carried out by performing small signal analysis.

### 5.1 Derivation of the State Space Model

The simplified model of Fig. 5 is divided into smaller parts as: ACG-1 dynamics, VSC-1 (rectifier) controller dynamics,

DC link dynamics, VSC-2 (inverter) controller dynamics and ACG-2 dynamics.

Assuming voltage at PCC-1 to be constant and in phase with direct-axis of synchronously rotating reference frame, writing equations for ACG-1:

$$\dot{i}_{s1,d} = -\frac{1}{L_{s1}}V_{c1,d} + \frac{1}{L_{s1}}V_{s1,d} - \frac{R_{s1}}{L_{s1}}i_{s1,d} + w_1 i_{s1,q}, \quad (17)$$

$$\dot{i}_{s1,q} = -\frac{1}{L_{s1}}V_{c1,q} - \frac{R_{s1}}{L_{s1}}i_{s1,q} - w_1 i_{s1,d} \quad (18)$$

where  $w_1$  is angular frequency of ACG-1,  $V_{s1}$  and  $V_{c1}$  are AC voltages at PCC-1 and converter bus-1, respectively,  $i_{s1}$  is current entering into converter bus-1 from PCC-1,  $R_{s1}$  and  $L_{s1}$  are resistance and inductance connected in between converter bus-1 and PCC-1.

Neglecting switching losses at rectifier,

$$i_{dc1} = \frac{V_{c1,d}i_{s1,d} + V_{c1,q}i_{s1,q}}{V_{dc1}}, \quad (19)$$

where  $i_{dc1}$  and  $V_{dc1}$  are current coming out of and voltage at rectifier.

The direct and quadrature components of converter voltage are obtained as output of ICC loop of VSC-1 as:

$$V_{c1,d} = V_{s1,d} - \left(k_{p,cc} + \frac{k_{i,cc}}{s}\right)(i_{s1,d}^{ref} - i_{s1,d}) - R_{s1}i_{s1,d} + L_{s1}w_1 i_{s1,q}, \quad (20)$$

$$V_{c1,q} = -\left(k_{p,cc} + \frac{k_{i,cc}}{s}\right)(i_{s1,q}^{ref} - i_{s1,q}) - R_{s1}i_{s1,q} - L_{s1}w_1 i_{s1,d}, \quad (21)$$

where  $k_{p,cc}$  and  $k_{i,cc}$  are proportional and integral gains of PI controllers used inside the ICC loop.

Having constant DC voltage control and constant reactive power control at rectifier station, the reference values of direct and quadrature axes currents which are obtained as output of the OC loop of VSC-1 controller as:

$$i_{s1,d}^{\text{ref}} = \left( k_{p,Vdc} + \frac{k_{i,Vdc}}{s} \right) (V_{dc1}^{\text{ref}} - V_{dc1}), \quad (22)$$

$$i_{s1,q}^{\text{ref}} = \frac{Q_{\text{rect}}^{\text{ref}}}{\frac{3}{2} V_{sd1}} + \left( k_{p,Qrect} + \frac{k_{i,Qrect}}{s} \right) (Q_{\text{rect}}^{\text{ref}} - Q_{\text{rect}}) \quad (23)$$

where  $k_{p,Vdc}$  and  $k_{i,Vdc}$  and  $k_{p,Qrect}$  and  $k_{i,Qrect}$  are proportional and integral gains of PI controllers for DC voltage control and reactive power control, respectively, used inside the outer loop.

In the DC grid, performing KCL at node-1 of Fig. 5,

$$\dot{V}_{dc1} = \frac{1}{C_1} i_{DC1} - \frac{1}{C_1} i_L, \quad (24)$$

$$\dot{V}_{dc2} = -\frac{1}{C_2} i_{DC2} + \frac{1}{C_2} i_L \quad (25)$$

After applying KVL in the DC link,

$$\dot{i}_L = \frac{1}{L_{12}} (V_{dc1} - V_{dc2}) - \frac{R_{12}}{L_{12}} i_L. \quad (26)$$

Similar to ACG-1, the equations for ACG-2 can be written as:

$$\dot{i}_{s2,d} = -\frac{1}{L_{s2}} V_{c2,d} - \frac{1}{L_{s2}} V_{s2,d} - \frac{R_{s2}}{L_{s2}} i_{s2,d} + w_2 i_{s2,q}, \quad (27)$$

$$\dot{i}_{s2,q} = \frac{1}{L_{s2}} V_{c2,q} - \frac{R_{s2}}{L_{s2}} i_{s2,q} - w_2 i_{s2,d} \quad (28)$$

The real and reactive power output of the inverter being represented as:

$$P_{\text{inv}} = V_{c2,d} i_{s2,d}, \quad (29)$$

$$Q_{\text{inv}} = V_{c2,q} i_{s2,q} \quad (30)$$

Similar to VSC-1, writing equations for inner current control loop for VSC-2 controller as:

$$V_{c2,d} = V_{s2,d} + \left( k_{p,cc} + \frac{k_{i,cc}}{s} \right) (i_{s2,d}^{\text{ref}} - i_{s2,d}) + R_{s2} i_{s2,d} - L_{s2} w_2 i_{s2,q}, \quad (31)$$

$$V_{c2,q} = \left( k_{p,cc} + \frac{k_{i,cc}}{s} \right) (i_{s2,q}^{\text{ref}} - i_{s2,q}) + R_{s2} i_{s2,q} + L_{s2} w_2 i_{s2,d}. \quad (32)$$

Inverter station having power–frequency control along with constant reactive power control, the output equations for the OC loop are written as:

$$i_{s2,d}^{\text{ref}} = \frac{P_{\text{inv}}^{\text{rated}} + K_f(f_2^{\text{rated}} - f_2)}{\frac{3}{2} V_{sd2}} + \left( k_{p,Pinv} + \frac{k_{i,Pinv}}{s} \right) (P_{\text{inv}}^{\text{rated}} - P_{\text{inv}}), \quad (33)$$

$$i_{s2,q}^{\text{ref}} = \frac{Q_{\text{inv}}^{\text{ref}}}{\frac{3}{2} V_{sd2}} + \left( k_{p,Qinv} + \frac{k_{i,Qinv}}{s} \right) (Q_{\text{inv}}^{\text{ref}} - Q_{\text{inv}}) \quad (34)$$

Writing Swing's equation for alternators in ACG-2,

$$\Delta \dot{f}_2 = \frac{-1}{2H_2} (\Delta P_{L2} - \Delta P_{\text{inv}} - \rho_{AC2} \Delta f_2), \quad (35)$$

$$\Delta P_{SG2,\text{net}} = \Delta P_{L2} - \Delta P_{\text{inv}}, \quad (36)$$

where  $H_2$  is the equivalent inertia constant and  $\rho_{AC2}$  is equivalent governor droop coefficient of ACG-2.

A state space model is established by performing small signal stability analysis on the grid and the converter equations from (17) to (36). The state space model is derived as:

$$\dot{X} = AX + BU, \quad (37)$$

$$Y = CX + DU, \quad (38)$$

where  $[X]$  is state vector,  $[A]$  is state matrix,  $[B]$  is input matrix,  $[C]$  is output matrix,  $[D]$  is feedforward matrix and  $[Y]$  is output vector. The currents passing via phase reactor of the AC grids, DC voltage across the capacitors, current passing via DC link and frequency of ACG 2 are considered as states of the system. And output vector comprises of DC current and reactive power output of the rectifier, active and reactive power output of the inverter station and net active power of alternators of ACG 2. And reference values of DC voltage and reactive power at rectifier station and nominal value of frequency of ACG-2 along with reference values of active and reactive power at the inverter station controller comprises of the input vector.

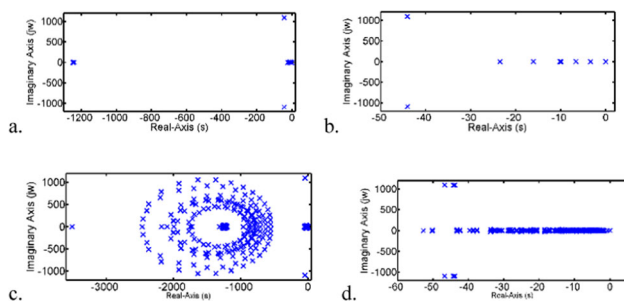
The system transfer function required for obtaining root locus plot is calculated as:

$$\frac{Y(s)}{X(s)} = C \cdot [sI - A]^{-1} B + D. \quad (39)$$

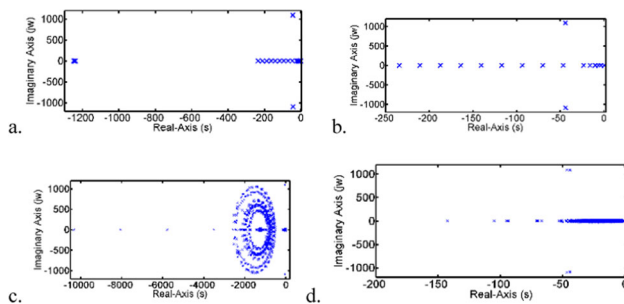
## 5.2 Effect of Varying PFC

The stability of the system is analyzed by eigenvalues of the state space matrix  $A$ . By analyzing the trajectories of the





**Fig. 6** With PFC equal to 3 pu/Hz. **a** Location of eigenvalues. **b** Enlarged version of location of eigenvalues. **c** Root locus plot. **d** Enlarged version of the root locus plot



**Fig. 7** With PFC varying from 3 to 30 pu/Hz. **a** Location of eigenvalues. **b** Enlarged version of eigenvalues location. **c** Root locus plot. **d** Enlarged version of the root locus plot

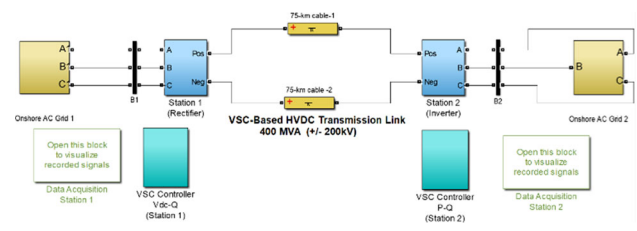
eigenvalues in response to increase in values of PFC, system stability is analyzed.

The location of the eigenvalues with PFC equal to 3 pu/Hz is shown in Fig. 6a. And Fig. 6b shows the enlarged version of this. Also, the root locus plot obtained using (39) and its enlarged version are shown in Fig. 6c, d, respectively, for these values of PFC. As observed, system eigenvalues as well as the roots of the system's transfer function remain in stable region when PFC is maintained at 3 pu/Hz.

Figure 7a shows the eigenvalue movement with increase in value of PFC from 3 to 30 pu/Hz in ten steps each of 3 pu/Hz. For this range of PFC, the root locus plot along with its zoomed version are also shown in Fig. 7c, d.

With change in value of PFC, location of all the eigenvalues except one remains constant. The movement of this eigenvalue continue to move towards negative real axis. This indicates that for the considered range of PFC, system eigenvalues continue to remain in the stable region. Similar pattern is observed in the root locus plot.

Based on the system modeling presented above, the small signal stability study indicates that for performing increase in values of PFC, the eigenvalues tend to move toward the negative real axis which results in improvement of system stability.



**Fig. 8** Test system layout

**Table 4** Parameters for proposed frequency regulation scheme

Considered parameters	Values	Calculated parameters	Values
$i_{d,upper}$	1.1 pu	$M$	1.0864
$i_{d,rated}$	1.0125 pu	$i_{vsc,upper}$	1.3602 pu
$i_{q,upper}$	0.8 pu	$i_{d,new1}$	1.3561 pu
$i_{q,rated}$	0.1055 pu	$k$	1.34
$i_{q,min}$	0.02 pu	$p$	0.104
$V_d$	0.99 pu	$d$	5.23
$K_f$	12	$i_{d,new2}$	1.3454 pu

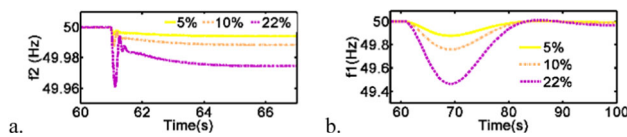
## 6 System Simulation Results

To justify fast and efficient frequency regulation offered by the proposed scheme, the network topology shown in Fig. 1 is considered and using data from Table 2, it is simulated on MATLAB/Simulink platform as shown in Fig. 8. Since all the five alternators of ACG 2 are loaded equally, thus, due to space constraints, output of G1 is only depicted in the figures following. Operations of over-frequency and under-frequency relays for load management are envisaged within frequency range of 48.5–51.5 Hz. To justify the effectiveness of the proposed technique, three cases have been considered for comparison:

- Case A: VSC-HVDC system using conventional vector current control.
- Case B: VSC-HVDC system using conventional vector current control supplemented with  $P$ - $f$  droop control.
- Case C: Proposed frequency regulation technique using VSC-HVDC system enabled with dynamic converter current modulation scheme.

Parameters like converter currents,  $d$ -axis component of voltage at PCC-2, PFC, etc., that are considered in simulation and parameters which are needed to be calculated for the proposed control scheme are listed in Table 4.

To demonstrate small, medium and large frequency excursion scenarios, different percentages of load increase are



**Fig. 9** **a** Frequency of ACG 2 and **b** frequency of ACG 1 for various scenarios

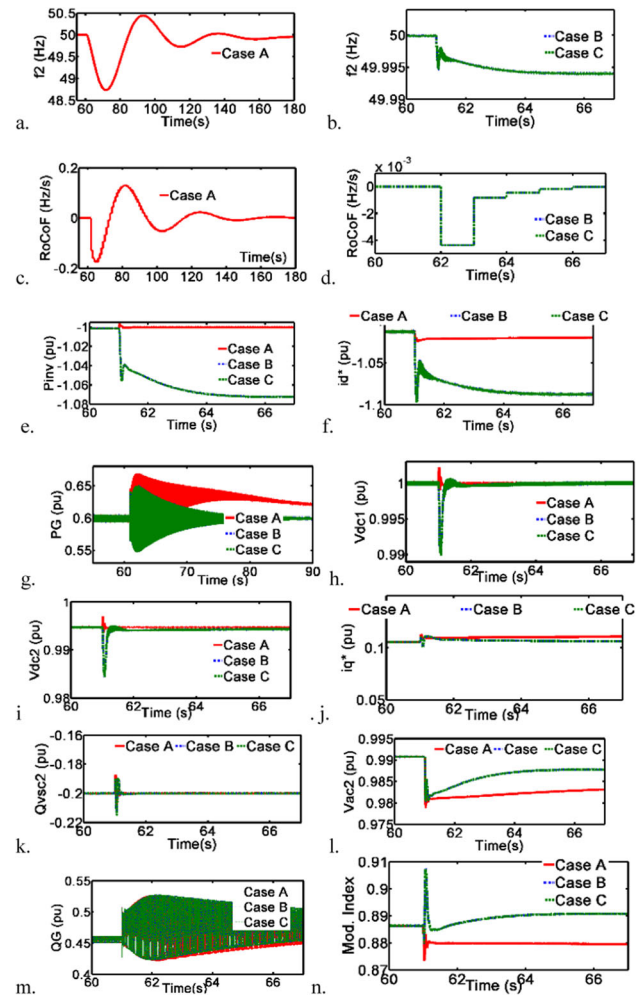
performed in ACG 2. The simulation results depicting frequency deviations (using proposed technique) in ACG 2 and ACG 1 are shown in Fig. 9 for the three scenarios. The actual values of frequency deviations of the two grids, obtained from the simulations run, are compared with calculated values obtained from the derived expressions, (7) and (11). The comparison is presented in Table 5.

It can be noted that actual values of frequency deviations obtained via simulation are more than those obtained analytically. This is because of the considered assumption of having negligible switching and line losses in the AC/DC system.

Using proposed technique,  $f_{2,nadir}$ , i.e., frequency nadir of ACG 2, as observed from Fig. 9 are given in Table 6, for the three scenarios. Also, the limiting values of frequency deviations for the three possible scenarios viz.,  $\Delta f_{2,small}$ ,  $\Delta f_{2,medium}$  and  $\Delta f_{2,large}$  (as calculated Table 4 data and using (14), (15) and (16)) are also mentioned in Table 6. Here,  $f_{2,lim}$  is calculated by adding rated ACG 2 frequency (i.e., 50 Hz) into value of  $\Delta f_{2,small}$  for small excursion scenario and likewise. It can be observed that the obtained nadir of ACG 2 is always maintained better than  $f_{2,lim}$  for all the three scenarios. Thus, frequency deviations are always limited to less than the respective predefined values from (14), (15) and (16).

## 6.1 Small Frequency Excursion

To illustrate a small frequency excursion scenario, a 5% load increase is performed in ACG 2 at  $t = 61$  s. Figure 10 shows the resultant outputs for HVDC/AC network. Before disturbance application, the entire system was operating at steady state in which ACG 2 was operating at near to its nominal frequency of 50 Hz. In this scenario, performances of Case B and Case C are exactly same. This is because the conventional  $P-f$  control has inherent frequency support capability by allowing  $i_d^*$  to take value up to  $i_{d,upper}$ . With 5% load increase, Case A undergoes serious frequency deviation below 49 Hz (Fig. 10a) as no active power support is available from the



**Fig. 10** **a** Frequency of ACG 2: Case A. **b** Frequency of ACG 2: cases B and C. **c** RoCoF of ACG 2: Case A. **d** RoCoF of ACG 2: Cases B and C. **e** Active power inverted by VSC-HVDC into ACG 2. **f** Inverter's reference direct-axis current. **g** Active power output of G1. **h** DC voltage maintained at rectifier. **i** DC voltage at inverter. **j** Quadrature-axis reference current of inverter. **k** Inverter's reactive power output. **l** AC voltage at PCC-2. **m** Reactive power output of G1. **n** Inverter's modulation index for small frequency excursion scenario

inverter while for Cases B and C, frequency of ACG 2 dips to 49.994 Hz (Fig. 10b). It should also be noted that RoCoF for cases B and C is less than 0.05 Hz/s (Fig. 10d.) in contrast to Case A where it is around 0.2 Hz/s (Fig. 10c.). Also, with absence of power support from VSC-HVDC in case A, its frequency takes longer to settle to a steady state, while quick frequency restoration is observed in the other two cases.

**Table 5** Frequency deviations in ACG 1 and ACG 2

Scenario	$\Delta P_{L2}$	$\Delta f_{1,cal}$ (Hz)	$\Delta f_{1,act}$ (Hz)	$\Delta f_{2,cal}$ (Hz)	$\Delta f_{2,act}$ (Hz)
Small	5%	– 0.005975	– 0.006065	– 0.0054	– 0.00565
Medium	10%	– 0.01154	– 0.01160	– 0.01031	– 0.011
Large	22%	– 0.02395	– 0.0256	– 0.02325	– 0.02455





**Table 6** Frequency nadir and limiting values of frequency deviations using proposed scheme

Scenario	$\Delta P_{L2}$	$f_{2,nadir}$ (Hz)	$\Delta f_{2,small/medium/large}$ (Hz)	$f_{2,lim}$ (Hz)
Small	5%	49.993	− 0.01084	49.989
Medium	10%	49.987	− 0.0425	49.9575
Large	22%	49.96	− 0.04304	49.957

Even after switching on additional 5% of system load, the system frequency drops to just 49.994 Hz for the proposed scheme and Case B. This is justified because as soon as the frequency deviation occurs, inverter releases additional power support immediately. Due to the considered magnitude of PFC (12 pu/Hz) and magnitude of applied load increase and from (7), the magnitude of frequency deviation obtained is smaller and, thus, lies in the permissible designed range only. In fact, the ability to release sufficient amount of power support instantaneously gives the proposed control an upper edge over various control techniques listed in literature.

To regulate system frequency, in cases B and C, the inverter of VSC-HVDC has released additional 7.2% active power output over its rated output of 1 pu (Fig. 10e) as compared to case A where it remains fairly constant. This has been possible due to generation of increased  $i_d^*$  (− 1.09 pu from rated value of − 1.0125 pu, Fig. 10f), in the OC loop. Due to such response from the inverter for cases B and C, active power output of the alternator, G1 of ACG 2 (Fig. 10g.), remains almost same as its pre-disturbance value (0.6 pu), thus, causing minimum frequency deviation in ACG 2. However, in Case A, the burden of providing frequency support is solely on the alternators of ACG 2 which rise their output from 0.6 to 0.65 pu. DC voltage is maintained to rated value at the rectifier station as shown in Fig. 10h. Except a transient dip in DC voltage at both the stations (Fig. 10i) at the instant of load switching, the DC voltage is maintained as constant throughout.

### 6.1.1 Effect of Dynamic Current Modulation on AC Voltage of ACG 2 During Small Excursion Scenario

In this scenario,  $i_d^*$  is not allowed to increase beyond  $i_{d,upper}$  (1.1 pu) and  $i_q^*$  kept almost same (0.106 pu) as depicted in Fig. 10f, j, respectively. As a result, reactive power support from the inverter to ACG 2 remains fairly constant (Fig. 10k). However, AC voltage at PCC-2 (Fig. 10l) suffers a dip of 0.01 pu at the switching instant which recovers quickly afterward. A small difference between pre-disturbance value (0.99 pu) and post-disturbance value (0.9865 pu) of AC voltage is also noted. This is justified because of switching load, additional 25 MVar are demanded by it while only 0.02 pu, i.e., net 20 MVar are supplied by all the five alternators of ACG 2 (Fig. 10m). The modulation index at inverter station is shown

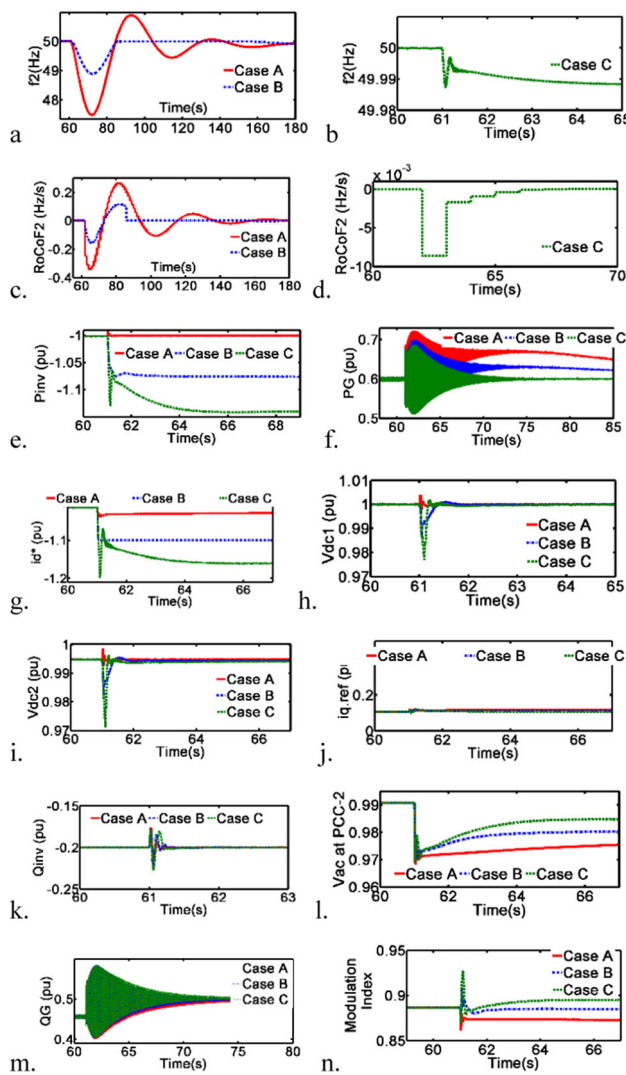
in Fig. 10n, which depicts that with the proposed scheme, its value is better than Case A.

## 6.2 Medium Frequency Excursion

A medium frequency excursion scenario is exhibited by initiating a load increase of 10%, i.e., 80 MW, 0.85 pf into ACG 2 at  $t = 61$  s. Before switching on this load, the frequency of ACG 2 was 50 Hz. Frequency of ACG 2 for cases A and B is shown in Fig. 11a. It is evident that this switching would result in tripping of under-frequency relays in ACG 2 when VSC-HVDC system is controlled with Case A. Using Case B also results in serious drop in system frequency below 49 Hz. In contrast when the proposed frequency regulation scheme is utilized as in Case C, system frequency (Fig. 11b) only drops to bit less than 49.99 Hz and settles quickly to a steady state value in about 5 s. In contrast, 20 s are taken in Case B for frequency to settle. Without having any active power support from VSC-HVDC in Case A, it takes longer and frequency keeps on oscillating around the nominal value for more than 100 s. The quick frequency restoration achieved by proposed scheme is also justified via RoCoF of ACG 2 shown in Fig. 11d which is less than 0.01 Hz/s in contrast to Case A having more than 0.3 Hz/s and for Case B having around 0.18 Hz/s as shown in Fig. 11c.

The inverter's output is shown in Fig. 11e. With proposed control, inverter is able to supply additional power output of 13.5% over its rated output, while for Case B, only 7.2% of additional output is inverted into ACG 2. This enhanced active support by the proposed control has been viable due to its ability to modulate its  $i_d^*$  to − 1.2 pu (Fig. 11g.) which is more than  $i_{d,upper}$  value of − 1.1 pu. However, for Case B, as  $i_d^*$  could not exceed beyond  $i_{d,upper}$ , thereby limiting its active power support capability. Without any frequency support from HVDC in Case A, its inverter's output and  $i_d^*$  are almost constant.

With limited support from the inverter in Case B and no support in Case A and in order to cater the additional load demand, G1 and other alternators of ACG 2 have increased their active power output from 0.6 to 0.62 pu in Case B and to 0.67 pu in Case A (Fig. 11f). While for proposed Case C, due to ample support available from VSC-HVDC via dynamic current modulation, active power output of alternators almost remains same. DC voltage maintained at rectifier and inverter stations is shown in Fig. 11h, i, respectively. For case C, the



**Fig. 11** **a** Frequency of ACG 2: Case A. **b** Frequency of ACG 2: cases B and C. **c** RoCoF of ACG 2: Case A. **d** RoCoF of ACG 2: Cases B and C. **e** Active power inverted by VSC-HVDC into ACG 2. **f** Inverter's reference direct-axis current. **g** Active power output of G1. **h** DC voltage maintained at rectifier. **i** DC voltage at inverter. **j** Quadrature-axis reference current of inverter. **k** Inverter's reactive power output. **l** AC voltage at PCC-2. **m** Reactive power output of G1. **n** Inverter's modulation index for medium frequency excursion scenario

DC voltage deviation is deeper comparatively but only at the instant of switching load and recovers quickly afterward.

### 6.2.1 Effect of Dynamic Current Modulation on AC Voltage of ACG 2 in Medium Excursion Scenario

As explained in Sect. 4 and as shown in Fig. 11j, value of  $i_q^*$  does not change much in this scenario. As a result, the reactive power output of inverter prior and after performing the current modulation remains unchanged (Fig. 11k). However, for all the three cases, AC voltage of system at PCC 2 drops nearly to 0.97 pu from prior value of 0.99 pu

(Fig. 11l). This is due to additional reactive power demand of 49.6 MVar by the switched load, in response to which only 0.04 pu that is 40 MVar additionally are compensated by all the five alternators of ACG 2 as shown in Fig. 11m. Nevertheless, AC voltage profile is better for proposed control with respect to both the cases A and B. It should be noted that dip in AC voltage is more in comparison to the small frequency excursion scenario which is because of the increased reactive power demand made by the switched load in comparison to the prior scenario. The modulation index for the inverter station is also shown in Fig. 11n having highest value for Case C and least for Case A.

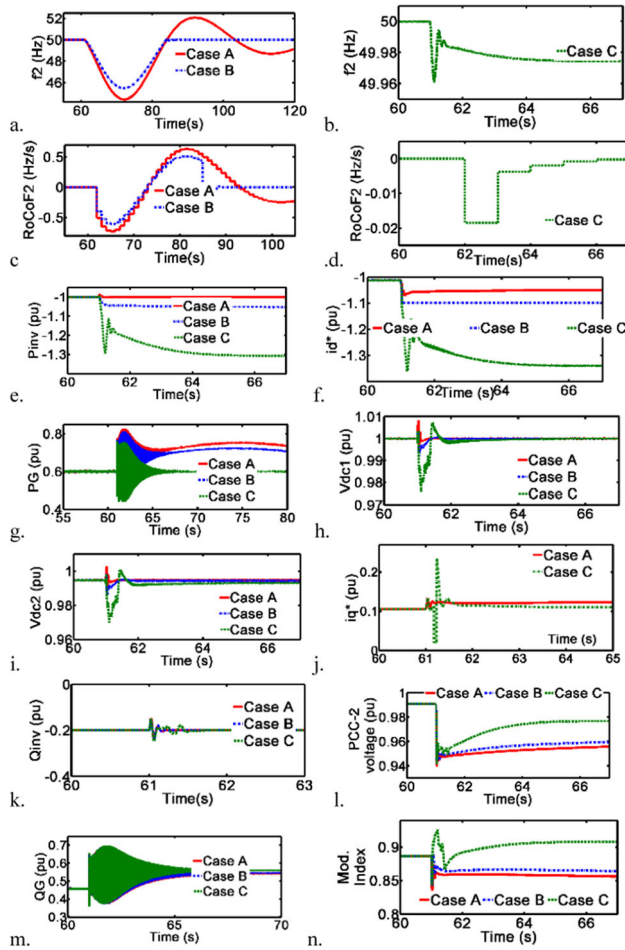
## 6.3 Large Frequency Excursion

To depict system performance for a large frequency excursion, a load increase of 22% is performed in ACG 2. Frequency of ACG 2 with cases A and B is shown in Fig. 12a and for Case C in Fig. 12b, respectively. Due to sudden load increase of a high value, cases A and B suffer serious frequency excursions. While with the proposed scheme used in Case C, the frequency nadir observed is not just of a very nominal value of 49.96 Hz but also settles to steady state within five seconds. In addition, as shown in Fig. 12d, its RoCoF value is better in comparison to its competitors. Active power output of inverter is depicted in Fig. 12e. The quick frequency restoration with proposed control has been possible due to ability of its inverter to provide additional active power support up to 30% above its rated output of 1 pu. Such significant active power support offered by inverter has only been feasible because it has modulated its  $i_d^*$  to take value equal to  $-1.36$  pu which is much near to  $i_{vsc,upper}$ , as shown in Fig. 12f, while in Case B,  $i_d^*$  still could not be increased beyond  $i_{d,upper}$  despite of such increased load demand and large frequency deviation. Thus, in Case A and B, recovery of system frequency is done by additional support from G1 and other four alternators which increase their output from 0.6 to 0.7 pu and 0.69 pu, respectively, as shown in Fig. 12g, whereas the additional active power output of inverter, for Case B, is even less than 10% above the rated, thus, causing increased frequency deviation in it. DC voltage maintained at rectifier and inverter ends are shown in Fig. 12h, i, respectively.

### 6.3.1 Effect of Dynamic Current Modulation on AC voltage of ACG 2 in Large Excursion

With proposed control, while  $i_q^*$  gets reduced to minimum value of  $i_{q,min}$  (0.02 pu) as shown in Fig. 12j so that simultaneously,  $i_d^*$  can be increased to  $i_{d,new,2}$ , i.e.,  $-1.36$  pu (Fig. 12f). Due to reduction of  $i_q^*$ , inverter's reactive power support gets compromised (Fig. 12k). However, this modulation last for less than one second and as soon as frequency





**Fig. 12** **a** Frequency of ACG 2: Case A. **b** Frequency of ACG 2: cases B and C. **c** RoCoF of ACG 2: Case A. **d** RoCoF of ACG 2: Cases B and C. **e** Active power inverted by VSC-HVDC into ACG 2. **f** Inverter's reference direct-axis current. **g** Active power output of G1. **h** DC voltage maintained at rectifier. **i** DC voltage at inverter. **j** Quadrature-axis reference current of inverter. **k** Inverter's reactive power output. **l** AC voltage at PCC-2. **m** Reactive power output of G1. **n** Inverter's modulation index for large frequency excursion scenario

of ACG 2 starts recovering,  $i_q^*$  regains its pre-disturbance value and hence, initial reactive power support offered by VSC-HVDC system gets restored.

Also, AC voltage at PCC-2 (Fig. 12l) drops to 0.94 pu from 0.99 pu but quickly starts recovering within a period of one second. This drop in AC voltage is due to additional reactive power demand of 111.6 MVar by the switched load. In response, net 100 MVar are injected by alternators of ACG 2 additionally (Fig. 12m). Even though reactive power support compromise, though for a short period, is made in proposed scheme, still its AC voltage recovery is faster and better with respect to cases B and A.

## 7 Conclusion

In this paper, an efficient and dynamic control scheme has been proposed for quickly regulating frequency of the interconnected AC grid using VSC-HVDC system. In fact, it is shown that by exploiting primary frequency reserves available with supporting grid at the other end and by proper utilization of converter's permissible operating range via proposed technique, disturbed grid's frequency can be limited within pre-defined values. It is shown that for load/generation unbalances up to 22%, VSC-HVDC can invert additional 30% power above its rated power into disturbed grid. Its ability to supply additional power when already carrying rated power so as to mitigate frequency excursion quickly without any prerequisites or causing time delays imparts it an upper edge over the listed control techniques. Nevertheless, converter's reactive power support gets compromised for a very short period but still better AC voltage profile is obtained in comparison to conventional control while maintaining frequency stability throughout. The proposed scheme is tested by introducing various credible load/generation unbalances in a IEEE14-bus system modelled on MATLAB/Simulink platform. It is shown that irrespective of severity of the disturbance, system frequency is restored quickly within an acceptable range. Additionally, load–frequency dynamics of both disturbed and supporting grids at the two ends of VSC-HVDC system is demonstrated and verified analytically as well as via simulations. Also, system stability analysis by tracking trajectories of eigenvalues in response to increase in values of power frequency coefficient has also been analyzed.

## References

1. Taneja, A.; Saha, R.; Singh, M.: A case study on VSC-HVDC converter outage using current modulation approach in AC-DC grids to restore frequency stability. *International Conference for Advancement in Technology (ICONAT)* **2022**, 1–6 (2022). <https://doi.org/10.1109/ICONAT53423.2022.9725902>
2. R. Saha, M. Singh and A. Taneja, "An Adaptive Master-Slave Technique using VSC current Modulation in VSC-based MTDC System," 2021 6th International Conference for Convergence in Technology (I2CT), 2021, pp. 1–8, doi: <https://doi.org/10.1109/I2CT51068.2021.9418191>.
3. A. Taneja, R. Saha and M. Singh, "Frequency Regulation Technique in AC-DC Grid using Converter Current Modulation in VSC-HVDC System," 2020 IEEE 17th India Council International Conference (INDICON), 2020, pp. 1–8.
4. C.E. Spallarossa, Y. Pipelzadeh and T. C. Green, "Influence of frequency-droop supplementary control on disturbance propagation through VSC HVDC links", IEEE Power and Energy Society General Meeting, 2013, Vancouver, Canada, 21–25 Jul, pp. 1–5.
5. Adeuyi, O.D.; Cheah-Mane, M.; Liang, J.; Jenkins, N.: Fast Frequency Response From Offshore Multiterminal VSC-HVDC Schemes. *IEEE Trans. Power Delivery* **32**(6), 2442–2452 (2017)





6. L. Orellana, V. Matilla, S. Wang, O. D. Adeuyi and C. E. Ugalde-Loo, "Fast frequency support control in the GB power system using VSC-HVDC technology," 2017 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), 2017, pp. 1–6.
7. Sanz, M.; Judge, P.D.; Spallarossa, C.E.; Chaudhuri, B.; Green, T.C.: Dynamic Overload Capability of VSC HVDC Interconnections for Frequency Support. *IEEE Trans. Energy Convers.* **32**(4), 1544–1553 (2017)
8. Sun, K.; Xiao, H.; Liu, S.; Liu, Y.: Machine learning-based fast frequency response control for a VSC-HVDC system. *CSEE Journal of Power and Energy Systems* **7**(4), 688–697 (2021). <https://doi.org/10.17775/CSEEJPES.2020.01410>
9. Zhu, J.; Booth, C.D.; Adam, G.P.; Roscoe, A.J.; Bright, C.G.: Inertia Emulation Control Strategy for VSC-HVDC Transmission Systems. *IEEE Trans. Power Syst.* **28**(2), 1277–1581 (2013)
10. Shen, Z., et al.: Variable-Inertia Emulation Control Scheme for VSC-HVDC Transmission Systems. *IEEE Trans. Power Syst.* **37**(1), 629–639 (2022). <https://doi.org/10.1109/TPWRS.2021.3088259>
11. Zhu, J., et al.: Inertia Emulation and Fast Frequency-Droop Control Strategy of a Point-to-Point VSC-HVdc Transmission System for Asynchronous Grid Interconnection. *IEEE Trans. Power Electron.* **37**(6), 6530–6543 (2022). <https://doi.org/10.1109/TPEL.2021.3139960>
12. Guan, M., et al.: The Frequency Regulation Scheme of Interconnected Grids With VSC-HVDC Links. *IEEE Trans. Power Syst.* **32**(2), 864–872 (2017). <https://doi.org/10.1109/TPWRS.2015.2500619>
13. Guan, M.; Pan, W.; Zhang, J.; Hao, Q.; Cheng, J.; Zheng, X.: Synchronous Generator Emulation Control Strategy for Voltage Source Converter (VSC) Stations. *IEEE Trans. Power Syst.* **30**(6), 3093–3101 (2015). <https://doi.org/10.1109/TPWRS.2014.2384498>
14. Phulpin, Y.: Communication-Free Inertia and Frequency Control for Wind Generators Connected by an HVDC-Link. *IEEE Trans. Power Syst.* **27**(2), 1136–1137 (2012)
15. Kirakosyan, A.; El-Saadany, E.F.; Moursi, M.S.E.; Salama, M.M.A.: Selective Frequency Support Approach for MTDC Systems Integrating Wind Generation. *IEEE Trans. Power Syst.* **36**(1), 366–378 (2021)
16. C. Du and E. Agneholm, "Investigation of Frequency/AC Voltage Control for Inverter Station of VSC-HVDC," IECON 2006 - 32nd Annual Conference on IEEE Industrial Electronics, 2006, pp. 1810–1815.
17. Alshammari, K.; Alrajhi, H.; El-Shatshat, R.: Optimal Power Flow for Hybrid AC/MTDC Systems. *Arab J Sci Eng* **47**, 2977–2986 (2022). <https://doi.org/10.1007/s13369-021-05983-z>
18. Liu, Y.; Chen, Z.: A Flexible Control Method of VSC-HVDC Link for Enhancement of Effective Short-Circuit Ratio in a Hybrid Multi-Infeed HVDC System. *IEEE Trans. Power Syst.* **28**(2), 1568–1581 (2013)
19. Li, Z.; Wei, Z.; Zhan, R.; Li, Y.; Tang, Y.; Zhang, X.-P.: Frequency Support Control Method for Interconnected Power Systems Using VSC-MTDC. *IEEE Trans. Power Syst.* **36**(3), 2304–2313 (2021). <https://doi.org/10.1109/TPWRS.2020.3026035>
20. Kabsha, M.M.; Rather, Z.H.: A New Control Scheme for Fast Frequency Support From HVDC Connected Offshore Wind Farm in Low-Inertia System. *IEEE Transactions on Sustainable Energy* **11**(3), 1829–1837 (2020)
21. Lu, Z.; Ye, Y.; Qiao, Y.: An Adaptive Frequency Regulation Method With Grid-Friendly Restoration for VSC-HVDC Integrated Offshore Wind Farms. *IEEE Trans. Power Syst.* **34**(5), 3582–3593 (2019)
22. Wen, Y.; Zhan, J.; Chung, C.Y.; Li, W.: Frequency Stability Enhancement of Integrated AC/VSC-MTDC Systems With Massive Infeed of Offshore Wind Generation. *IEEE Trans. Power Syst.* **33**(5), 5135–5146 (2018)
23. Zhang, L.; Harnefors, L.; Nee, H.P.: Modeling and Control of VSC-HVDC Links Connected to Island Systems. *IEEE Trans. Power Syst.* **26**(2), 783–793 (2011)
24. Sanz, I.M.; Judge, P.D.; Spallarossa, C.E.; Chaudhuri, B.; Green, T.C.; Strbac, G.: "Effective damping support through VSC-HVDC links with short-term overload capability,"; IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe). Turin, Italy **2017**, 1–6 (2017). <https://doi.org/10.1109/ISGTEurope.2017.8260280>
25. S. Norrga, L. A. ngquist, and K. Ilves, "Operating region extension for multilevel converters in HVDC applications by optimisation methods, in AC and DC Power Transmission (ACDC 2012), 10th IET International Conference on, Dec 2012, pp. 1–6
26. Recommended Practice for Excitation System Models for Power System Stability Studies", IEEE Standard 421.5–1992, August 1992.
27. IEEE Working Group on Prime Mover and Energy Supply Models for System's Dynamic Performance Studies, "Hydraulic Turbine and Turbine Control Models for Dynamic Studies", IEEE Transactions on Power Systems, Vol. 7, No. 1, February 1992, pp.167–179.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



# A High PFC Integrated AC-DC Circuit With Inherent Lossless LCD Snubber

Vinod Kumar Yadav, *Member, IEEE*

**Abstract**—This research presents an isolated single-stage-single-switch integrated LED driver with high efficiency, high input power factor (IPF), and in compliance to the stringent IEC61000-3-2 and ENERGY STAR standards. The proposed integration combines the power factor correction (PFC) SEPIC and power control (PC) flyback converters. The flyback converter in the PC stage operates in DCM to achieve a higher IPF. The PFC stage consists of an inductor-capacitor-diode (LCD) which together functions as an inherent lossless snubber. The LCD clamps the peak voltage spike of the switch to a lower voltage during the switching transient. In addition the inherent LCD snubber recycles the leakage inductance energy which is shared between the PFC stage and the load. The inherent snubber operation reduces the extra cost and space that would incur because of an additional auxiliary snubber circuit. Unlike a conventional AC-DC system with a snubber, the driver works in a non-resonant mode. The proposed system's operating principles, mathematical analysis, and theoretical conclusion are discussed. The simulation results are verified by using a 60W/65V hardware prototype. The proposed topology achieve an efficiency of 91.8% with an IPF of 0.995 with less than 5% input current harmonics.

**Index Terms**—Integrated, light emitting diode, LCD snubber, power factor correction (PFC), flyback, SEPIC, single stage.

## I. INTRODUCTION

Light Emitting Diode (LED) history is over a century old, and the development of the blue LED in 1994 by the Nichia company opened the door for its vast applications for commercial and practical purposes. LEDs applications are not just restricted to illumination, it is nowadays used to control the light with WiFi/GPRS/Zigbee/IoT/Bluetooth, it is used to transfer data, and in some circumstances it helps plants to grow and create carbon credit. It is also used in healthcare facilities and can detect motion. When compared to other old lighting methods, LED has several advantages, including a long life cycle, a wide colour range, better thermal capability and control, better resistance to mechanical shock, small sizes, packages and cost, and strict adherence to RoHS regulations [1]-[3].

The only disadvantage with LEDs is that they cannot be directly linked to the power supply. Thus, it is essential to drive the LED through a current-controlled power utility (LED driver circuit). The LED driver's goal is not only to drive the LED but also to meet all of the prerequisite stringent criteria for its safe and cost-efficient operation. While working with a LED load, it is fundamental to meet the IEC61000-3-2 class C standard for harmonic content in input current with IPF as per the Energy Star program's minimum requirement [4]. Usually, a LED driver circuit consists of a rectifier connected to a power factor correction (PFC) circuit in order to improve the power quality in terms of harmonic content and power factor at the

supply side and a power control (PC) DC-DC converter to regulate the load power and obtain the desired power quality at the load side. Generally, a single-stage LED driver does not meet the prerequisite power quality criteria. A two-stage LED driver is best suited to meet the desired power quality standards such as high IPF, low total harmonic distortion (THD), low crest factor and a regulated output. However, a two-stage driver is not ideal for low-power applications since it is bulky, expensive and complicated due to the presence of multiple switches [5]. An integrated stage LED driver is used to address this size and cost issue. In an integrated stage LED driver system, the PFC stage and the PC stage share a common switch; they work the same way as a two-stage converter, but with only one switch. As a result, switching losses are reduced, and with a minimum switch driving circuit, it is simple and cost-effective exhibiting the power quality characteristics of a two-stage LED driver while preserving the benefits of a single-stage LED driver [6].

In switch-mode operation for the integrated LED driver circuits, the switches share both the PFC and PC stage and are thus subjected to over-current and over-voltage stress, resulting in higher switching losses. Among the many integrated LED drivers, flyback converters are the most widely used. In general, it is observed that the leakage inductance energy trapped in them results in a significant voltage spike at the switch. Therefore, a turn-off snubber is inevitable to limit the spikes and overcome the EMI issue as well. With an aim to reduce the size, weight and increase the power density, an increase in switching frequency also aggravated the above shortcomings.

An RCD snubber reduces the turn-off switching losses and switching voltage spike [8]. However, due to the energy dissipation in the snubber resistor, it fails to recycle all of the leakage energy back to the supply. Additionally, these snubber circuits have large circulating currents, resulting in increased power dissipation in snubber components. The RCD snubber recovers the leakage energy and redirects it to the input power supply. This process delays the flyback secondary side's conduction time and also increases the flyback circuit's rms current. An active snubber in flyback integrated topologies is a better solution to this issue [9]. However, this comes at a cost of additional switching elements and a complex algorithm for the controller [11]-[12]. An active LC-based snubber is reported in [13]; however, this topology's working is complex and costly due to additional switches. An alternate approach to active clamp the leakage energy is given in [14], where two flyback converters are cascaded with 180° phase shift; the suggested circuit achieves a high efficiency under ZVS and recycles the energy from the leakage flux but loses the

galvanic isolation. Another cascaded approach of two flyback circuits to mitigate the leakage energy issue mentioned in [15], requires two switches and thus topology becomes bulky. Two cross-coupled inductor-capacitor-diode (LCD) snubbers where the primary flyback is linked to the secondary snubber and vice versa is reported in [16] to address the leakage energy issue; however, it adds a level of complexity to the topology and control. A simple passive lossless turn-off snubber with an attempt to achieve high power conversion efficiency at a low-cost with a design to minimize switching losses is presented in [17], but the attained efficiency is low. In [18], an inherent lossless snubber circuit based single-stage-single-switch AC-DC-LED driver topology is proposed. However the achieved IPF and input current THD are less appreciable. Considering all of the above-mentioned efforts, and to resolve the prevailing issues of deteriorated power quality in isolated integrated flyback LED driver circuits, there is a necessity to develop an inherent snubber-based high-power quality integrated LED drivers with lower component counts and reduced control complicity.

On this line, this paper presents a single-stage-single-switch AC-DC system for LEDs where the PFC SEPIC and a flyback PC stage are integrated, operating under the discontinuous condition mode (DCM). The proposed integration results in the formation of an inherent LCD snubber that recycles the energy of the leakage inductance and shares between the PFC stage and the load. The leakage energy absorbed by the PFC stage ensures a constant voltage snubber operation; this clamps the switch's voltage, reducing power loss and making the system more efficient. This paper is arranged as follows: Section II introduces the integrated LED driver circuit and depicts the proposed integrated driver's operating principles, as well as the converter analysis. Section III discusses the design process and mathematical considerations, whereas Section IV brings out the discussion on the snubber circuit. Section V includes the simulation and experimental outcomes. Finally, Section VI concludes the proposed work.

## II. ANALYSIS OF THE PROPOSED INTEGRATED PFC TOPOLOGY

Fig. 1(a) and Fig. 1(b) portrays the circuit diagram of the proposed LED driver and its simplified version respectively. The combine operation of  $S_1$  and  $S_2$  is achieved through a single switch  $S_w$  resulting in reduced gate driver requirement and control complexity. The proposed integration of SEPIC and the flyback converter performs the function of both PFC and PC. Structurally, the circuit comprises of slow recovery bridge rectifier diodes ( $D_1 - D_4$ ) with  $L_{in}$  and  $C_{in}$  as filter components,  $L_1$  and  $L_2$  as energy-storage inductors,  $D_5$ ,  $D_6$ ,  $D_7$  and  $D_o$  are intermediate and load side diodes,  $C_{Link}$ ,  $C_1$ ,  $C_2$  and  $C_o$  are DC offset, intermediate and load side capacitors respectively. The inherent lossless LCD snubber is composed of  $L_2$ ,  $C_2$ , and  $D_6$ . As seen in Fig. 1(b), the capacitors  $C_1$  and  $C_2$  clamps the undesired voltage spike of the flyback winding during turn-off. Further the arrangement of  $L_2$ ,  $C_2$ , and  $D_6$  forms an inherent lossless LCD snubber working in a non-resonant mode, thereby, the need for an additional snubber

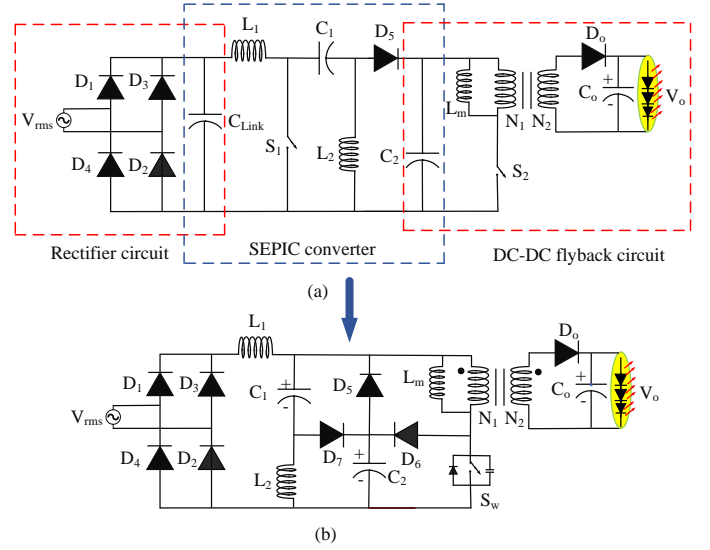


Fig. 1. (a) The proposed LED driver topology. (b) Its' simplified integrated version.

is eliminated. The proposed LED driver circuit's steady-state theoretical and operating modes are given in Fig. 2 and Fig. 3 respectively. The following assumptions are used to simplify the steady-state analysis: All the diodes and switches are ideal and without any parasitic components, inductors and capacitors are ideal. Leakage inductance  $L_{lk}$  which is much smaller than magnetizing inductance  $L_m$  are added to an ideal transformer to represent a non-ideal transformer with a turn ratio  $n = N_p/N_s$ . The output capacitor is large enough to maintain a constant output. The switching frequency is much higher than the line frequency, hence input to the PFC stage is assumed a constant. The capacitors  $C_1$  and  $C_2$  are initially charged and their voltages ( $V_{C1}$ ,  $V_{C2}$ ) are constant and sufficient enough to reset the flyback transformer. Following eight modes describe the complete working of the proposed topology.

**Mode 1 [ $t_1$ - $t_2$ ]:** This mode begins with switch ( $S_w$ ) turn-on, the voltages  $V_{DC}$ ,  $V_{C1}$  and  $V_{C2}$  are applied to inductor  $L_1$ ,  $L_2$  and  $L_m$ , and their currents increase linearly. During this mode, diode  $D_6$  and  $D_7$  are reverse biased and withstanding a voltage of  $V_{C2}$  and  $V_{C1} + V_{C2} - V_{DC}$ . The polarity of potential across  $C_1$  and  $C_2$  turns on the diode  $D_5$ . The duration of this mode is short and ends with the current reversal in  $L_2$ .

**Mode 2 [ $t_2$ - $t_3$ ]:** During this mode, currents in inductors  $L_1$ ,  $L_2$ , and  $L_m$  reach their peak values. The load capacitor  $C_o$  and magnetizing energy from primary winding via diode  $D_o$  supplies the load. The voltage stress across  $L_m$  and  $L_{lk}$  is  $V_{DC}$ . Switch turn-off marks the end of this mode.

**Mode 3 [ $t_3$ - $t_4$ ]:** Switch turn-off decreases the instantaneous current in  $L_1$  and  $L_2$ , reverse biases the diode  $D_5$ , while  $D_7$  remains off and the diode  $D_6$  starts to conduct. Voltage across  $D_5$  and  $D_7$  clamps to  $V_{C1} - V_{C2}$  and  $V_{C1} + V_{C2} - V_{DC}$ . The total voltage across  $L_m$  and  $L_{lk}$  reduces to  $V_{DC} - V_{C2}$  and  $V_{C2}$  is the voltage across the  $S_w$ . The energy of leakage inductance  $L_{lk}$  is partially absorbed in the PFC stage by  $C_1$  and  $C_2$  and is partially transfer to the secondary side. The energy transfer maintains the a constant voltage across  $C_1$  and  $C_2$  resulting

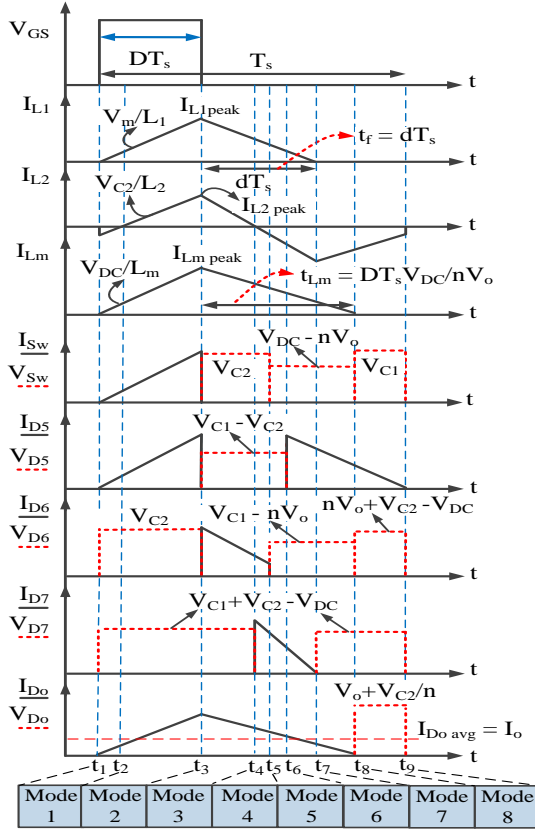


Fig. 2. Steady-state theoretical waveform of the proposed converter.

in reduced voltage stress and spike across the switch.

**Mode 4** $[t_4-t_5]$ : This state commences as the leakage inductor current ceases. The magnetizing inductance  $L_m$  discharges linearly into the output capacitance  $C_o$  through the transformer secondary, while the snubber circuit remains idle, preserving the snubber capacitor voltage  $V_{C2}$ . Here, the current  $I_{L2}$  reverses, thus the instantaneous polarity change in the voltage across  $L_2$  turns on diode  $D_7$ , while current in  $L_1$  and  $L_m$  further decreases linearly.

**Mode 5** $[t_5-t_6]$ : In this mode diode  $D_5$  and  $D_6$  are off withstanding a voltage of  $V_{C1} - V_{C2}$  and  $V_{C1} - nV_o$  respectively, while diodes  $D_7$  and  $D_o$  remains conducting.

**Mode 6** $[t_6-t_7]$ : With diodes  $D_5$  and  $D_7$  conducting, current  $I_{L2}$  reaches its negative peak. The magnetising inductance  $L_m$  of the transformer discharges through diode  $D_o$  and feeds current to the LED via diode  $D_o$ . The zero current through  $L_1$  marks the end of this mode.

**Mode 7** $[t_7-t_8]$ : This mode starts with the reverse biasing of diodes  $D_6$  and  $D_7$  withstanding a voltage  $V_{C1} - nV_o$  and  $V_{C1} + V_{C2} - V_{DC}$  respectively. While diode  $D_5$  remains conducting, the zero current through magnetizing inductor  $L_m$  marks the end of this mode.

**Mode 8** $[t_8-t_9]$ : During this mode, except for diode  $D_5$ , the rest of the semiconductor devices are off. The diode  $D_o$  withstands a voltage of  $V_o + V_{C2}/n$ , while the output capacitor  $C_o$  feed the load current. This mode lasts till the end of the switching cycle.

### III. INTEGRATED LED DRIVER CIRCUIT DESIGN CONSIDERATIONS

This section provides a complete analysis and mathematical descriptions of the proposed LED driver circuit. The input voltage is sinusoidal and given as  $V_{in}(t) = V_m \sin(w_L t)$ , where  $w_L = 2\pi f_L$ . Assuming the converter operates in DCM with respect to inductor  $L_1$ , the average source current ( $I_{in avg}$ ) over the switching period is given as,

$$I_{in avg}(t) = \frac{1}{T_s} \frac{1}{2} (DT_s + t_f) I_{L1 peak} \quad (1)$$

where the peak value inductor current  $I_{L1 peak}$  is given by,

$$I_{L1 peak} = \frac{(V_m)DT_s}{L_1} |\sin w_L t| \quad (2)$$

The fall time ( $t_f$ ) of current in  $L_1$  is estimated as follows:

$$dT_s = t_f = \frac{V_m |\sin w_L t| DT_s}{V_{C1} + V_{C2} - V_m |\sin w_L t|} \quad (3)$$

Using (2) and (3), the simplified version of (1) is,

$$I_{in avg} = \frac{V_m D^2 T_s}{2L_1} \frac{V_{C1} + V_{C2}}{V_{C1} + V_{C2} - V_m |\sin w_L t|} \quad (4)$$

From (4) it is observed that the voltage across  $C_1$  and  $C_2$  predominantly affects the nature of the input current wave-shape and hence is crucial to analyze the effect of capacitor voltages on the input current and IPF. Therefore, equation (4) is modified as,

$$i(t) = K^* \frac{V_{C1} + V_{C2}}{V_{C1} + V_{C2} - V_m |\sin w_L t|} \sin(w_L t) \quad (5)$$

As per the Fourier series expansion, (5) is expressed as:

$$i(t) = I_o + \sum_{n=1}^{\infty} I_n \sin(nw_L t + \phi_n) \quad (6)$$

In general the IPF is expressed as follows,

$$IPF = \frac{I_{1rms}}{I_{rms}} \cos \phi = \frac{P_{in}}{V_{rms} * I_{rms}} = \frac{\sqrt{(2)P_{in}}}{V_m \sqrt{\frac{1}{\pi} \int_0^\pi i^2(t) dt}} \quad (7)$$

$$I_{rms} = \sqrt{I_o^2 + \sum_{n=1}^{\infty} \frac{I_n^2}{2}} \quad (8)$$

where  $I_{1rms}$  and  $I_{rms}$  are the fundamental rms value and total rms value of  $i(t)$ ,  $\phi$  is phase deviation between input voltage  $V_{in}(t)$  and input current  $i(t)$ . The relation between total harmonic distortion (THD) and distortion factor is given below,

$$\text{Distortion factor} = \frac{I_{1rms}}{I_{rms}} \propto \frac{1}{\sqrt{1 + \text{THD}^2}} \quad (9)$$

The fundamental component ( $i_1(t)$ ) of input current  $i(t)$  is expressed as,

$$i_1(t) = \alpha_1 \cos w_L t + \beta_1 \sin w_L t \quad (10)$$

Owing to the symmetry of  $i(t)$ ,  $\alpha_1$  will be zero. With  $\cos \phi = 1$ , the coefficient  $\beta_1$  is computed using the following relations:

$$\beta_1 = \frac{2}{T_L} \int_0^{T_L} i(t) * \sin(w_L t) dt \quad (11)$$

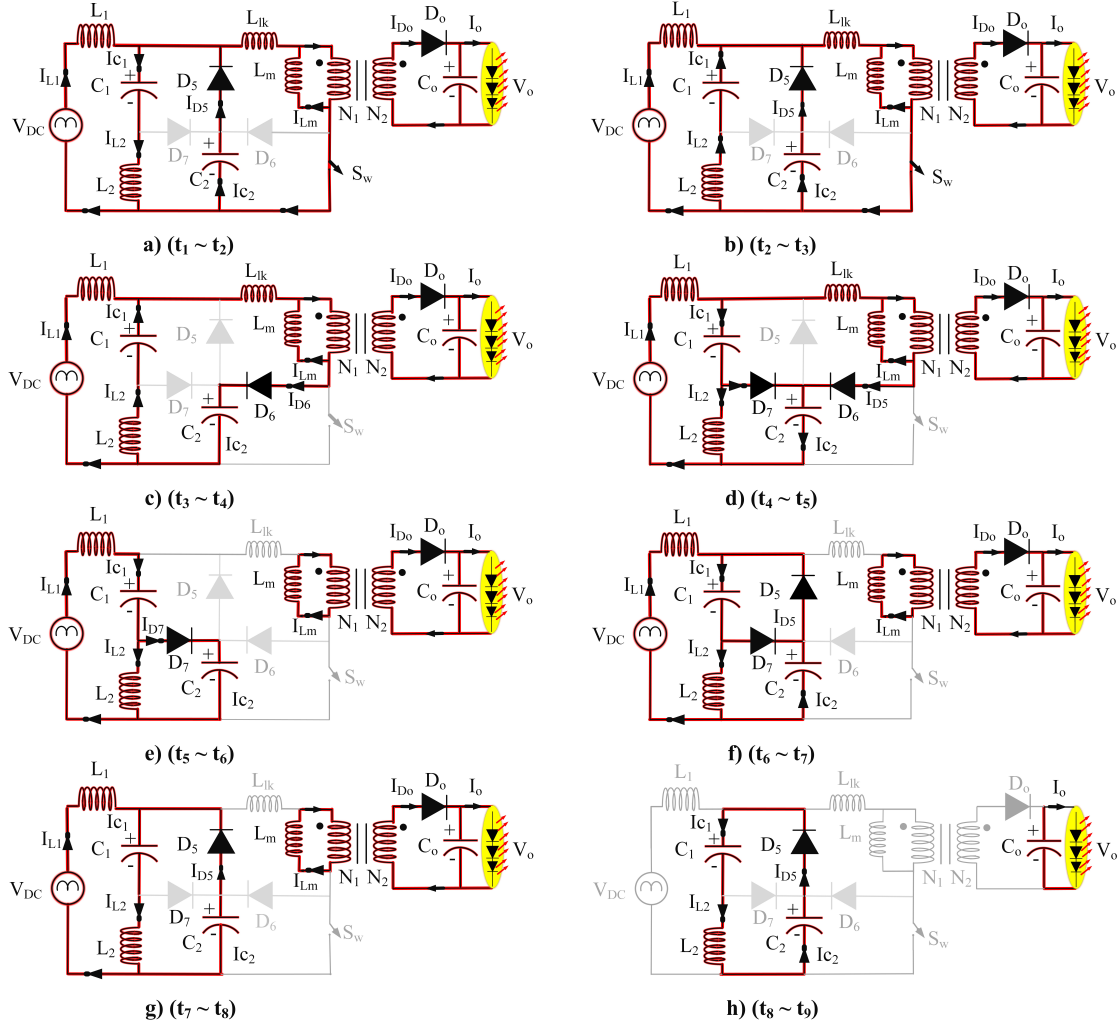


Fig. 3. Steady-state operating modes of proposed integrated PFC AC-DC system. (a) Mode-1 ( $t_1 < t \leq t_2$ ), (b) Mode-2 ( $t_2 < t \leq t_3$ ), (c) Mode-3 ( $t_3 < t \leq t_4$ ), (d) Mode-4 ( $t_4 < t \leq t_5$ ), (e) Mode-5 ( $t_5 < t \leq t_6$ ), (f) Mode-6 ( $t_6 < t \leq t_7$ ), (g) Mode-7 ( $t_7 < t \leq t_8$ ), (h) Mode-8 ( $t_8 < t \leq t_9$ ).

Substituting (5) in (11),  $\beta_1$  is expressed as:

$$\beta_1 = \frac{2K^*}{T_L} \int_0^{T_L} \frac{(V_{C1} + V_{C2}) \sin^2(w_L t)}{V_{C1} + V_{C2} - V_m |\sin w_L t|} d(w_L t) \quad (12)$$

$$\beta_1 = \frac{2K^*}{\pi} \int_0^\pi \frac{(V_{C1} + V_{C2}) \sin^2(\theta)}{V_{C1} + V_{C2} - V_m |\sin \theta|} d(\theta) \quad (13)$$

$$k = \int_0^\pi \frac{(V_{C1} + V_{C2}) \sin^2 \theta}{V_{C1} + V_{C2} - V_m |\sin \theta|} d\theta \quad (14)$$

$$\beta_1 = \frac{2K^* k}{\pi} = \frac{2W}{\pi} \quad (15)$$

According to equation (14), the constant  $k$  is function of  $V_m$ ,  $V_{C1}$  and  $V_{C2}$ , the value of  $k$  can be analyzed for different values of  $V_{C1} + V_{C2}$  for a constant  $V_m$ . Similarly the IPF can be numerically evaluated using equations (4) to (15). Further the relationships between  $k$ , IPF,  $V_m$ ,  $V_{C1}$  and  $V_{C2}$  is pictorially depicted through numerical evaluation in Fig. (4) for different values of  $V_{C1} + V_{C2}$  for a  $V_m$  of 210  $V_{rms}$ . The IPF approaches unity for large values of  $V_{C1} + V_{C2}$ ; thus, selecting  $V_{C1} + V_{C2}$  as large as possible is necessary to achieve a high IPF. However,

this results in high voltage stress on the switch, and a trade-off exists between IPF and switch voltage stress so that the  $V_{C1} + V_{C2}$  can be chosen to meet the desired IPF while minimizing switch voltage stress.

From (3), the  $t_f$  cannot exceed  $T_s/2$  and to avoid transformer the saturation, the duty cycle ( $D$ ) cannot be greater than 50%. Applying these constraints the limit on  $D$  is derived as follows;

$$dT_s = t_f = \frac{V_m |\sin w_L t| DT_s}{V_{C1} + V_{C2} - V_m |\sin w_L t|} \leq T_s/2 \quad (16)$$

$$\frac{V_m DT_s}{V_{C1} + V_{C2} - V_m} < T_s/2 \quad (17)$$

$$D < \frac{1}{2} \left( \frac{V_{C1} + V_{C2}}{V_m} - 1 \right) \quad (18)$$

For worst case scenario i.e.,  $D=100\%$  and using (18), the aggregated maximum voltage of  $C_1$  and  $C_2$  is,

$$(V_{C1} + V_{C2})_{\max} = 3V_m \quad (19)$$



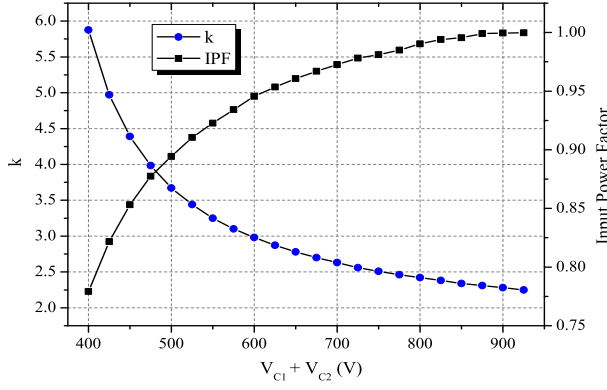


Fig. 4. Calculated values of IPF and  $k$  for different  $V_{C1} + V_{C2}$  values.

Applying the volt-sec balance for  $L_1$ , the average voltage across  $C_2$  is obtained as,

$$V_{DC}D + (V_{DC} - V_{C2})d = 0 \quad (20)$$

$$V_{DC} \left(1 + \frac{D}{d}\right) = V_{C2} \quad (21)$$

Similarly, applying volt-sec balance for  $L_2$ , the average voltage across  $C_1$  is obtained as,

$$(V_{DC} - V_{C1})D + V_{C2}d + (V_{C2} - V_{C1})(1 - (D + d)) = 0 \quad (22)$$

$$V_{C1} = \frac{V_{DC}D + V_{C2}(1 - D)}{1 - d} \quad (23)$$

Considering  $V_{DC} = 2V_m/\pi$  and using (18), (21) and (23), the relationship between the charging ( $D$ ) and discharging ( $d$ ) duty cycle for inductor  $L_1$  in DCM is evaluated as given below,

$$\text{Re} \left( (2\pi D + \pi - 2)d^2 + (4 - 2D - 2\pi D - \pi)d + 4D - 2D^2 > 0 \right) \quad (24)$$

The relationship among the  $V_{C1}$ ,  $V_{C2}$  and  $V_{C1} + V_{C2}$  for different values of  $D$  and  $d$  is shown in Fig. 5. The LED driver's input power ( $P_{in}$ ) is given by,

$$P_{in} = \frac{1}{T_L} \int_0^{T_L} V_{in}(t) * I_{inavg}(t) dt \quad (25)$$

where  $T_L = \frac{2\pi}{w_L}$  and using (4), the following relation are derived,

$$P_{in} = \frac{V_m^2 D^2 T_s}{2\pi L_1} \int_0^\pi \frac{(V_{C1} + V_{C2}) \sin^2 \theta}{V_{C1} + V_{C2} - V_m |\sin \theta|} d\theta \quad (26)$$

$$P_{in} = \frac{k V_{DC}^2 D^2 T_s}{8\pi L_1} \quad (27)$$

where

$$k = \int_0^\pi \frac{(V_{C1} + V_{C2}) \sin^2 \theta}{V_{C1} + V_{C2} - V_m |\sin \theta|} d\theta \quad (28)$$

Considering the ideal condition and neglecting the losses with  $P_{in} = P_o$ , the collateral association of inductor  $L_1$  and  $L_2$  is calculated by following equations,

$$L_1 = \frac{k V_{DC}^2 D^2 T_s}{8\pi P_o} = \frac{2.23 * 189^2 * 0.3^2}{8\pi * 92000 * 60} \quad (29)$$

The peak value of current ( $I_{L2 \text{ peak}}$ ) through  $L_2$  is given by:

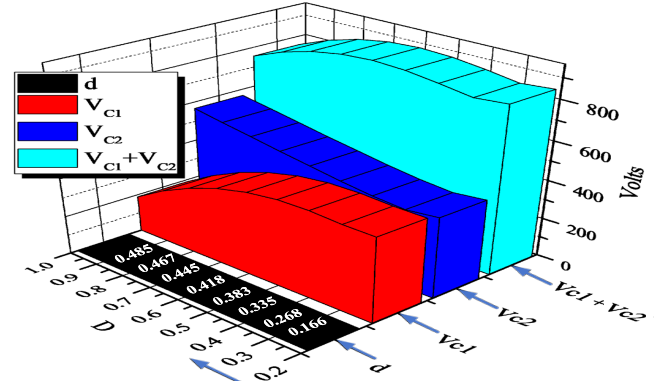


Fig. 5. Calculated values of discharging duty cycle ( $d$ ),  $V_{C1}$ ,  $V_{C2}$  and  $V_{C1} + V_{C2}$  for different charging duty cycle ( $D$ ) values.

$$I_{L2 \text{ peak}} = \frac{V_{C2} D T_s}{L_2} \quad (30)$$

Using (2), (3) and (30), the expression for the current through capacitor  $C_1$  during charging ( $I_{C1 \text{ charge}}$ ) and discharging ( $I_{C1 \text{ discharge}}$ ) is as follows,

$$I_{C1 \text{ charge}} = \frac{I_{L1 \text{ peak}} t_f}{2T_s} = \frac{V_m^2 |\sin^2 w_L t| D^2 T_s}{2L_1 (V_{C1} + V_{C2} - V_m |\sin w_L t|)} \quad (31)$$

$$I_{C1 \text{ discharge}} = \frac{I_{L2 \text{ peak}} D T_s}{2T_s} = \frac{V_{C2} D^2 T_s}{2L_2} \quad (32)$$

Applying the ampere-sec balance to  $C_1$  over the line frequency is suitable for computing  $L_2$  and is given as follows,

$$\int_0^{\pi/w_L} I_{C1 \text{ charge}} dt = \int_0^{\pi/w_L} I_{C1 \text{ discharge}} dt \quad (33)$$

Substituting (31) and (32) in (33), the following is obtained.

$$\frac{V_m^2 D^2 T_s}{2L_1} \int_0^\pi \frac{|\sin^2 w_L t|}{V_{C1} + V_{C2} - V_m |\sin w_L t|} d(w_L t) = \frac{V_{C2} D^2 T_s}{2L_2} \quad (34)$$

Further, (34) is simplified and expressed as,

$$\frac{V_m^2 D^2 T_s}{2L_1 w_L} \frac{k}{V_{C1} + V_{C2}} = \frac{2\pi V_{C2} D^2 T_s}{4L_2 w_L} \quad (35)$$

Using (35), the value of  $L_2$  is computed and given below.

$$L_2 = \frac{\pi V_{C2} (V_{C1} + V_{C2}) L_1}{2k V_m^2} = \frac{\pi * 400 * 840 * 60 * 10^{-6}}{2 * 2.23 * 297^2} \quad (36)$$

For an ideal case the output power ( $P_o$ ) is calculated as,

$$P_o = \frac{1}{2T_s} L_m I_{Lm \text{ peak}}^2 \quad (37)$$

The peak value of current ( $I_{Lm \text{ peak}}$ ) through  $L_m$  is given by,

$$I_{Lm \text{ peak}} = \frac{V_{DC} D T_s}{L_m} \quad (38)$$

Using (37) and (38), the output power is evaluated as:

$$P_o = \frac{V_{DC}^2 D^2 T_s}{2L_m} \quad (39)$$

From (37), (38) and (39), the design constraint with respect to the magnetizing inductance  $L_m$  is given by,

$$L_m > \frac{V_{DC}^2 D^2 T_s}{2P_o} = \frac{189^2 * 0.3^2}{2 * 60 * 92000} \quad (40)$$

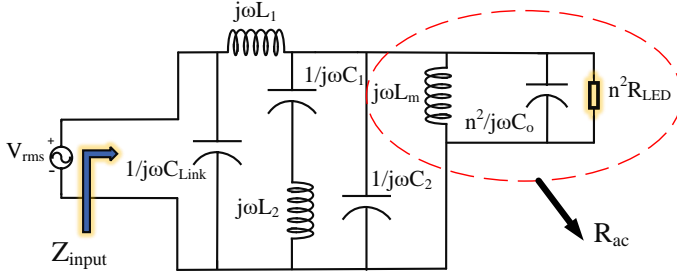


Fig. 6. Equivalent input impedance diagram in mode I

Applying volt-sec balance on magnetizing winding  $L_m$  yields in,

$$V_{DC}DT_s - t_{Lm}nV_o = 0 \quad (41)$$

Thus the fall time  $t_{Lm}$  of magnetizing winding  $L_m$  using (41) is given below,

$$t_{Lm} = \frac{V_{DC}}{nV_o}DT_s \quad (42)$$

$$t_{Lm} + DT_s < T_s \quad (43)$$

For DCM operation, the flyback converters' turn ratio must be selected such as  $I_{Lm}$  reaches to zero before the start of next cycle. Using (42) and (43), the turn ratio is given as below,

$$n > \frac{V_{DC}}{V_o} \frac{D}{1-D} > \frac{189}{65} \cdot \frac{0.3}{1-0.3} \quad (44)$$

Over the switching cycle  $T_s$ , the voltage across  $C_1$  is assumed to be constant but  $C_1$  must also follow any line frequency variation. Thus for a given resonant frequency  $f_{rs}$ , the capacitance of  $C_1$  is numerically evaluated using the below relation [19].

$$C_1 = \frac{1}{(2\pi f_{rs})^2(L_1 + L_2)} \quad (45)$$

Where  $f_{rs}$  is a resonant frequency which depends on the value of  $L_1$ ,  $L_2$  and  $C_1$ . The  $f_{rs}$  must be greater than the line frequency ( $F_L$ ) to avoid the input current oscillations. Furthermore,  $f_{rs}$  must be substantially lower than the switching frequency  $F_{sw}$  to ensure a constant voltage over a switching cycle. Since the supply time period is much higher than the switching time period, mode I operation of the proposed AC-DC system is sufficient to calculate the capacitance of  $C_2$  (the effect of input filters is disowned). As the system is assumed to operate at unity IPF, the impedance offered by the system must be resistive. Fig. 6 shows the equivalent impedance diagram for mode I, and by calculating input impedance ( $Z_{input}$ ) and equating imaginary terms to zero, the capacitance of  $C_2$  is obtained.

$$Z_{input} = \frac{1}{j\omega C_{Link}} \parallel \left( j\omega L_1 + \left( \frac{1}{j\omega C_1} + j\omega L_2 \right) \parallel \frac{1}{j\omega C_2} \parallel R_{ac} \right) \quad (46)$$

$$C_2 \equiv \frac{1}{(2\pi F_{sw})R_{ac}} \equiv \frac{1}{2\pi * 92000 * 19.84} \quad (47)$$

The output load ( $R_{LED}$ ) is used to calculate the equivalent resistance  $R_{ac}$  as given in [20].

$$R_{ac} = \frac{8n^2}{\pi^2} R_{LED} = \frac{8 * 2^2}{\pi^2} * 6.12 = 19.84\Omega \quad (48)$$

TABLE I  
KEY PARAMETERS OF THE EXPERIMENTAL PROTOTYPE

Parameters	Values
Input voltage	210V <sub>rms</sub> , 50Hz
Output voltage	65V
Output current	923mA
Output power	60W
Switching frequency $F_{sw}$	92kHz
Operating duty cycle	30%
Switch $S_w$	IRFP460
Diode ( $D_1 - D_o$ )	MURT460JX
Capacitor $C_1, C_2$	334J400/EC-1205
Magnetizing inductor $L_m$	400μH
Leakage inductance $L_{lk}$	40μH
Inductor $L_{in}, L_1, L_2$	1mH, 60μH, 160μH
Transformer turns ratio	10:5
Capacitor $C_{Link}, C_o$	470μF CD2934
LED panel	SM1944 (RDL) 6S10P 170*100mm
$R_{LED}$	6.12Ω

The output capacitor is seized in a manner similarly to a conventional DCM DC-DC flyback converter and is given below,

$$C_o = \frac{(nV_{DC}DT_s - L_m I_o)^2}{2n^2 L_m V_o \Delta V_o} = \frac{2 * 189 * 0.3 * 1.08 - 4 * 9.23}{2 * 2^2 * 40 * 65 * 0.02} \quad (49)$$

The DC-link capacitor ( $C_{Link}$ ) must supply current to the PFC stage for about a half cycle without dropping the DC bus voltage below 85%. The optimal value of the DC-link capacitor can be evaluated in terms of the desired DC-link voltage ripple ( $\Delta V_{DC}$ ), switching frequency ( $F_{sw}$ ) and DC offset voltage ( $V_{DC}$ ) is given as below.

$$C_{Link} = \frac{D^2 |V_m|^2}{8\pi^2 (2f_L) F_{sw} \Delta V_{DC} * V_{DC} L_{in}} \quad (50)$$

#### IV. SALIENT FEATURES

In this section inherent snubber operation during the turn-off period of the proposed integrated AC-DC converter is described. Unlike an additional smaller snubber capacitors, the capacitor  $C_1$  and  $C_2$  are large enough to maintain a constant voltage across the switch and thereby can handle a higher voltage spikes and also supports the PFC operations. During the switch transition from on to off, the capacitors  $C_1$ ,  $C_2$  and inductor  $L_2$  are connected in series to captures the transformer leakage energy. During which, voltage across the switch ( $V_{sw}$ ) and the magnetising inductance ( $V_{LM}$ ) using (21) and (23) is given by,

$$V_{sw} = V_{C2} = V_{DC} \left( 1 + \frac{D}{d} \right) \quad t_3 \leq t < t_5 \quad (51)$$

$$V_{LM} = V_{C1} - V_{C2} = \frac{V_{DC}(d + d^2 - D^2)}{d(1-d)} \quad (52)$$

Unlike the snubber based conventional flyback converter, the diode  $D_o$  and the energy of the magnetizing inductance is synchronized, thereby any delay in the output current due to the reversal of the secondary diode is avoided. As per [7], the

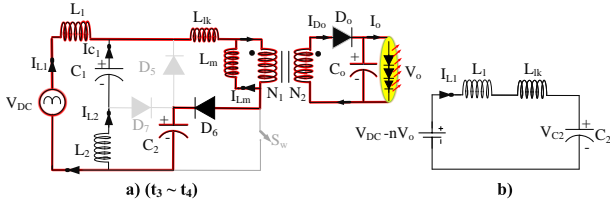


Fig. 7. a) Sub-topological state highlighting the snubber action, b) its equivalent circuit.

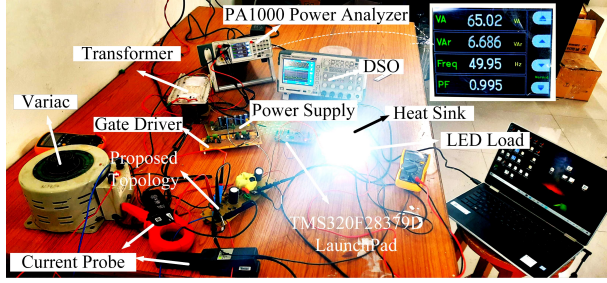


Fig. 8. The experimental setup for the proposed integrated AC-DC system.

necessary condition for a constant voltage snubber to reset the magnetizing winding  $L_m$  is given as,

$$2V_{C2} > nV_o \quad (53)$$

In view of  $V_{C2}$  fulfilling the condition mentioned above, it inherently enables a constant capacitor snubber voltage operation and thereby, eliminates the occurrence of any voltage spike. In addition, the constant voltage across  $C_2$  also ensures the swift demagnetizing of  $L_m$ . Which avoids the adverse consequences like increased current in flyback, and demand for higher-rated clamp diodes  $D_6$ . With  $D \leq 50\%$  for DCM flyback, from (53) this snubber fits the limit design. During switching transient the energy stored in the leakage inductance ( $E_{lk}$ ) and capacitor ( $\Delta E_{C2}$ ) is given by:

$$E_{lk} = \frac{L_{lk}}{2} [I_{lkmax}^2 - I_{lkmin}^2] = L_{lk} I_{sw,max}^2 \quad (54)$$

$$\Delta E_{C2} = \frac{C_2}{2} [(V_{C2} + \Delta V_{C2})^2 - V_{C2}^2]$$

If the leakage energy is flown during the turn-off, then,

$$\Delta E_{C2} = E_{lk} \quad (55)$$

Leakage inductance is undesirable and considered very small; in most cases, it is less than 10% of the magnetizing inductance. Using (55) to analyze the relationship between leakage inductance and capacitance  $C_2$ , below condition is derived,

$$L_{lk} < \frac{C_2 [(V_{C2} + \Delta V_{C2})^2 - V_{C2}^2]}{2I_{sw,max}^2} \quad (56)$$

Fig. 7(a) and (b) depict the inherent snubber operation and snubber equivalent circuit of the proposed converter respectively. The set of differential equations pertaining to the switch transition is given below to analyze the snubber capacitor voltage and leakage inductance current.

$$i_{C2} = C_2 \frac{dv_{C2}(t)}{dt} \quad (57)$$

$$v_{C2}(t) + (L_{lk} + L_1) \frac{di_{lk}(t)}{dt} - V_{DC} + nV_o = 0 \quad (58)$$

$$v_{C2}(t) + (L_{lk} + L_1) C_2 \frac{dv_{C2}^2(t)}{dt} - V_{DC} + nV_o = 0 \quad (59)$$

$$v_{C2}(t) = (v_{C2}(t_3) - V_{DC} + nV_o) \cos(w_1 t) + i_{lk}(t_3) * Z_1 \sin(w_1 t) + V_{DC} - nV_o \quad (60)$$

$$i_{lk}(t) = \frac{V_{DC} + nV_o - v_{C2}}{Z_1} \sin(w_1 t) + i_{lk}(t_3) * Z_1 \cos(w_1 t) \quad (61)$$

where  $w_1$  is the angular frequency, and  $Z_1$  is the characteristic impedance.

$$Z_1 = \sqrt{\frac{L_1 + L_{lk}}{C_2}}, \quad w_1 = \frac{1}{\sqrt{(L_1 + L_{lk})C_2}} \quad (62)$$

The amount of charge (Q) flown through capacitor  $C_2$  during switch off transient is :

$$Q = \frac{I_{lkmax}(t_5 - t_3)}{2} \quad (63)$$

Wherein, the snubbing state time period ( $t_5 - t_3$ ) follows the relation given below,

$$(t_5 - t_3) \equiv \frac{\pi}{2w_1} = \frac{\pi}{2} \sqrt{(L_1 + L_{lk})C_2} \quad (64)$$

In order to evaluate the snubber voltage stress ( $\Delta v_{C2}$ ) we apply the method given in [11]; where (60) and (61) denote the second order snubber circuits' normalized form equations. According to [11], the expression for the maximum snubber capacitor voltage during mode 3 is given below,

$$V_{C2,max} = nV_o + \sqrt{(V_{C2,min} - nV_o)^2 + (Z_1 I_{lk,max})^2} \quad (65)$$

By inspection from (65),  $\Delta V_{C2}$  is given as:

$$\Delta V_{C2} \equiv Z_1 I_{lk,max} \quad (66)$$

The maximum current across leakage inductance is maximum switch current. Further, using (54), (56) and (65) the relation of leakage inductance and inductor  $L_1$  is expressed as,

$$L_{lk} < \frac{L_{lk} + L_1}{2} \quad (67)$$

In summary, the ability of the proposed converter in dispatching the leakage energy partially to the PFC stage and the partially to the load fulfills the criteria of lossless inherent LCD snubber.

The various parameters employed for fabrication are listed in Table I while, Table II shows the comparative scaling among a few of the prevailing integrated LED drivers; the comparison is on a standard scale of the number of components counted in the circuit, their reported efficiency, input power factor, and switching stress subjected to their input voltage. As compared to [9]-[10] and [20]-[26], the proposed topology showcases an inbuilt lossless LCD snubber, which helps to protect the switch and other solid-state components from overshoot transients. It helps to eliminate the need for external snubber protection and thus reduces the cost. In general, the proposed converter provides acceptable performance with a manageable number of components, high IPF and efficiency, and lower THD.



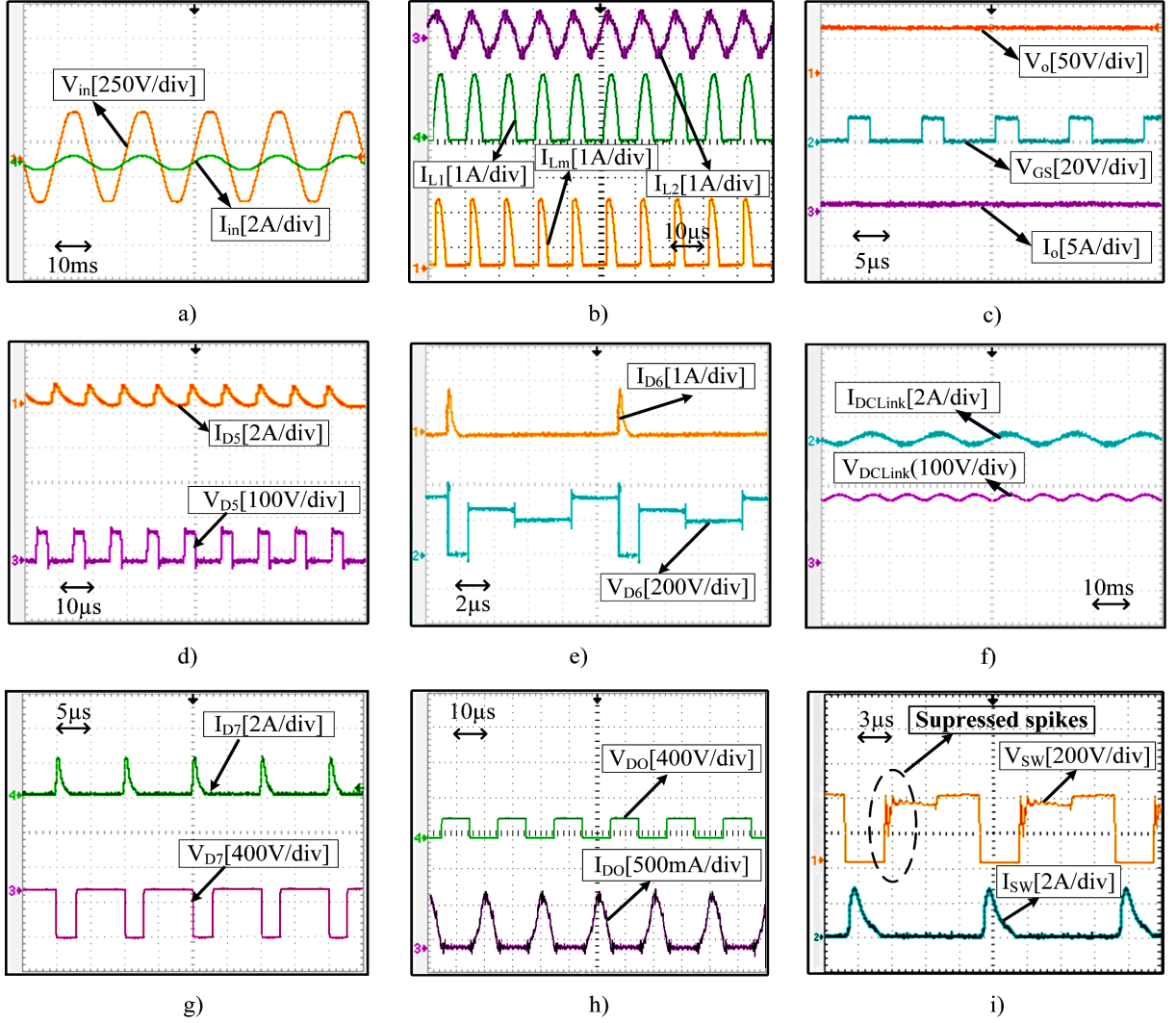


Fig. 9. Experimental waveform for (a) input voltage and current, (b) current in inductor  $L_1$ ,  $L_2$  and  $L_m$ , (c) output LED voltage, output LED current and gate to source voltage, (d) voltage and current across diode  $D_5$ , (e) voltage and current across diode  $D_6$ , (f) DC Link capacitors' voltage and current, (g) voltage and current across diode  $D_7$ , (h) voltage and current across diode  $D_o$ , (i) voltage and current across switch  $S_w$ .

## V. EXPERIMENTAL RESULTS AND DISCUSSION

To validate the theoretical analysis and conclusion of the proposed topology, a 65V/60W LED laboratory prototype feeding from a 210V/50Hz ac input is designed and implemented. The experimental bench of the test setup is shown in Fig. 8. The switching frequency is kept at 92kHz with duty ratio of 30%. The gate pulse is generated using micro-controller TMS320F28379D with CCS6.0 and the TI control suite. The sizing of inductors and capacitors is done to obtain the DCM operation as per the previously mentioned design equations and further the practically available component values were chosen. The value of inductor  $L_1$  is calculated by (29) and it is  $51.6\mu\text{H}$  we have taken it as  $60\mu\text{H}$ . From (36) the value of inductor  $L_2$  is coming  $160.98\mu\text{H}$  we have taken  $L_2$  value as  $160\mu\text{H}$ . The equation (40) gives magnetizing inductance  $L_m$  values as  $291\mu\text{H}$ , we have kept  $L_m$  as  $400\mu\text{H}$ . As per equation (44) the value of turn ratio  $n$  to support DCM must be greater than 1.24, we have kept  $n$  value to 2, which gives value of  $t_{Lm}$  from (42) as  $5.1\mu\text{sec}$ . Equation

(45) gives the value of capacitor  $C_1$  as  $0.018\mu\text{F}$  while (47) gives  $C_2$  value as  $0.0871\mu\text{F}$ , so the next available practical value of the capacitor is taken in this topology as  $0.33\mu\text{F}$ , also by inspection and to enhance the PFC stage operation both capacitor taken same ( $C_1 = C_2$ ). The output capacitor value for theoretical 2% ripple as per (49) is  $313\mu\text{F}$ , so the next practical available value is taken for  $C_o$  as  $470\mu\text{F}$ , from (50) the value of the DC link capacitor is coming to  $385\mu\text{F}$  for 15% offset ripple; thus the next available value of  $470\mu\text{F}$  is taken as  $C_{Link}$ . The experimental results for input power utility, inductors  $L_1$ ,  $L_2$ , and  $L_m$  currents, output power utility, current and the voltage across diodes  $D_5$ ,  $D_6$ ,  $D_7$ ,  $D_o$ , DC link capacitors' voltage and current, switch current, and gate to source voltage are shown in Fig. 9(a)-(i). It is seen from the result that the input current follows the footsteps of the input voltage, thus depicting a high IPF as per IEC61000-3-2 standards. The output voltage and current are well regulated, while the switch voltage stress is limited to twice the input. A Tektronix PA1000 power analyzer is used to capture all the

TABLE II  
COMPARISON WITH FEW OF THE PREVAILING INTEGRATED LED DRIVERS

Integrated Topology	Duty cycle	Diodes	Switches	Capacitors	Inductor/ Coupled inductor	Total components	IPF	THD (%)	Efficiency (%)	Switching stress <sup>1</sup>
[6]	0.3	6	1	6	3/1	17	0.997	-	90.8	Twice
[10]	-	7	1	4	2/1	15	0.95	-	90.8	Twice
[18]	0.32	8	1	5	2/1	17	0.98	25	89.3	Twice
[20]	-	8	1	5	2/2	17	0.99	5.10	90.3	Twice
[21]	0.32	8	1	5	2/1	17	0.99	-	89.3	Twice
[22]	0.7	9	1	5	2/2	18	0.995	-	91.6	Quaternary
[23]	0.5	8	2	6	5/1	22	0.99	8.0	92.8	Thrice
[24]	0.1	7	1	4	3/1	16	0.992	12.6	91.6	Thrice
[25]	-	10	1	6	2/1	20	0.99	5.20	91.2	Twice
[26]	-	9	2	5	1/1	18	-	-	91.08	Twice
Proposed converter	0.30	8	1	5	3/1	18	0.995	4.84	91.8	Twice

Note:<sup>1</sup> Switching stress is voltage at switch and it is compared with respect to the input voltage RMS value.

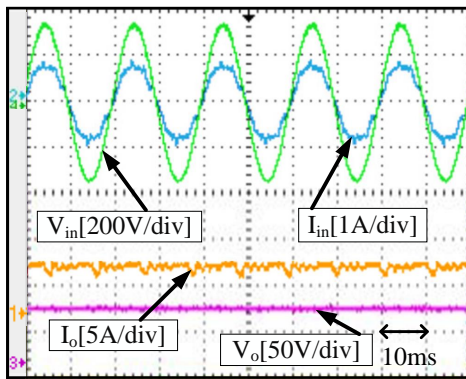


Fig. 10. Input current, Input voltage ( $210V_{rms}$ ), output current (825mA) and voltage (60V) for output load at 50W.

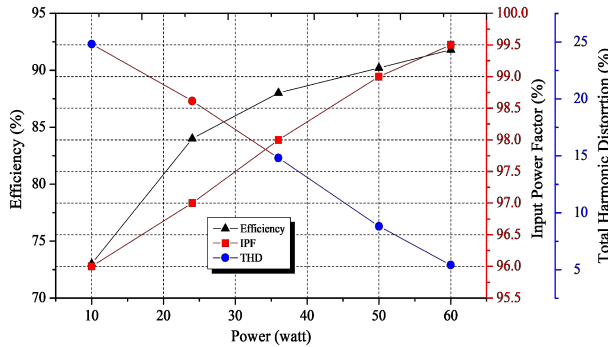


Fig. 11. Efficiency, IPF and THD at different load conditions.

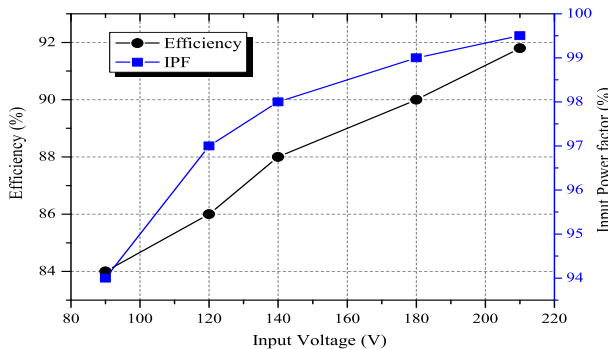


Fig. 12. Efficiency and IPF at different input voltage conditions.

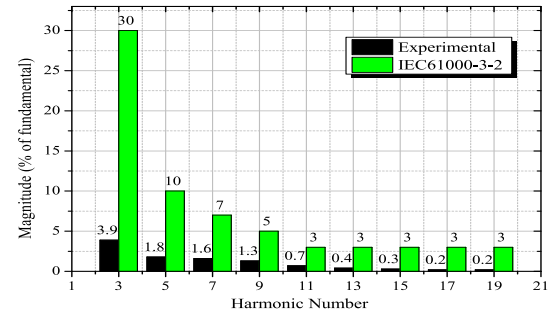


Fig. 13. Harmonics in the input current (in%) compared to the IEC61000-3-2 class-C standard.

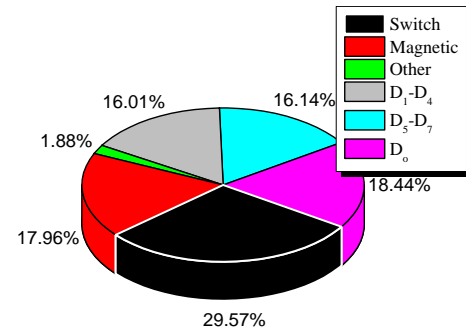


Fig. 14. Loss distribution of the proposed converter

power quality indices. The obtained input power is 66.59W and the IPF is 0.995 with an input current THD(%) of 4.84. The maximum efficiency of 91.8% and an IPF of 0.995 is achieved at the rated output power of a 60W LED module. Fig. 10 gives experimental results for input voltage, input current, output voltage, and output current for a 50 W LED load. The obtained IPF at 50W load is 0.99. Fig. 11 describes the power quality indices at different output power ranging from 16% to 100% (To draw the scaling in Fig. 11, the various loads used are OSRAM's 10W (e27), 24W (SM1820 RDL-12S2P), 36W (SM1824 RDL-6S6P), 50W (SM1823 RDL-10S5P), and 60W (SM1944 RDL-6S10P) LED modules.) with an input voltage of  $210V_{rms}$ . Fig. 12 describes the efficiency, THD and IPF curves for varying input voltage. The maximum efficiency

is achieved at input of  $210V_{rms}$ . Fig. 13 compares the measured input current harmonics against the values mentioned in IEC61000-3-2 standards, and the results are for a 60W LED load with a  $210V_{rms}$  input. Fig. 14 gives power loss distribution among the elements of the proposed converter.

## VI. CONCLUSION

An integrated single-stage-single-switch AC-DC system driving LED with high efficiency and nearly unity power factor is presented. The presence of an intermediate inductor, capacitor and diode in the PFC stage inherently works as a lossless snubber. During the switch transition, the capacitor  $C_2$  limits the rise of voltage spikes across the switch. Thus, switching losses are limited. Because of the inherent LCD snubber, some of the energy stored in leakage inductance is taken up by the PFC stage, and some is sent to the load and thereby performs as a lossless LCD snubber. As a result, the efficiency of the system increases, and the voltage across the capacitors remains constant. The experimental results are satisfactory and comply with IEC61000-3-2 and ENERGY STAR standards.

## REFERENCES

- [1] D. Salazar-Pérez, M. Ponce-Silva, J. M. Alonso, J. A. Aquí-Tapia and C. Cortés-García, "A Novel High-Power-Factor Electrolytic-Capacitorless LED Driver Based on Ripple Port," in *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 9, no. 5, pp. 6248-6258, Oct. 2021.
- [2] V. K. Yadav, M. Kumar, J. Kundu and V. Sharma, "LED: An Optimistic Solution for a Brighter Future," in *IEEE Journal of Emerging and Selected Topics in Industrial Electronics*, vol. 4, no. 1, pp. 299-308, Jan. 2023.
- [3] G. Z. Abdelmessih, J. M. Alonso, M. A. Dalla Costa, Y. -J. Chen and W. -T. Tsai, "Fully Integrated Buck and Boost Converter as a High Efficiency, High-Power-Density Off-Line LED Driver," in *IEEE Transactions on Power Electronics*, vol. 35, no. 11, pp. 12238-12251, Nov. 2020.
- [4] ENERGY STAR® Program Requirements Product Specification for Luminaires (Light Fixtures) – Eligibility Criteria, V. 2.0, Jan. 2016.
- [5] V. K. Yadav, A. Kumar Verma and U. R. Yaragatti, "Modelling and Control of Two Stage High PFC LED Driver Circuit using Average Current Control Method Driven by Vienna Rectifier," 2020 IEEE 9th Power India International Conference (PIICON), SONEPAT, India, 2020, pp. 1-6.
- [6] V. K. Yadav, A. K. Verma and U. R. Yaragatti, "An Integrated Single-Stage Single-Switch Topology With Reduced Nonlinear Components for LED," in *IEEE Journal of Emerging and Selected Topics in Industrial Electronics*, vol. 4, no. 1, pp. 317-326, Jan. 2023.
- [7] E. Dzhunusbekov and S. Orazbayev, "A New Passive Lossless Snubber," in *IEEE Transactions on Power Electronics*, vol. 36, no. 8, pp. 9263-9272, Aug. 2021.
- [8] N. M. Mukhtar and D. D. Lu, "A Bidirectional Two-Switch Flyback Converter With Cross-Coupled LCD Snubbers for Minimizing Circulating Current," in *IEEE Transactions on Industrial Electronics*, vol. 66, no. 8, pp. 5948-5957, Aug. 2019.
- [9] G. Y. Jeong and S. Kwon "Improved Single-Stage AC-DC LED-Drive Flyback Converter using the Transformer-Coupled Lossless Snubber" in *Journal of Electrical Engineering and Technology*. Vol.11. no.3, pp.644-652, May 2016.
- [10] Y. Wang, S. Zhang, J. M. Alonso, X. Liu and D. Xu, "A Single-Stage LED Driver With High-Performance Primary-Side-Regulated Characteristic," in *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 65, no. 1, pp. 76-80, Jan. 2018.
- [11] A. Abramovitz, C.-S. Liao, and K. Smedley, "State-plane analysis of regenerative snubber for flyback converters," *IEEE Trans. Power Electron.*, vol. 28, no. 11, pp. 5323-5332, Nov. 2013.
- [12] Y. Ma, F. -I. Chou, P. -Y. Yang, J. -T. Tsai, Z. -Y. Yang and J. -H. Chou, "Optimal Parameter Design by NSGA-II and Taguchi Method for RCD Snubber Circuit," in *IEEE Access*, vol. 8, pp. 182146-182158, 2020.
- [13] L. Chen, H. Hu, Q. Zhang, A. Amirahmadi and I. Batarseh, "A Boundary-Mode Forward-Flyback Converter With an Efficient Active LC Snubber Circuit," in *IEEE Transactions on Power Electronics*, vol. 29, no. 6, pp. 2944-2958, June 2014.
- [14] H. Cheng, Y. Chang, C. Chang, S. Hsieh and C. Cheng, "A Novel High-Power-Factor AC/DC LED Driver With Dual Flyback Converters," in *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 7, no. 1, pp. 555-564, March 2019.
- [15] H. -L. Cheng, Y. -N. Chang, H. -C. Yen, C. -C. Hua and P. -S. Su, "An Interleaved Flyback-Typed LED Driver With ZVS and Energy Recovery of Leakage Inductance," in *IEEE Transactions on Power Electronics*, vol. 34, no. 5, pp. 4497-4508, May 2019.
- [16] N. M. Mukhtar and D. D. Lu, "A Bidirectional Two-Switch Flyback Converter With Cross-Coupled LCD Snubbers for Minimizing Circulating Current," in *IEEE Transactions on Industrial Electronics*, vol. 66, no. 8, pp. 5948-5957, Aug. 2019.
- [17] S. Jung and G. Cho, "Transformer Coupled Recycle Snubber for High-Efficiency Offline Isolated LED Driver With On-Chip Primary-Side Power Regulation," in *IEEE Transactions on Industrial Electronics*, vol. 61, no. 12, pp. 6710-6719, Dec. 2014.
- [18] S. Lee and H. Do, "A Single-Switch AC-DC LED Driver Based on a Boost-Flyback PFC Converter With Lossless Snubber," in *IEEE Transactions on Power Electronics*, vol. 32, no. 2, pp. 1375-1384, Feb. 2017.
- [19] D. S. L. Simonetti, J. Sebastian and J. Uceda, "The discontinuous conduction mode Sepic and Cuk power factor preregulators: analysis and design," in *IEEE Transactions on Industrial Electronics*, vol. 44, no. 5, pp. 630-637, Oct. 1997.
- [20] Y. Wang, F. Li, Y. Qiu, S. Gao, Y. Guan and D. Xu, "A Single-Stage LED Driver Based on Flyback and Modified Class-E Resonant Converters With Low-Voltage Stress," in *IEEE Transactions on Industrial Electronics*, vol. 66, no. 11, pp. 8463-8473, Nov. 2019.
- [21] Q. Luo, J. Huang, Q. He, K. Ma and L. Zhou, "Analysis and Design of a Single-Stage Isolated AC-DC LED Driver With a Voltage Doubler Rectifier," in *IEEE Transactions on Industrial Electronics*, vol. 64, no. 7, pp. 5807-5817, July 2017.
- [22] S. Zhang, X. Liu, Y. Guan, Y. Yao and J. M. Alonso, "Modified zero voltage-switching single-stage LED driver based on Class E converter with constant frequency control method," in *IET Power Electronics*, vol. 11, no. 12, pp. 2010-2018, 16 10 2018.
- [23] Y. Wang, X. Deng, Y. Wang and D. Xu, "Single-Stage Bridgeless LED Driver Based on a CLCL Resonant Converter," in *IEEE Transactions on Industry Applications*, vol. 54, no. 2, pp. 1832-1841, March-April 2018.
- [24] B. Poorali and E. Adib, "Analysis of the Integrated SEPIC-Flyback Converter as a Single-Stage Single-Switch Power-Factor-Correction LED Driver," in *IEEE Transactions on Industrial Electronics*, vol. 63, no. 6, pp. 3562-3570, June 2016.
- [25] Y. Wang, J. Huang, G. Shi, W. Wang and D. Xu, "A Single-Stage Single Switch LED Driver Based on the Integrated SEPIC Circuit and Class-E Converter," in *IEEE Transactions on Power Electronics*, vol. 31, no. 8, pp. 5814-5824, Aug. 2016.
- [26] K. Cao, X. Liu, M. He, X. Meng and Q. Zhou, "Active-Clamp Resonant Power Factor Correction Converter With Output Ripple Suppression," in *IEEE Access*, vol. 9, pp. 5260-5272, 2021.



**Vinod Kumar Yadav** received the B.E. degree in Electrical Engineering from the Jabalpur Engineering College, Jabalpur, India, in 2015, the M.Tech. degree in Power Electronics and Drives from the National Institute of Technology Rourkela, Orissa, India, in 2017, and the Ph.D. degree in the Department of Electrical Engineering, Malaviya National Institute of Technology, Jaipur, India, in 2023. He is currently working as a postdoctoral fellow in the Department of Sustainable Energy Engineering, Indian Institute of Technology, Kanpur, India. He worked as an assistant professor in the Department of Electrical Engineering at the Government Engineering College Bikaner, Rajasthan, India. He also worked as a Research Associate at Delhi Technological University, Delhi, India, and Motilal Nehru National Institute of Technology, Allahabad, India. His research interests include AC-DC PFC converters, LED driver circuits, high-gain boost converters, EV chargers, Hydrogen fuel technology, and soft switching.



# A multilevel authentication-based blockchain powered medicine anti-counterfeiting for reliable IoT supply chain management

Neetu Sharma<sup>1</sup> · Rajesh Rohilla<sup>1</sup>

Accepted: 4 September 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

## Abstract

In this paper, a novel blockchain-based multilevel security and authentication application is proposed to address the problem of counterfeiting in the medicine supply chain. In this blockchain-enabled application, multiple computers owned by different supply chain organizations collaborate in a distributed fashion to establish a trustworthy network. By adopting a resource-efficient Hyperledger Fabric framework, the proposed system aims to enhance transparency, integrity, and traceability in the pharmaceutical supply chain. Hyperledger Fabric blockchain also adds network security, privacy preservation, real-time tracking, and reliability features to the proposed system. Moreover, the proposed system also incorporates a blockchain-based QR code watermarking layer for authentication and verification. A validation module is designed that empowers buyers to validate the identity and history of serialized medicinal products, and buyer identity validation ensures that only legitimate buyers can access the products. Simulation outcomes and performance measurements demonstrate the effectiveness of the proposed system, including location traceability, QR code authentication, and verification. The Caliper tool has been used to measure the performance of the proposed system in terms of execution time, throughput, latency, and resource statistics for up to 100,000 transactions and 20 peers. The highest throughput achieved is 417.5 TPS (transactions per second) with 100,000 transactions and eight peers. The QR code authentication performance is tested under various noisy, cropped, and blurred attacks. We presented simulation outcomes of the implemented supply chain as well as chain code algorithms. The results indicate improved scalability, validation mechanisms, throughput, latency, and resource consumption compared to existing schemes.

**Keywords** Counterfeiting · Blockchain · Hyperledger Fabric · Network security · Supply chain · Privacy-preserving

---

Extended author information available on the last page of the article

Published online: 21 September 2023

Springer



# 1 Introduction

Medicine counterfeiting has severe impacts on health and lives, leading to a high number of deaths annually [1]. The pharmaceutical industry has long grappled with drug counterfeiting issues that can be tackled using the Hyperledger Fabric framework [2]. This paper aims to design a multilevel authentication and security approach using Hyperledger Fabric and blockchain-based QR code watermarking. By combining blockchain encryption with QR verification, the reliability of anti-counterfeiting measures can be enhanced. A decentralized approach is preferable to combat counterfeiting, as centralized solutions are vulnerable to attacks by intermediaries in the supply chain. Moreover, it improves reliability and fault tolerance. Blockchain technology offers the pharma industry opportunities to improve traceability, transparency, and drug safety [3]. Implementing a method to track product authenticity from manufacturing to consumption is crucial [4]. Blockchain enables automatic verification within the supply chain, providing built-in trust and supporting supply chain sustainability [5–7]. Blockchain platforms programmatically enable trust, assuring consumers of product quality [8]. Research confirms the potential benefits of blockchain adoption in supply chain processes for both consumers and manufacturers [9]. Major technology companies such as IBM and Cisco are actively exploring blockchain solutions for health care, including medicine counterfeiting [10]. Start-ups have already introduced blockchain-based designs for tracking medicines in the supply chain [11]. However, practical implementation faces significant challenges [12]. Enterprise solutions require fast transactions, privacy, security, and scalability. Analyzing the interplay between barriers to blockchain adoption in global trade is essential [13].

## 1.1 Role of blockchain technology in supply chain management

Blockchain applications are being explored across various sectors, including life sciences and pharmaceuticals, to manage health records [14], practitioner information [15], drug supply chains, and more. Blockchain can offer an effective solution for supply chain traceability. By digitizing assets and maintaining a decentralized, immutable ledger of transactions, pharma companies can track product journeys from manufacturing to delivery [16]. Blockchain, combined with IoT, can bolster the drug supply chain against medicine counterfeiting. Mobile apps with integrated security platforms authenticate and track medicine units using QR codes/barcodes. IoT connects physical items via the internet, assigning unique identities through physical identifiers. Blockchain's attributes, such as transparency, traceability, accountability, privacy, security, immutability, integrity, and provenance, make it applicable to diverse domains. Investments in blockchain have surged tenfold in the past 5 years [11], emphasizing accountability and

confidentiality [17]. Real-world applications face challenges in privacy, audibility, and authorization [18]. As a decentralized infrastructure and secure authorization mechanism [19], blockchain enhances governing administrations' efficiency by ensuring traceability throughout the medicine supply chain.

## 1.2 Motivation

This research aims to develop a transparent and trustworthy blockchain-based solution for managing the medicine supply chain, effectively combating counterfeiting and saving lives. Logistics, encompassing transportation and storage, is a critical aspect of the supply chain, and blockchain can enhance its integrity by establishing programmatic trust among participants. By addressing the limitations of centralized systems, blockchain technology offers a robust solution. Although still in its nascent stage, efforts are underway to address challenges related to speed, scalability, and throughput. Access control restrictions would greatly benefit use cases such as the medicine supply chain. In this work, we leverage the Hyperledger Fabric framework to design and implement a proposed medicine supply chain solution. Given the growing popularity of QR code security algorithms, this research focuses on QR code authentication and evaluates its performance against relevant attacks in the specified context.

## 1.3 Contributions

The main contributions of this research are as follows:

1. We designed a Hyperledger blockchain-based medicine supply chain integrated with IoT to address vulnerabilities in the existing supply chain system.
2. We developed novel chain code algorithms with access privilege restrictions and multiple validations to govern the entire workflow of the medical supply chain.
3. We introduced encrypted QR codes and proposed a special QR code authentication method using blockchain-based invisible watermarking to enhance reliability. Conducting an analysis of attacks for evaluation purposes.
4. Empowered buyers to validate the identity of serialized medicinal products in real time by scanning encrypted QR codes and accessing transaction history. Real-time notification is provided in case of duplicate QR codes. Also, implemented a buyer identity validation mechanism to ensure that only legitimate buyers can receive the medicinal products.
5. We demonstrated the effectiveness of the chain code algorithms by presenting simulation results and evaluating performance metrics for up to 100,000 transactions. The proposed design outperforms existing schemes with a superior validation mechanism, achieving higher throughput, lower latency, and requiring fewer computational resources.

## 2 Prior work

In this review, the focus is on existing research related to the usability of blockchain technology in combating medicine counterfeiting, improving supply chain management, addressing public health concerns, and blockchain watermarking.

### 2.1 Survey related to usability of blockchain in supply chain

One of the studies [1] presented a novel medicine supply chain management system that utilizes blockchain technology on the Hyperledger Fabric platform to securely share medicine supply chain records. This study lacks drug tracking and validation mechanisms. Table 1 provides a summary of existing drug counterfeiting and blockchain-based supply chain schemes. It reveals that many previous works lack traceability, scalability, and validation mechanisms, and some have limited design development. Blockchain-based supply chain design with traceability mechanisms was demonstrated in [3]. This study also lacks validation mechanisms. Another study [20] introduced the Gcoin blockchain for transparent transactions in the medicine domain to prevent falsified medicines. This study presents theoretical concepts only, not design implementation. The importance of blockchain technology in various sectors, including medicine and food, was discussed in [21]. The authors assessed the suitability of blockchain for multi-criteria decision-making (MCDM) problems in supply chain management using hesitant fuzzy sets (HFSs). This study also presented theoretical contributions rather than solution implementation. Supply chain verification based on GPS location traceability was proposed in [22], but the proposed design lacks validation mechanisms.

The challenges associated with existing health-care blockchains were identified in [23], which highlights the medical supply chain as a promising area for research. In [24], various traditional and emerging technologies for addressing medicine counterfeiting were compared, with blockchain technology identified as the most promising solution. This study conducted a review of existing literature in the field without introducing any novel contributions. An improved supply chain management framework based on blockchain technology is proposed in [25]. This work lacks design implementation. Performance evaluations of blockchain-based health-care systems using the Hyperledger Fabric framework are conducted in [26] and [27], with transaction rates of up to 250 and 10,000, respectively. These works lack scalability. It is worth noting that the testing in [26] was limited to a maximum of 250 transaction rates and was computationally expensive. Regarding [27], their performance evaluation was limited to 10,000 transactions, and their system was unable to serve more than 4 users within this range.

In contrast, we have successfully tested our proposed system up to 100,000 TPS, demonstrating its scalability and improved performance. The proposed work offers a product validation mechanism that complements access control, tracking, privacy, scalability, and multilevel authentication features, which are crucial for ensuring security in medicine supply chain management.

**Table 1** Summary of existing blockchain-based drug counterfeiting and supply chain schemes

Existing scheme	Objective	Limitation
[1]	To propose a novel medicine supply chain model for a smart hospital	Lacks drug tracking and validation mechanisms
[3]	To design Hyperledger-enabled drug supply chain	Lacks drug validation mechanism
[11]	To use blockchain in medicine	Only theoretical concepts
[20]	To introduce Gcoin blockchain for smooth transactions of medicine	Design concept only, not solution development
[21]	To explore the usage of blockchain for multi-criteria decision-making in supply chain	Implementation is not available
[22]	To design Hyperledger-enabled supply chain for textile industry	Lacks design algorithms, scalability, and product validation mechanism
[24]	To review the role of emerging technologies in fight against fake medicines	Only theoretical concepts
[25]	To use blockchain in tracking process of supply chain	Lacks design implementation
[26]	To evaluate the performance of blockchain-based health-care system	Lacks scalability and computationally expensive
[27]	To propose a Hyperledger-enabled health-care system	Lacks scalability
[28]	To propose a blockchain-based counterfeit-proof supply chain	Less scalable and privacy-preserving
[29]	To propose an approach to combine IOT data collection modules with blockchain using fuzzy logic model	Less scalable and privacy-preserving
[30]	To propose a user authentication mechanism	Limited performance evaluation
Proposed	Multilevel authentication-based blockchain-based medicine anti-counterfeiting approach	Authentication, access control, product validation, scalability, design implementation, and improved performance



In [28], the authors proposed a blockchain-based counterfeit-proof supply chain approach that leverages radio frequency identification (RFID) and Ethereum blockchain technology. This work is less scalable and privacy-preserving due to the use of the public blockchain. The performance evaluation is limited to 600 transactions, with maximum throughput up to 60 TPS and latency up to 1600 ms. Also, the use of RFID reduces the range of product traceability. In contrast, the proposed work utilizes the Hyperledger framework and IoT technology to improve privacy, scalability, and real-time monitoring over longer distances. Moreover, the performance of the proposed work is evaluated up to 100,000 transactions, with maximum throughput up to 417.5 TPS and latency up to 0.15 s. Further, we also evaluated performance metrics such as execution time, throughput, and latency in terms of the number of peers.

In [29], the authors combined the IoT data collection modules with the Ethereum blockchain by utilizing a fuzzy logic model. This work is also less scalable and privacy-preserving due to use of the public blockchain. In [30], the authors presented a user authentication mechanism by integrating blockchain with access control and a physical unclonable function (PUF). This work is not focused on product traceability. Also, the performance evaluation is limited to 1000 transactions. In comparison with these Ethereum-based works, the proposed Hyperledger-based design offers more scalability and better performance.

In [31], a product certification scheme for automotive supply chains was proposed. This work leverages the Hyperledger Besu blockchain, which is an Ethereum client developed under the Hyperledger umbrella. However, the study reported that the deployed blockchain network experienced delays and incurred high transaction costs. In [32], a deep learning-based attack prediction mechanism for supply chain management was introduced. The focus of this study was on detecting attacks rather than formulating a protection strategy. A blockchain–IoT-based decentralized storage approach for data protection is studied in [33], emphasizing the security and privacy aspects of the system. In response to the COVID-19 pandemic, a personal protective equipment supply chain model using blockchain technology is suggested in [34], aiming to enhance efficiency and traceability in the distribution of essential supplies.

Overall, prior work indicates that emerging digital technologies such as IoT and blockchain have the potential to enhance the efficiency, transparency, and trustworthiness of supply chain systems.

## 2.2 Survey related to blockchain watermarking

A significant amount of research has been conducted to enhance the accuracy and security of QR code anti-counterfeiting techniques. In [35], the authors discussed the use of anti-counterfeiting watermarking techniques as an alternative method to prevent the duplication or fabrication of QR barcodes. They highlighted important algorithms that enable the creation of digital QR barcodes. For printed QR barcodes, a multi-channel-resistant watermarking system based on the discrete wavelet transform (DWT) was proposed for QR code verification in [36]. This method aimed

to improve the security of printed QR codes. In [37], an anti-counterfeiting method was presented that utilized statistical evaluation of a single related feature differential sequence of a critical region. The method involved the use of inks to create random delicate texture patterns and employed a supervised and assisted segmentation technique and a bone width transformation algorithm to identify critical parts of sample photos. In [38], a QR code-based watermarking method for securing digital images was explored. The features of the QR code framework, such as error correction and high information capacity, made it advantageous for watermarking. The study focused on the security and robustness of the QR code watermarking system against common digital picture attacks.

In the context of digital rights management (DRM), QR code-based image watermarking using the DWT transform domain was demonstrated in [39]. This approach aimed to protect digital images by embedding watermarks. A dual anti-counterfeiting strategy for QR codes was proposed in [40], incorporating encryption and digital watermarking frameworks. The method involved encrypting authorization data using RSA-based encryption, generating a respective watermark image from the encrypted data, and employing DWT and singular value decomposition (SVD)-oriented techniques for anti-print extraction and embedding image watermarking. The utilization of QR codes to facilitate data transmission in health-care settings was discussed in [41]. The study focused on developing a platform for medical data in health files and employed advanced encryption standards to secure the information contained in QR codes.

DWT-based watermarking techniques for color image security were presented in [42], utilizing edge detection and the combination of bit plane complexity (block-chain). The HH element of the DWT transform domain was used for watermark embedding. The performance of the method was evaluated under various watermark attacks. Efficient digital color image watermarking techniques with encryption utilizing blockchain and DWT edge coefficients were proposed in [43], aiming to enhance the security of digital images. Similarly, edge detection-based watermarking techniques were proposed in [44], emphasizing the importance of incorporating edge information for watermark embedding.

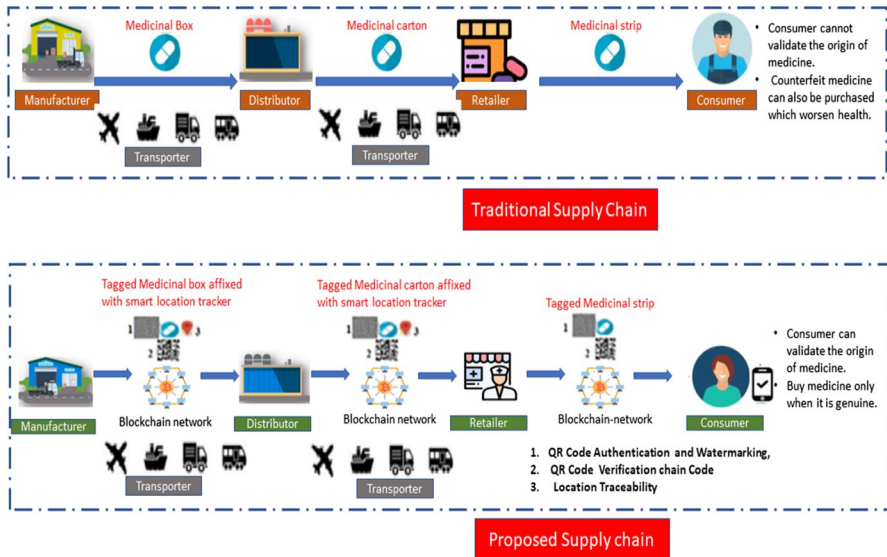
Overall, these studies have contributed to advancing the field of QR code anti-counterfeiting and watermarking, exploring various techniques and evaluating their security, robustness, and performance under different conditions and attacks.

### 3 Proposed medicine supply chain management design

In this section, we provide an overview of the proposed medicine supply chain design, outlining its key elements and layer architecture.

#### 3.1 Overview

Counterfeiting poses significant challenges in the medicine supply chain, causing substantial financial losses and damaging the reputation of pharmaceutical



**Fig. 1** Traditional vs. proposed multilayer security-based medicine supply chain schemes

companies, and also posing significant risks to public health. To address this issue, our proposed work incorporates the integration of blockchain technology and IoT to enhance the management of the medicine supply chain. In this blockchain-enabled distributed application, multiple computers owned by different supply chain organizations collaborate in a distributed fashion to establish a trustworthy network. Figure 1 illustrates a comparison between traditional approaches and our proposed multiple security-based, reliable blockchain-based medicine supply chain schemes. The figure clearly demonstrates that our proposed approach provides multilayer security measures. The system incorporates three levels of verification for medicines, including QR code authentication, verification, and traceability.

The key stakeholders involved in the supply chain system are manufacturers, distributors, retailers, consumers, and transporters. Manufacturers, representing pharmaceutical companies, serve as sellers in the system. Distributors, known as medicine wholesalers, act as intermediaries, purchasing medicinal products in bulk directly from manufacturers and functioning as both sellers and buyers. Retailers, which include small stores and pharmacies, acquire medicine consignments from distributors and also operate as both sellers and buyers. Consumers, as end users, purchase medicines from retailers and solely fulfill the role of buyers. Transporters play a crucial role in handling the shipping and logistics of consignments within the supply chain.

In both the traditional and proposed systems, the flow of medicine starts from the manufacturer and reaches the consumer. However, in the traditional scenario, consumers lack the ability to validate the origin of medical products, which puts them at risk of unknowingly purchasing fake or counterfeit medicine that could worsen their health conditions.

In the proposed system, every consumer can verify the origin of a medicine product by simply scanning its QR code using a smartphone. The use of blockchain and IoT technology ensures that the system can detect invalid or duplicate codes, providing an additional layer of security.

The proposed scheme involves the serialization of all medicines by the manufacturer, who then labels them with encrypted QR (E-QR) codes containing their serial numbers. Each shipment prepared by the manufacturer consists of medicinal boxes, each labeled with an E-QR code that identifies the cartons inside the box, and the cartons, in turn, are labeled with codes representing the set of product IDs for all the medicines they contain. The boxes and cartons are equipped with smart location trackers to facilitate the tracking of medicine products throughout the supply chain.

In this system, distributors are only able to purchase medicinal boxes, retailers can purchase cartons, and consumers can purchase medicine strips. The manufacturer initiates the process by sending a labeled box, along with a location tracker, to the distributor. The location tracker continuously transmits location data to the blockchain network, enabling real-time tracing and tracking of medicinal products. The distributor can access the E-QR code (1) on the box, decrypt it using blockchain de-watermarking, and then validate the history of the medicinal products inside the box by scanning the QR code (2) through the supply chain app from their registered location (3). The distributor then sends labeled cartons, also tagged with smart location trackers, to the retailer. Retailers can view and validate the history of the medicinal products received by scanning the QR code on the carton. Finally, when retailers sell labeled medicine strips, consumers can also validate the authenticity of the medicinal products. At each point of exchange, the unique product ID is verified against the information stored on the distributed blockchain platform, ensuring the genuineness of the product.

### 3.2 Building components

The proposed medicine supply chain system consists of four main components: organizations, medicinal assets, smart contract modules, and smart devices. The organizations involved in the system are manufacturers, distributors, retailers, consumers, and transporters. The medicinal assets are the medicines produced by manufacturers. The smart contract modules play a crucial role in the system and provide various transaction functionalities. The following smart contract modules are implemented:

1. *Onboarding*: This module handles user registration requests and validates them. All organizations, except consumers, are required to register on the blockchain network.
2. *Inventory*: The inventory module serializes all medicinal products after each production phase and creates an inventory of medicines. Its main purpose is to ensure traceability, protect against counterfeiting, and manage uncertainties.

3. *Ordering*: The ordering module facilitates the collection and processing of purchase orders for medicines. Buyers can create purchase orders through this module.
4. *Shipping*: This module enables sellers to create shipments of medicines. It handles the process of preparing and dispatching medicines for delivery.
5. *Validation*: The validation module is responsible for authenticating stakeholders and allowing authorized buyers to validate the identity of medicinal products. It also enables transporters to update the shipment status and retailers to retail medicines after successful validation.
6. *View*: The view module provides functionality to trace the events in the life cycle of a medicinal product, starting from its origin. It allows stakeholders to track and monitor the movement of medicines.

Smart devices used in the system include location trackers and scanners (or smartphones). Location trackers are utilized to maintain real-time information about the location of medicines, ensuring tracking. Smartphones or scanners are used by stakeholders to validate the authenticity of medicinal assets. The proposed supply chain design follows a layered architecture, as shown in Fig. 2. It incorporates blockchain-based watermarking (BCW) with encryption/decryption capabilities as the third layer of authentication, enhancing the reliability of the system. This layered architecture describes the procedure for adding medicine supply chain transactions to the blockchain when clients invoke smart contract functionalities.

The proposed layered architecture in Fig. 2 provides clear authority for each member of the supply chain to verify medicines. The trunk, box, or strip of medicine are labeled with a BCW code, which needs to be decrypted to generate a QR code. Before accessing the medicine, the QR code must be verified. This verification process ensures the traceability and authenticity of the medicine throughout the supply chain. The layered architecture consists of three layers: the user layer,

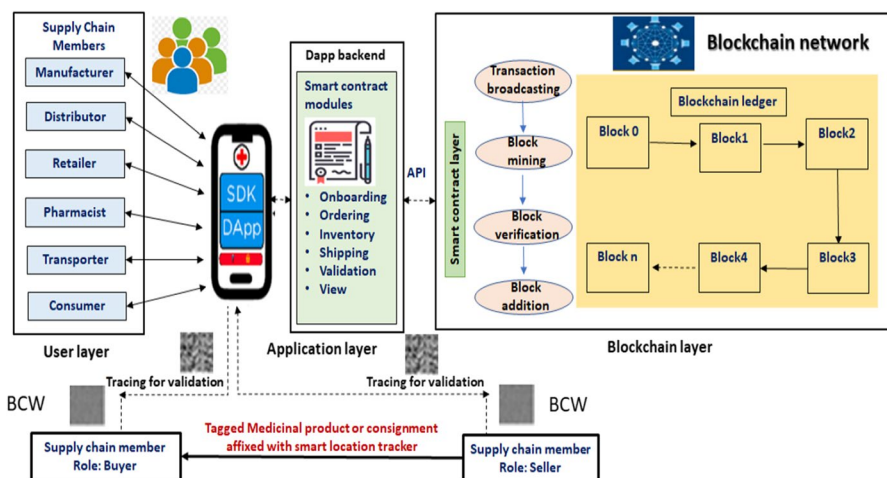


Fig. 2 Layered architecture of the proposed supply chain

the application layer, and the blockchain layer. The user layer includes the supply chain members who connect with the application layer to send transaction requests and receive transaction states. The distributed application (Dapp) layer serves as an interface between the blockchain clients (users) and the blockchain peer nodes. It enables transaction uploading and retrieval and provides APIs for invoking smart contract services within the blockchain network. The Dapp layer also returns transaction states obtained from the blockchain network to the users.

The blockchain layer contains a decentralized database that stores information in a sequential chain of linked blocks. Its purpose is to create trust among participants in the blockchain network. This layer performs tasks such as broadcasting transactions to all peer nodes, mining transaction blocks, verifying transaction blocks, and appending blocks into the blockchain ledger. Each block represents a specific set of transactions and is added to the blockchain ledger in a sequential manner. In this layered architecture, the transfer of medicine occurs from the seller to the buyer. Users first send registration requests to become part of the blockchain-based supply chain network with different roles. The onboarding module processes the user details and updates user profiles in the ledger upon validation. After each production phase, the manufacturer serializes and registers each medicinal product on the ledger. The manufacturer then attaches a smart location tracker to the medicine shipment.

When a buyer places an order through the ordering module, the seller creates a consignment shipment for the transporter. The transporter can update the shipment status as the consignment is delivered to the buyer's registered location. Upon receiving the shipment, the buyer verifies its authenticity by scanning the QR code with a smartphone Dapp. Only a genuine buyer with a valid identity can receive the shipment at the designated location. The use of the smartphone Dapp allows buyers to identify any instances of counterfeiting. If the QR code is a copy from a different transaction or not registered with the blockchain network, it will be deemed invalid. The BCW authentication layer provides an additional layer of security to enhance the reliability of the system. Overall, the proposed layered architecture ensures the transparency, traceability, and authenticity of medicines in the supply chain.

## 4 Process implementation

In this section, we delve into the fundamentals of Hyperledger Fabric, explore chain-code modules, and proceed with the implementation of the proposed Hyperledger-based medicine supply chain framework.

### 4.1 Fundamentals of Hyperledger Fabric blockchain and its consensus mechanism

In the proposed work, a private Hyperledger blockchain is used, which is a more efficient and energy-saving blockchain than a public blockchain. Moreover, Hyperledger Fabric blockchain also provides network security, privacy preservation, real-time tracking, reliability, and fault tolerance features. Moreover, Hyperledger

blockchains also provide the scalability, privacy, and access restrictions that cannot be achieved using public blockchains. Hyperledger is an open-source blockchain project hosted under the Linux Foundation, encompassing various frameworks tailored for different use cases. Among these frameworks, Hyperledger Fabric stands out as an enterprise-ready, open-source blockchain solution. To measure the performance of frameworks within the Hyperledger ecosystem, the Caliper tool is employed.

In a Hyperledger network, several elements play crucial roles:

- *Channel*: Facilitates data compartmentalization between stakeholders in the network.
- *Assets*: Represents the data being tracked and stored on the network.
- *Transaction*: Enables the alteration of asset state within the network.
- *Ledger*: Maintains a list of transactions and the corresponding asset states.
- *World state*: Represents the current state of all assets and their associated transactions.
- *Smart contract (Chain code)*: Contains the logic responsible for executing transactions and modifying asset states.
- *Peer*: A computing resource that participates in the network.
- *Ordering service*: Receives transactions from peers, sequences them into blocks, and writes them onto the ledgers of each peer.
- *MSP (Membership Service Provider)*: Offers credentials or IDs required by applications running on peers to interact with the network. As Hyperledger operates as a permissioned network, each network element must possess identification.
- *Certificate authority (CA)*: Issues certificates to distinguish different elements within the network.
- *Anchor peer*: Each organization designates one of its peers as the Anchor peer, responsible for maintaining communication with Anchor peers from other organizations.
- *Committer*: Responsible for writing or committing blocks published by the ordering service onto the ledger. Any peer belonging to the channel where the transaction was initiated can act as a committer for that transaction.
- *Endorser*: Receives incoming transactions from applications, simulates them, and forwards them to the ordering service. A peer becomes an endorser in the network if it has a deployed chain code.

A key component of any blockchain network is the consensus mechanism, which is a protocol used to achieve agreement among multiple participants on the state of the ledger. To ensure secure and timely transaction validation while consuming minimal resources, Hyperledger blockchains employ efficient and time-effective consensus algorithms. In Hyperledger Fabric, consensus is achieved through a pluggable



consensus architecture. Users of Hyperledger Fabric can choose from a number of consensus mechanisms based on the needs of their network. Many public blockchains, on the other hand, use a single consensus algorithm, such as Proof of Work (PoW) or Proof of Stake (PoS). This adaptability is particularly useful in enterprise scenarios, where different networks may have different trust models and performance requirements. Hyperledger Fabric supports a number of consensus mechanisms, including Solo, Raft, and Kafka, which are listed below:

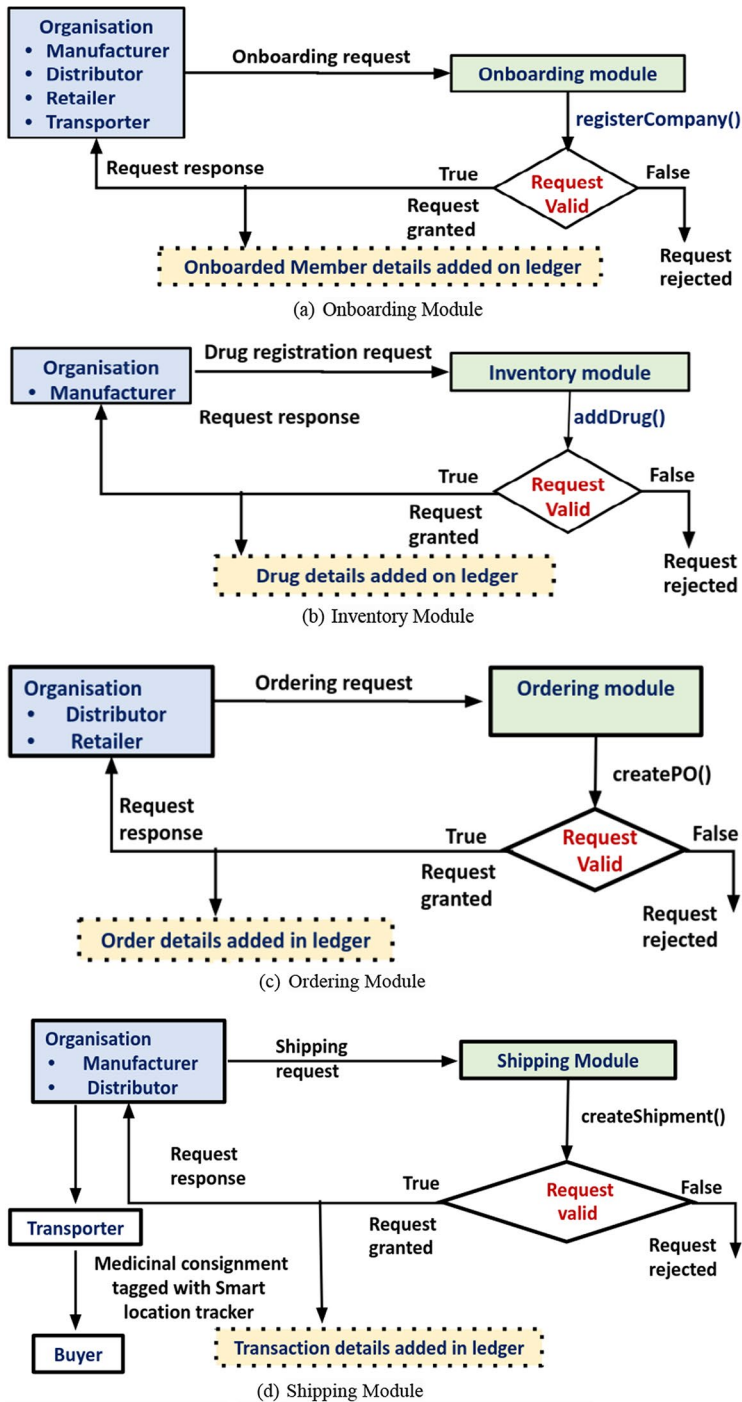
- **Solo consensus:** Hyperledger Fabric provides a simple, single-node consensus mechanism called Solo. It does not require multiple nodes to reach consensus and is primarily used for development and testing. It is unsuitable for production networks due to a lack of fault tolerance and Byzantine fault tolerance.
- **Raft consensus:** Raft is a fault-tolerant, replicated consensus algorithm. It ensures that the network nodes agree on the ledger state. The Raft Hyperledger Fabric implementation allows users to configure a group of nodes to act as orderers, which collectively order transactions and generate consistent blocks.
- **Kafka consensus:** It uses the Kafka messaging system to order and distribute transactions to peers. This has the potential to provide high throughput and scalability, making it appropriate for high transaction rate networks.

## 4.2 Workflow through chaincode modules

The proposed system's workflow is divided into six chaincode modules: onboarding, inventory, ordering, shipping, validation, and viewing. These modules handle the transactional flow of medicinal products from the manufacturer to the consumer. Figure 3 provides an illustration of the onboarding, inventory, ordering, and shipping modules within the system.

The onboarding chaincode module is responsible for collecting and validating enrollment requests submitted by various entities. Upon validation, the module registers the new entity, and if the details are found to be invalid, the request is rejected. Additionally, the onboarding module updates the details of onboarded members in the ledger. Algorithm 1 presents the chaincode of the onboarding module, which encompasses the following functionalities:

- **Register company:** This function is utilized to enroll a new user on the ledger, assigning them one of the following roles: manufacturer, distributor, retailer, or transporter. The function includes fields such as company ID, company CRN, company name, location, organization role, and hierarchy key. The company ID field is a composite key derived from the combination of the company CRN and company name. The hierarchy key field is assigned a value of 1 for a manufacturer, 2 for a distributor, and 3 for a retailer organization. However, the hierarchy key is not defined for transporters.



**Fig. 3** The workings of onboarding, inventory, ordering, and shipping modules

**Algorithm 1:** Algorithm of the Onboarding chaincode module.

```

1. Input : company Name, company CRN, Location, organisation Role
2. function REGISTERCOMPANY (company CRN, Company Name, Location,
   Organisation Role)
3.   fetch user company details through company CRN
4.   if (user's company details not exist) then
5.     create new user company ID
6.     if (organisation Role == "Manufacture" ||
       organisation Role == "Distributor"
       || organisation Role == "Retailer") then
7.       if (organisation Role == "manufacturer") then
8.         hierarchy Key = 1;
9.       else if (organisation Role == "Distributor") then
10.        hierarchy Key = 2;
11.      else if (organisation Role == "Retailer") then
12.        hierarchy Key = 3;
13.      else
14.        Not buyer or seller Organisation
15.      end if
16.      else if (organisation Role == "Transporter") then
17.        No hierarchy key is needed for transporter
18.      else Invalid Organisation
19.        add user company details on the ledger
20.      end if
21.      else user's company already registered
22.    end if
23.  end function
24. OUTPUT: company IDs (manufacturer IDs, distributor IDs, retailer IDs, consumer IDs
   and transporter IDs), name, location, organisation Role, hierarchy Key.

```

The inventory chaincode module is responsible for registering medicinal products on the ledger. Valid drug registration requests result in the addition of a new drug to the blockchain network, and invalid requests are not registered within the system. Algorithm 2 presents the chaincode of the inventory module, which encompasses the following functionality:

- **addDrug:** This function is designed to register a new drug, but only authorized manufacturers can invoke this transaction on the blockchain network. The function includes fields such as product ID, drug name, serial number, manufacturer, manufacturing date, expiration date, owner, and shipment. The product ID field serves as a composite key derived from the combination of the drug name and serial number. Initially, the shipment field is empty, and the owner is set as the manufacturer invoking this transaction.

**Algorithm 2 :-** Algorithm of the inventory chaincode module

1. **Input:** drug Name, Serial No., mfg Date, Exp Date, Company CRN
2. **function** ADDDRUG (Drug Name, Serial Name, mfg Date, Exp date, Company CRN)
3.     Fetch company details through company CRN
4.     **if** (Company organisation Role === manufacturer) **then**
5.         Register medicine on the ledger
6.     **else**
7.         Invalid transaction sender
8.     **end if**
9. **end function**
10. **OUTPUT** – product ID, drug Name, mfg Date, Exp. Date, Owner, shipment

The ordering module is responsible for collecting and validating order requests placed by enrolled buyers. Valid orders result in the generation of purchase order IDs, and invalid requests are rejected.

Algorithm 3 presents the chaincode of the ordering module, which includes the following functionality:

- **createPO:** This function enables buyer members, specifically distributors or retailers, to add purchase orders for buying medicine on the ledger. The function requires fields such as purchase order ID, drug name, quantity, buyer ID, and seller ID. The purchase order ID is derived as a composite key, combining the drug name and buyer CRN.

**Algorithm 3:-** Algorithm of the ordering chaincode module

1. **INPUT:** buyer CRN, seller CRN, drug Name, quantity
2. **function** CREATEPO (buyerCRN, sellerCRN, drugName, quantity)
3.     fetch buyer and seller company details through buyer CRN and seller CRN
4.     **if** (buyer Details. Organisation Role === “Retailer”  
        && seller Details. Organisation Role === “Distributor”  
        || (buyer Details. Organisation Role === “Distributor”  
        && seller Details. Organisation Role === “Manufacturer”)) **then**
5.         create purchase order ID
6.         Add purchase order on the ledger
7.     **else**
8.         Invalid order request
9.     **end if**
10. **end function**
11. **OUTPUT:** purchase order ID, drug Name, Quantity, buyer, seller

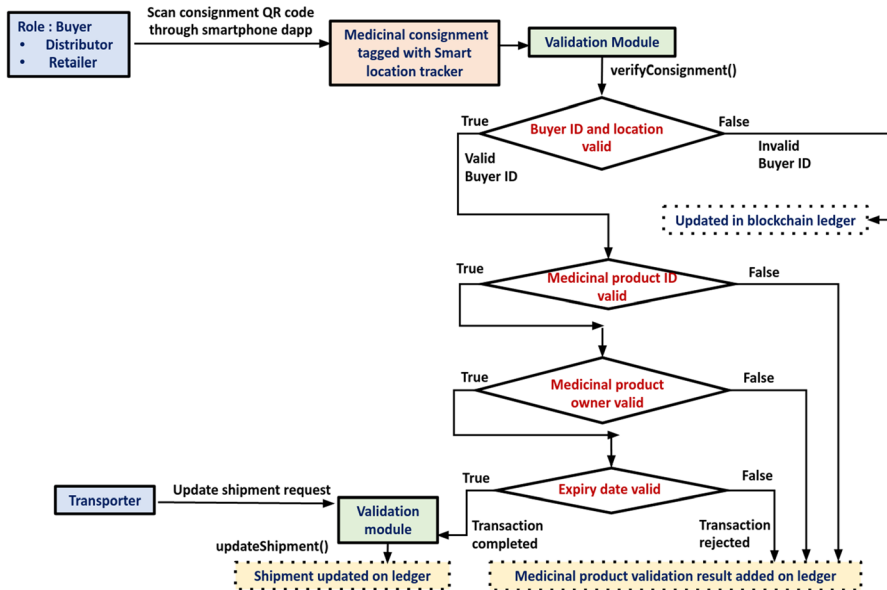
The shipping module is responsible for processing newly created purchase orders and facilitating the transportation of medicinal consignments through a designated transporter. Algorithm 4 outlines the chaincode of the shipping module, which includes the following functionality:

- **CreateShipment:** This function is designed to transport the medicinal consignment using a designated transporter. It is invoked by seller members in response to the purchase order function. The function requires fields such as shipment, transaction creator, list of assets, transporter, and status. The shipment field is derived as a composite key, combining the buyer name and buyer CRN.

Algorithm 4: An Algorithm of the shipping Chaincode module	
1.	<b>INPUT:</b> buyer CRN, drug Name, list Of Assets, transporter CRN
2.	<b>functions</b> CREATESHIPMENT (buyerCRN, drugName, listOfAssets, transporterCRN)
3.	fetch purchase order details through buyerCRN and drugName
4.	<b>if</b> (listOfAssets. length == Purchase OrderDetails. quantity) <b>then</b>
5.	fetch assets details from list of assets
6.	create shipment with shipment ID, shipment status 'In transit', and add them on the ledger
7.	<b>else</b>
8.	Invalid shipment details
9.	<b>end if</b>
10.	<b>end function</b>
11.	<b>OUTPUT-</b> shipment ID, transaction creator, assets, transporter ID, shipment status

The validation chaincode module plays a crucial role in validating the buyer ID and the medicinal product ID in the proposed system. It ensures the authenticity of the medicine consignment before it is received by the buyer. Figure 4 illustrates the working of the validation module, which validates a medicine consignment. It verifies both the buyer’s identity and the product’s identity, ensuring that only genuine buyers receive the medicine. Figure 5 shows the working of the validation module for validating medicine strips and the view module for retrieving the transaction history of medicines.

In the proposed system, every buyer is required to scan the QR code on the medical strip or consignment from their registered location using the supply chain Dapp. This scanning process serves the purpose of validating both the buyer’s identity and the product’s identity. The blockchain network traces the transportation of medicinal products and captures the location of each buyer through the smart location tracker. To purchase and validate medicines, consumers can only do so from a registered buyer location by providing their unique identifier, such as a government-issued identification number, to the retailer. This validation process ensures that only genuine buyers with proper access credentials can validate medicinal products and receive the consignment. If the buyer fails to meet the validation criteria, the transporter returns the consignment, as there may be concerns about potential counterfeiting.

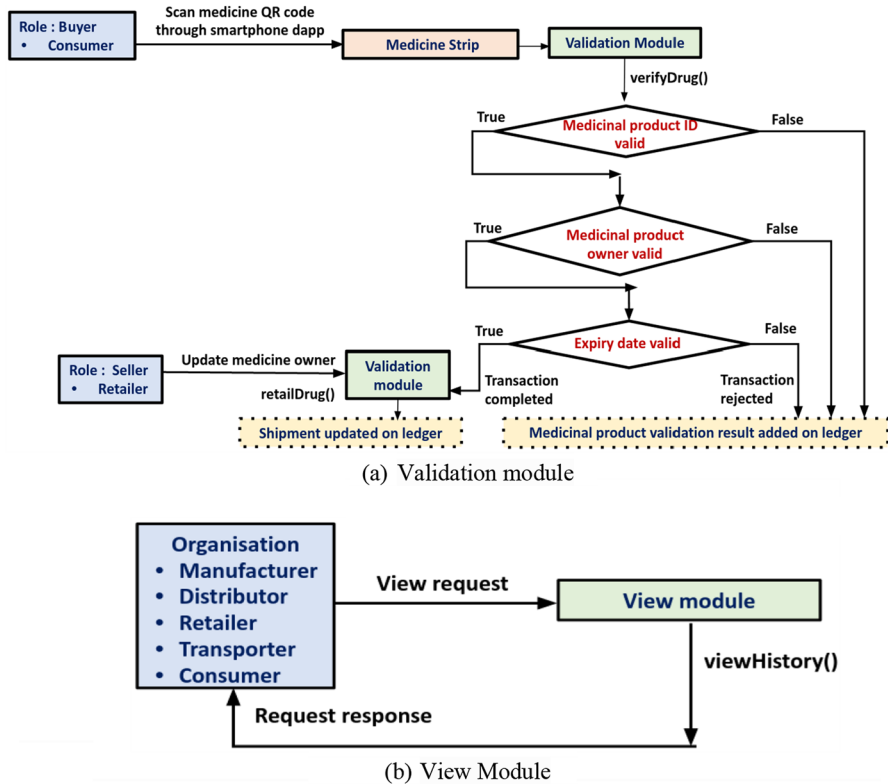


**Fig. 4** Working of the validation module to validate medicine consignment

The validation process for the ID of a medicinal product consists of three levels in the proposed system. Firstly, the product ID scanned by the buyer is verified by checking if it exists in the recorded transactions on the ledger. Secondly, if the product ID is valid, the second-level validation checks the current owner of the product. For example, if a consumer is purchasing medicine, the current owner should be the retailer. Lastly, the third-level validation verifies the expiration date of the medicinal product to ensure that expired products are not sold.

With these validation checks, if a medicinal product is deemed valid, the system will close the transaction upon successful sale. However, if the validation function returns false, indicating any failed checks, the system will reject the transaction due to possible counterfeiting. This empowers the consumer to verify the authenticity of medicinal products by scanning their IDs and helps prevent the circulation of counterfeit items. Algorithms 5 and 6 depict the chaincodes of the validation module for medicine consignment and medicine strip, respectively, with the following functionalities:

- **Verify consignment:** This function validates the consignment and can be invoked by a distributor or retailer. It includes fields such as the buyer's CRN, a list of assets (medical products), the current owner, the current expiry date, the current date, and the location. The output of this function is the validation result along with the verifier ID.
- **updateShipment:** This function updates the status of a shipment and is called by the transporter when the consignment is dispatched to a buyer. It includes fields such



**Fig. 5** Working of the validation module to validate the medicine strip and view module

as the buyer's CRN, drug name, and transporter's CRN. After the consignment is delivered, each medicinal product within the consignment is updated with the shipment information, and the owner field is also updated accordingly.

- **retailDrug:** This function allows the retailer to sell medicinal products to the consumer. It is invoked only by the retailer and includes fields such as the drug name, serial number, retailer's CRN, and the customer's UIN (unique identification number). The owner field of the corresponding drug is updated with the customer's UIN.
- **verifyDrug:** This function enables consumers to validate medicine strips. It contains fields such as the drug name, serial number, current owner, current expiry date, current date, and location. The function produces the validation result along with the verifier ID.



**Algorithm 5:** Algorithm of the validation chaincode module for medicine consignment

1. **INPUT:** buyer CRN, list Of Assets, current Owner, current expiry Date, current location
2. **function** VERIFY CONSIGNMENT (buyerCRN, list of Assets, currentOwner, current expiry date, currentDate, current location)
3.     fetch buyerID and buyer's registered location through buyer details
4.     **if** (buyerID != undefined || current location != buyer's registered location)
5.         **then**
6.             fetch all medicine details through list of assets
7.             **if** (allproductIDs != undefined)
8.                 **then**
9.                     All product IDs are invalid
10.                     **if** (all product IDs.owner == currentOwner) **then**
11.                         Genuine Medicinal products
12.                         **if** (allProducts.expiry >= currentDate) **then**
13.                             All medicinal products have valid expiry Date
14.                             Validation result "Valid Consignment" and verifier "buyerID"
15.                         **else**
16.                             Invalid medicinal products
17.                         **end if**
18.                     **else**
19.                         Invalid medicinal products IDs
20.                     **end if**
21.             **else**
22.                 invalid buyer
23.                 "Invalid consignment" and verifier "buyer ID"
24.             **end if**
25. **end function**
26. **OUTPUT:** Buyer identity and medicinal product validation result
27. **INPUT:** buyer CRN, drug Name, transporter CRN
28. **function** UPDATESHIPMENT (buyerCRN, drugName, transporterCRN)
29.     Fetch shipment details and drug details through buyer CRN and drugName
30.     Fetch transporterID through transporterCRN
31.     **if** (valid transporterID) **then**
32.         **return** update shipment status to 'delivered'
33.         Update shipment field and owner field in registered medicine details on the ledger
34.     **else**
35.         Invalid transporterID, buyer or drug details
36.     **end function**
37. **OUTPUT-** updated shipment details and owner of all medicine

### Algorithm 6:- Algorithm of the validation chaincode module for medicine strip

```

1. INPUT: drug Name. serial No. current Owner, current expiry Date, current
   Date, current location
2. function VERIFYDRUG (drugName, SerialNo, currentOwner,
   currentexpiryDate, currentDate, current location)
3.   fetch transaction under identify and store it with verifier Identify
   fetch medicinal product ID and its details through drug name and serial no.
4.   if (product ID!= undefined) then
5.     valid product ID
6.     if (productID.owner == currentOwner) then
7.       Genuine Medicinal product
8.       if (productid.expiry >= currentDate) then
9.         Valid medicinal product expiry date
10.        Validation result:“Valid Medicinal product” and verifier “consumer
        identity” updated on ledger.
11.      else
12.        Invalid medicinal product expiry date
13.      end if
14.    else
15.      medicine may be counterfeited
16.    end if
17.  else
18.    Validation result “Invalid medicine” and verifier “ consumer identify”
    updated on ledger.
19.  end if
20. end function
21. OUTPUT: Buyer identify and medicinal product validation result
22. INPUT: drug Name, serial NO, retailer CRN, customer UIN
23. function RETAILDRUG (drug Name, serialNo, retailer CRN, CustomerUIN)
24.   fetch retailerID through retailerCRN
25.   fetch medicine details through drug name and serial number
26.   if (valid retailerID) then
27.     Update owner field of the medicine details with consumer UIN
28.   else
29.     Invalid shipment
30.   end if
31. end function
32. OUTPUT- update the owner of sold medicine;

```

In addition to the validation module, the view module handles view requests from authorized users of any organization and retrieves the transaction history of medicines for transparency and traceability purposes. Algorithm 7 shows the chain code of the view module that has the following functionality:

- **viewHistory:** This function is defined to view the history of a medicinal product from its origin. It can be called by a user of any authorized organization. It contains fields such as drug name and serial no.

#### Algorithm 7: Algorithm of the view chaincode module

1. **INPUT:** Drug name, serial number
2. **function** VIEWHISTORY| drugName, serialNO
3.     Fetch entire history of medicine details through drug name and serial number
4. **end function**
5. **OUTPUT:** Medicinal product details and its history, buyerID, medicinal product validation result.

### 4.3 Process implementation

The proposed Hyperledger-based supply chain framework is developed and tested on a system with an Intel Core i3, 8th generation CPU. The operating system used is Ubuntu 20.04.2 LTS, a 64-bit version. The system is equipped with 12.00 GB of RAM to support the development and experimental processes (Table 2).

The implementation of the Fabric network is conducted in a Docker environment utilizing Docker-engine (version 19.03.15). Docker-compose (version 1.24.0) is employed for container and Docker image configuration. Hyperledger Fabric (v2.2.0) and Node (v10.19.0) are utilized for developing the Fabric–Node client SDK. Crypto-materials are generated using the crypto-config.yaml file, channel artifacts were generated using the configtx.yaml file, and docker-compose.yaml files are employed for running the Docker containers. The chain code and application files are built in JavaScript. Postman, a simulation software, is used for testing transaction proposals, with transaction requests written in JSON format for calls such as POST and GET.

The proposed supply chain in the Hyperledger Fabric environment consists of various network participants, including manufacturers, distributors, retailers, consumers, and transporters. Each participant has contributed two peers (peer 0 and peer

**Table 2** Summary of development environment

Name	Version
Ubuntu	20.04.2 LTS
Docker-engine	19.03.15
Docker-compose	1.24.0
Hyperledger Fabric	2.2.0
Node	10.19.0

Creating peer1.retailer.pharma-network...	Done						
Creating ca.distributor.pharma-network...	Done						
Creating ca.manufacturer.pharma-network...	Done						
Creating peer1.distributor.pharma-network...	Done						
Creating ca.code...	Done						
Creating peer1.manufacturer.pharma-network...	Done						
Creating peer2.distributor.pharma-network...	Done						
Creating peers.consumer.pharma-network...	Done						
Creating peer1.consumer.pharma-network...	Done						
Creating peer1.transporter.pharma-network...	Done						
Creating ca.consumer.pharma-network...	Done						
Creating ca.retailer.pharma-network...	Done						
Creating peer1.retailer.pharma-network...	Done						
Creating ca.transporter.pharma-network...	Done						
Creating peers.transporter.pharma-network...	Done						
Creating peer2.manufacturer.pharma-network...	Done						
Creating orderer.pharma-network...	Done						
Creating cll	Done						
CONTAINER ID	IMAGE	COMMAND	CREATED	STATUS	PORTS	NAME	
3cc59109da	hyperledger/fabric-tools:latest	"/bin/bash"	16 seconds ago	Up Less than a second		cli	
b3b4937652	hyperledger/fabric-peer:latest	"network"	About a minute ago	Up 17 seconds	0.0.0.0:15051->15051/tcp, 0.0.0.0:15053->15053/tcp	peer2.transporter.pharma-net	
19432ce0ba0	hyperledger/fabric-orderer:latest	"orderer"	About a minute ago	Up 25 seconds	0.0.0.0:7050->7050/tcp	orderer.pharma-network-c	
4dc0b15f817	hyperledger/fabric-peer:latest	"peer node start"	About a minute ago	Up 26 seconds	0.0.0.0:7051->7051/tcp, 0.0.0.0:7053->7053/tcp	peer2.manufacturer.pharma-net	
an1-network.com	hyperledger/fabric-ca:latest	"sh -c 'fabric-ca-se.."	About a minute ago	Up 41 seconds	0.0.0.0:11954->7854/tcp	ca.transporter.pharma-ne	
8837b72e3a3	hyperledger/fabric-peer:latest	"peer node start"	About a minute ago	Up 31 seconds	0.0.0.0:12051->12051/tcp, 0.0.0.0:12053->12053/tcp	peer1.retailer.pharma-ne	
8dc0053733c	hyperledger/fabric-peer:latest	"peer node start"	About a minute ago	Up 39 seconds	0.0.0.0:13051->13051/tcp, 0.0.0.0:13053->13053/tcp	peer2.consumer.pharma-ne	
8f5d3cd60f94	hyperledger/fabric-ca:latest	"sh -c 'fabric-ca-se.."	About a minute ago	Up 39 seconds	0.0.0.0:9054->7054/tcp	ca.retailer.pharma-netwo	

1) and one user. Peer 0 of all organizations serves as the anchor peer and endorsing peer. The architecture of the supply chain on the Hyperledger platform is depicted in Fig. 6. In the pre-setup phase of the Fabric network, crypto-materials were created for all elements, including peers, orderers, and certificate authorities (CAs), using the cryptogen tool. These crypto-materials consist of X.509 digital certificates for user identification. Channel artifacts were created using the configtxgen tool. During the final network setup phase, all the services required to run the Fabric network components were created on a local machine using Docker. Docker containers were initially created for all the peers of different organizations and their corresponding certificate authorities.

 Springer

From the CLI container, we perform the following tasks: creating channels, ensuring that all peers join the pharma channel, and updating the anchor peers for all organizations. The chain codes have been deployed on the endorsing peers to enable the registration of companies in different organizational roles, registration of drugs, ordering of drugs, shipping of drugs, validation, and viewing of drugs. Each peer within the network is aware of the other peers and their respective organizations through the identities issued by the Membership Service Provider (MSP). The MSP facilitates peer identification and issues credentials (public–private keys) to users who represent registered organizations from outside the network. Each organization in the network maintains its own certificate authority (CA). The manufacturer organization takes responsibility for configuring and starting the network. In this network, only one channel named "pharma channel" has been created. The ledger is maintained by the channel and is accessible only to the peers that have joined the channel. Every peer joins the pharma channel to access the channel's data and maintain a ledger.

The primary objective of the proposed blockchain-based system is to address the issue of medicine counterfeiting and ensure public health safety. To achieve this, the system employs restricted access controls, granting users only essential privileges to invoke chain code functions. Table 3 outlines the access controls for different organizations in the system, specifying the rights and permissions granted to various members of the supply chain. All authorized organizations can access the Fabric pharma channel network using their X.509 certificates (public–private keys) issued by the certificate authorities. Users belonging to the manufacturer organization have the authority to register a company, add new drugs, create shipments, and view the history of medicines. Users from distributor organizations can register a company, create purchase orders, create shipments, verify consignments, and view the history of medicines. Retailer organization users can register a company, create purchase orders, verify consignments, retail drugs, and view the history of medicines. Users of consumer organizations have the authority to verify medicines and view the history of medicines. Transporter organization users can register a company, update shipments, and view the history of medicines.

For the ordering service, we utilized a single-node ordering service named Solo consensus mechanism in our implementation. The client application initiates a transaction proposal by invoking smart contract functions. The Membership Service Provider (MSP) verifies the identity of the client to authorize their actions. The transaction proposal is then sent to the endorsers, who have deployed the chain codes. In order to achieve consensus, the endorsers simulate the transaction and determine whether it is valid based on the business logic of the chaincode. The ordering service captures these endorsement transactions, creates blocks, and broadcasts them to all peers on the network. Committer peers are responsible for transaction validation. The transaction details are propagated through the gossip protocol on the pharma channel Fabric network, allowing all peer nodes to validate them. Once validation is completed, transaction blocks are updated on the ledger of each committer's peer. Finally, the committer sends an asynchronous response back to the client application after executing the transaction. A transaction in Hyperledger Fabric is considered

**Table 3** Access privileges of organizations in proposed system

Organization	Users role	Access privileges	How accessed
Manufacturer Distributor	Seller	Register company, addDrug, create shipment, view history	Using public-private key
	Both seller and buyer	Register company, create purchase order, create shipment, verify consignment, view history	Using public-private key
Retailer	Both seller and buyer	Register company, create purchase order, verify consignment, retail drug, view history	Using public-private key
Consumer Transporter	Buyer	Verify drug, view history	Using public-private key
	Shipment handler	Register company, update shipment, view history	Using public-private key

final once it has been endorsed, ordered, and validated. This indicates that it has been completed in accordance with the chaincode guidelines and recorded on the ledger. The consensus mechanism ensures that all participating nodes agree on the final state of the ledger.

#### 4.4 Proposed blockchain-based watermarking (BCW)

This paper presents a secure QR code watermarking algorithm based on blockchain technology. The algorithm provides two levels of security:

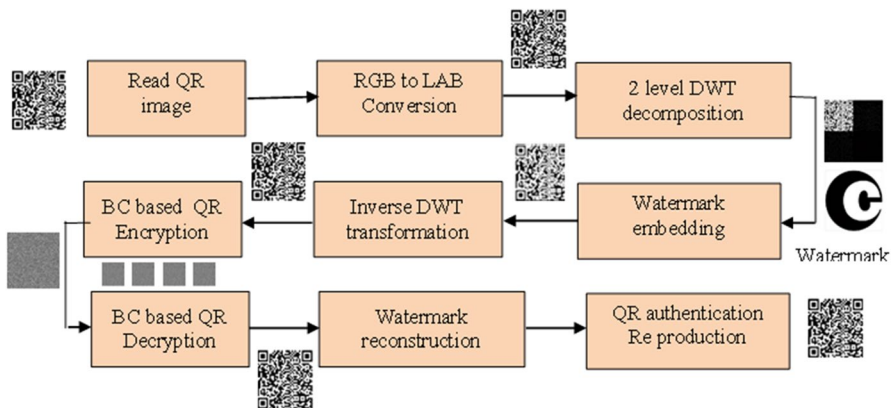
1. Embedding a watermark logo into the QR code: This level of security adds an additional visual identifier to the QR code, making it more resistant to counterfeiting and tampering.
2. Blockchain-based encryption/decryption using SHA256: The algorithm proposes using blockchain technology to encrypt and decrypt the QR code data, ensuring anti-counterfeiting measures and enhancing overall security.

Figure 8 illustrates the block diagram of the proposed blockchain-based watermarking system (BCW system).

##### 4.4.1 BCW embedding

This work proposes an undetectable ED-based QR code watermarking that adheres to the Sobel ED technique. The appropriate X and Y gradients masking are used to determine the ED, and they are;

$$F = [(Z7 + Z8 + Z9) - (Z1 + Z2 + Z3)] + [(Z3 + Z6 + Z9) - (Z1 + Z4 + Z6)] \quad (1)$$



**Fig. 8** Proposed block diagram of the blockchain-based QR code watermarking for supply chain



The DWT-LL sub-bands edge coefficient, as  $f(u, v)^{ed}$ , was suggested by the present watermark insertion algorithm. In order to determine the dilation coefficient  $f(u, v)^{di}$  of  $2 \times 2$  masks across edge components  $f(u, v)^{ed}$ , the difference between the dilatation and edge elements, which results in the reduced features, is used to generate the local invisible watermark as follows:

$$g(u, v) = f(u, v)^{di} - f(u, v)^{ed} \quad (2)$$

On the basis of careful scaling parameter estimation, which should be adjusted to 0.9 according to the best learning from this research, the watermark's invisibility could be maximized. As a 1st order scaled improved watermarking rule is discovered;

$$W_T(x, y) = (1 - \alpha) * (g(u, v)) + \alpha N_{x,y} \quad (3)$$

Let us to describe as  $LL_{x,y}^f$ , a LL coefficient, DWT's second-level feature coefficient  $N_{x,y}$  is utilized as AWGN noise that was added. Let  $W_E$  is embedded watermark as

$$W_E = LL_{x,y}^f + W_T(x, y) \quad (4)$$

Lastly, an inverse wavelet transform is used to produce the watermarked QR image. Figure 8 depicts the suggested block flow of the secured watermark embedding with decryption using DWT-blockchain. In a recent publication, the HH coefficient was changed to the LL coefficient, and watermarking is part of the LAB space.

The watermark has been mathematically defined as an additive signal or content that would be actually stored as a concealed watermark message  $b$  inside the boundaries of established undetected distortions graphically provided by mask  $M$ . This is expressed as follows:

$$Ow = x + w(M) \quad (5)$$

here  $x$  is original image or QRC

$$b = w(M) \quad (6)$$

#### 4.4.2 Blockchain encryption using SHA 256 hash code

The 256-bit cryptographic hash function used by the SHA 256 oriented hash algorithm essentially comprises two phases. To meet the dimensions of blockchain-based encryption as well as the genesis packet size in the initial step, the given image is automatically scaled to  $512 \times 512$ . Then, 512 bits are combined with input vector  $X$  by each chain block having dimensions of 32 bits. Each block of 512 bits in the second stage is individually represented by;

$$[X^{(1)}, X^{(2)}, \dots X^{(m)}]. \quad (7)$$

Message block numbers are then determined consecutively for 64 words or 32-bit apiece as follows:

$$H^k = H^{k-1} + C_{S^{(k)}}^f * H^{k-1} \quad (8)$$

These hash blocks are shuffled to produce the encrypted QR codes, which are placed on the outbound product in the supply chain.

## 5 Simulation results

This section presents the test results of the different chain code modules implemented in the proposed system. The simulation results of selected chain code functions are provided through snapshots. Clients interact with the system using Dapp APIs, which are invoked through the Postman software. Transaction requests are made in JSON format using POST and GET calls. The simulation results are divided into two parts. The first part focuses on simulating the security and traceability aspects of the Hyperledger chain code. The verification results obtained from these simulations are presented. In the second part, the simulation results of QR code watermarking using blockchain encryption are presented.

### 5.1 Test results of proposed Hyperledger Fabric design chaincodes

Firstly, we test a register company API that has been used to enroll various organizations on the ledger. Figure 9 shows the post call of the Dapp for registering the manufacturer company with the following details: name, CRN number, and location. Similarly, we have added and tested other organisations in either of the following roles—distributor, retailer, or transporter.

Figure 10 shows the backend of the Dapp listening to the client application and registering a new company. After testing the register company API, we test the addDrug API used to register a new drug only by registered manufacturers on the blockchain network.

Figure 11 shows the post call to the Dapp for registering a drug named Paracetamol with the following details: serial number, manufacturing date, expiry date, company CRN, and organization role.

Figure 12 shows the backend of the Dapp to create the drug asset with the following details: product ID, name, manufacturer, manufacturing date, expiry date, owner, and shipment. Next, we are testing a create purchase order API, which can be used by authorized buyer organisations to buy medicines.

Figure 13 shows the post call to the Dapp for creating a new purchase order with the following details: buyer CRN, seller CRN, drug name, quantity, and organization

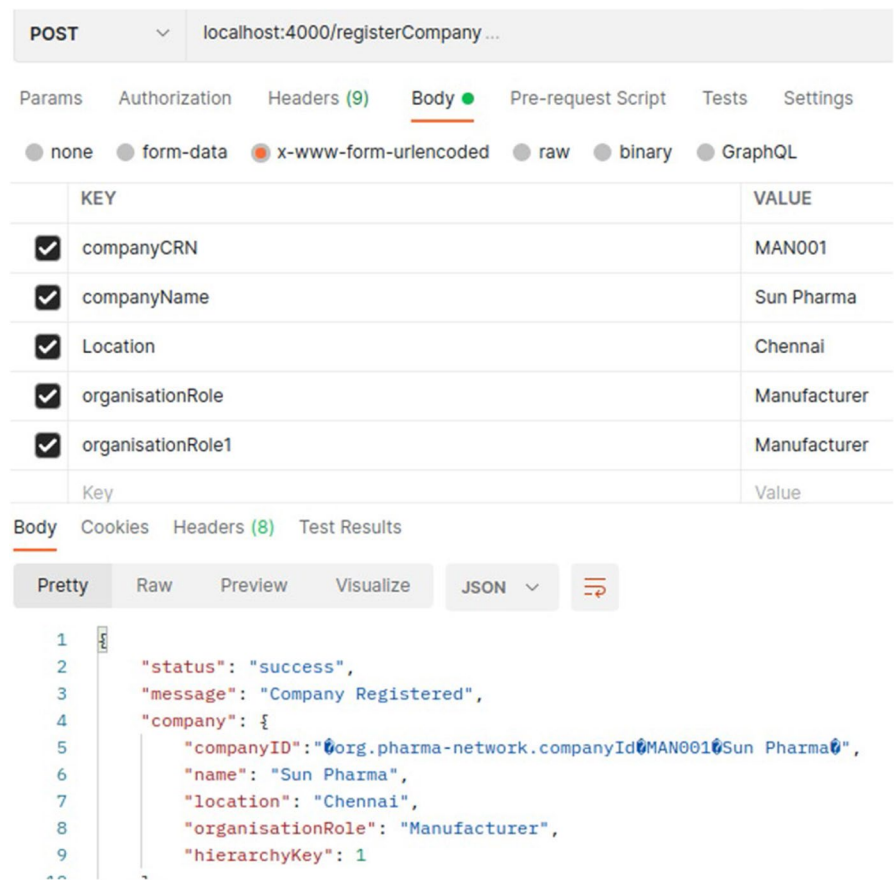


Fig. 9 Post call to register manufacturer company

role. Next, we are testing a create shipment API, which can be used by authorized seller organizations to create shipments for medicine consignments.

Figure 14 shows the backend of the Dapp to create a shipment with the following details: shipment ID, creator, list of assets, transporter, and status. Next, we test a retail drug API used to retail medicines by authorized retailers.

Figure 15 shows the post call to the Dapp for retailing medicine with the following details: drug name, serial number, retailer CRN, customer, and UIN (used Aadhar number).

Lastly, we are testing a view drug API used to view a history of medicines by any permissioned user. Figure 16 shows the backend of the Dapp that retrieved the history of medicines with the following details: drug name and serial number.

```

neetu@neetu-ubuntu:~/workspace/pharma-net/application$ node index.js
Drug counterfeiting App listening on port 4000!
Distributor identity added to wallet.
Retailer identity added to wallet.
Consumer identity added to wallet.
Transporter identity added to wallet.
Manufacturer identity added to wallet.
.....Connecting to Fabric network Gateway
.....Connecting to channel - pharma-channel
.....Connecting to pharma-net Smart Contract
New company registration request
Processing request
{
  companyID: '\x00org.pharma-network.companyId\x00MAN001\x00Sun Pharma\x00',
  name: 'Sun Pharma',
  location: 'Chennai',
  organisationRole: 'Manufacturer',
  hierarchyKey: 1
}
New company is registered
Disconnect from fabric network
.....Disconnecting from Fabric Gateway
Registering a Company
.....Connecting to Fabric network Gateway
.....Connecting to channel - pharma-channel
.....Connecting to pharma-net Smart Contract
New company registration request
Processing request
{
  companyID: '\x00org.pharma-network.companyId\x00TRA001\x00FedEx\x00',
  name: 'FedEx',
  location: 'Delhi',
  organisationRole: 'Transporter'
}

```

Fig. 10 Backend of the Dapp to register new company

## 5.2 Result of BCW

The results of the proposed ED wavelet domain invisible watermarking combined with blockchain-based encryption, without any attacks, for QR codes are presented in Fig. 17. The QR image undergoes watermarking using the DWT–SVD–HD method and subsequently encryption using a hash algorithm to enhance security, as depicted in Fig. 17. This figure highlights the significant contributions of the BCW-based QR code authentication method.

Figure 18 illustrates the sequential analyses of the crypto weights in terms of histograms. It can be observed that due to the QR image's limited color features, the histograms exhibit a high level of correlation. The encrypted information demonstrates a well-equalized (flat) histogram, indicating the quality of the encryption standard. Figure 19 shows the results of watermark extractions under various attacks, including noisy, filter, and motion blur. In this experiment, the attacks are applied to the blockchain-encrypted QR information. Since the histograms are already flat and exhibit closely related patterns, the proposed method performs consistently well across all attacks. These results are further supported by quantitative evaluations.

POST

localhost:4000/addDrug

Params

Authorization

Headers (9)

Body

Pre-request Script

Tests

Settings

none

form-data

x-www-form-urlencoded

raw

binary

GraphQL

<input type="checkbox"/>	productName	Paracetamol
<input checked="" type="checkbox"/>	serialNo	001
<input checked="" type="checkbox"/>	mfgDate	JAN2020
<input checked="" type="checkbox"/>	expDate	DEC2022
<input checked="" type="checkbox"/>	companyCRN	MAN001
<input checked="" type="checkbox"/>	organisationRole	Manufacturer
	Key	Value

Body

Cookies

Headers (8)

Test Results

Pretty

Raw

Preview

Visualize

JSON

```

1  {
2    "status": "success",
3    "message": "Drug Added successfully",
4    "drug": {
5      "productId": "org.pharma-network.productIdKey0001Paracetamol",
6      "name": "Paracetamol",
7      "manufacturer": "org.pharma-network.companyIdMAN001Sun Pharma",
8      "manufacturingDate": "JAN2020",
9      "expiryDate": "DEC2022",
10     "owner": "org.pharma-network.companyIdMAN001Sun Pharma",
11     "shipment": ""

```

Fig. 11 Post call to the Dapp for registering a new drug

```

New Drug Add request
Processing request
{
  productId: '\x00org.pharma-network.productIdKey\x00001\x00Paracetamol\x00',
  name: 'Paracetamol',
  manufacturer: '\x00org.pharma-network.companyId\x00MAN001\x00Sun Pharma\x00',
  manufacturingDate: 'JAN2020',
  expiryDate: 'DEC2022',
  owner: '\x00org.pharma-network.companyId\x00MAN001\x00Sun Pharma\x00',
  shipment: ''
}
New Drug is added

```

Fig. 12 Backend of the Dapp to create a drug asset

POST localhost:4000/createPO

Params Authorization Headers (9) **Body** Pre-request Script Tests Settings

none form-data x-www-form-urlencoded raw binary GraphQL

KEY	VALUE
<input checked="" type="checkbox"/> buyerCRN	RET002
<input checked="" type="checkbox"/> sellerCRN	DIST001
<input checked="" type="checkbox"/> drugName	Paracetamol
<input checked="" type="checkbox"/> quantity	2
<input checked="" type="checkbox"/> organisationRole	Manufacturer
Key	Value

Body Cookies Headers (8) Test Results

Pretty Raw Preview Visualize JSON

```

1  {
2    "status": "success",
3    "message": "Purchase order created successfully",
4    "purchaseOrder": {
5      "poID": "org.pharma-network.poIDKeyRET002Paracetamol",
6      "drugName": "Paracetamol",
7      "quantity": "2",
8      "buyer": "org.pharma-network.companyIdRET002upgrad",
9      "seller": "org.pharma-network.companyIdDIST001VG Pharma"

```

Fig. 13 Post call to the Dapp for creating a new purchase order

```

Purchase Order Created
Disconnect from fabric network
....Disconnecting from Fabric Gateway
Creating Purchase Order
....Connecting to Fabric Gateway
....Connecting to channel - pharmachannel
....Connecting to PHARMANET Smart Contract
New Shipmentrequest
Processing request
{
  shipmentID: '\x00org.pharma-network.shipmentKey\x00DIST001\x00Paracetamol\x00'
  creator: '\x00org.pharma-network.companyId\x00MAN001\x00Sun Pharma\x00',
  assets: [
    '\x00org.pharma-network.productIDKey\x00001\x00Paracetamol\x00',
    '\x00org.pharma-network.productIDKey\x00002\x00Paracetamol\x00',
    '\x00org.pharma-network.productIDKey\x00003\x00Paracetamol\x00'
  ],
  transporter: '\x00org.pharma-network.companyId\x00TRA001\x00FedEx\x00',
  status: 'in-transit'
}

```

Fig. 14 Backend of the Dapp to create a shipment



**POST** localhost:4000/retailDrug

Params Authorization Headers (9) **Body** Pre-request Script Tests Settings

none form-data **x-www-form-urlencoded** raw binary GraphQL

KEY	VALUE
drugName	Paracetamol
serialNo	001
retailerCRN	RET002
customerAadhar	AAD001
organisationRole	Retailer

Key Value

Body Cookies Headers (8) Test Results

Pretty Raw Preview Visualize JSON

```

1  {
2    "status": "success",
3    "message": "Drug details updated successfully",
4    "drug": {
5      "productID": "org.pharma-network.productIDKey00010Paracetamol0",
6      "name": "Paracetamol",
7      "manufacturer": "org.pharma-network.companyId0MAN0010Sun Pharma",
8      "manufacturingDate": "JAN2020",
9      "expiryDate": "DEC2022",
10     "owner": "AAD001",
11     "shipment": "org.pharma-network.shipmentKey0RET0020Paracetamol0"

```

Fig. 15 Post calls to the Dapp for retailing medicine

### 5.3 Performance evaluation of proposed Hyperledger-based design

For measuring the performance of a developed Hyperledger-based design, the Hyperledger Caliper 0.4.2 framework has been used. The Caliper tool gives performance metrics such as success rate, throughput, latency, and resource consumption (e.g., CPU, memory, and IO) for a system under test. The performance metrics are as follows:

*Execution time:* This is the time between the point of transaction request submission and execution.

*Throughput:* It is the ratio of a total successful transaction to the total time duration in seconds.



```

View Drug History processed
Disconnect from fabric network
....Disconnecting from Fabric Gateway
View history of transaction on the drug
....Connecting to Fabric Gateway
....Connecting to channel - pharmachannel
....Connecting to PHARMANET Smart Contract
View Drug current state Initialized
Processing View Drug current state
{
  productID: '\x00org.pharma-network.productIDKey\x00001\x00Paracetamol\x00',
  name: 'Paracetamol',
  manufacturer: '\x00org.pharma-network.companyId\x00MAN001\x00Sun Pharma\x00',
  manufacturingDate: 'JAN2020',
  expiryDate: 'DEC2022',
  owner: 'AAD001',
  shipment: '\x00org.pharma-network.shipmentKey\x00RET002\x00Paracetamol\x00'
}

```

Fig. 16 Backend of the Dapp to retrieve history of medicine

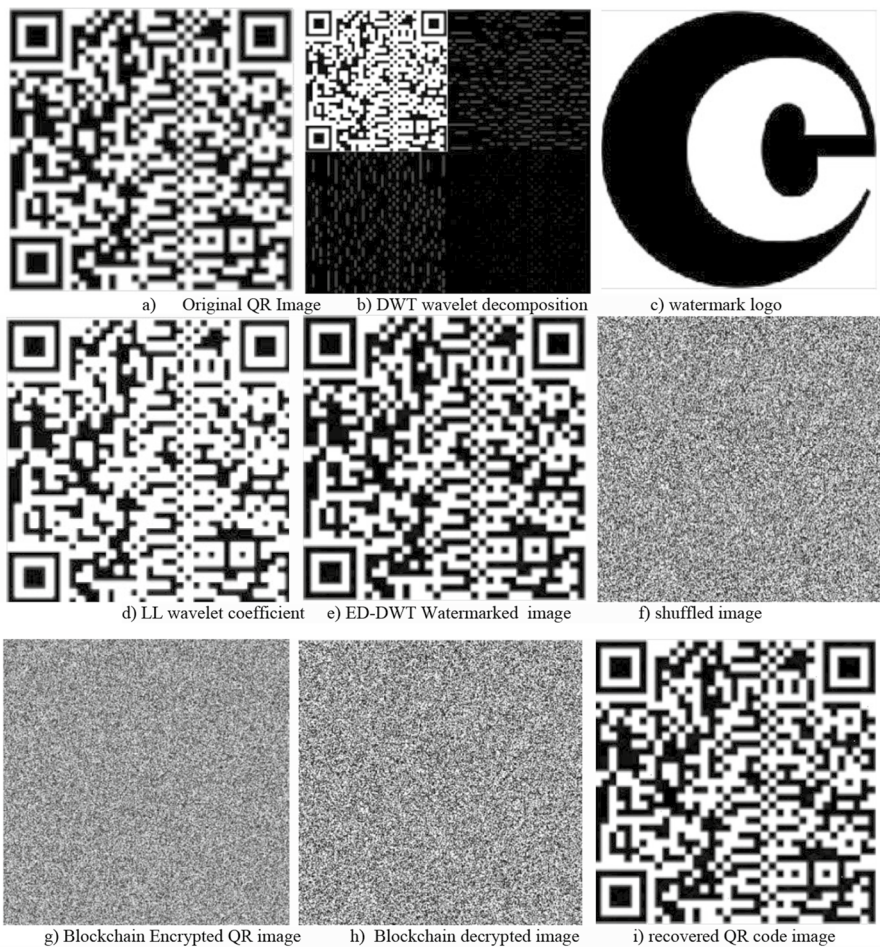
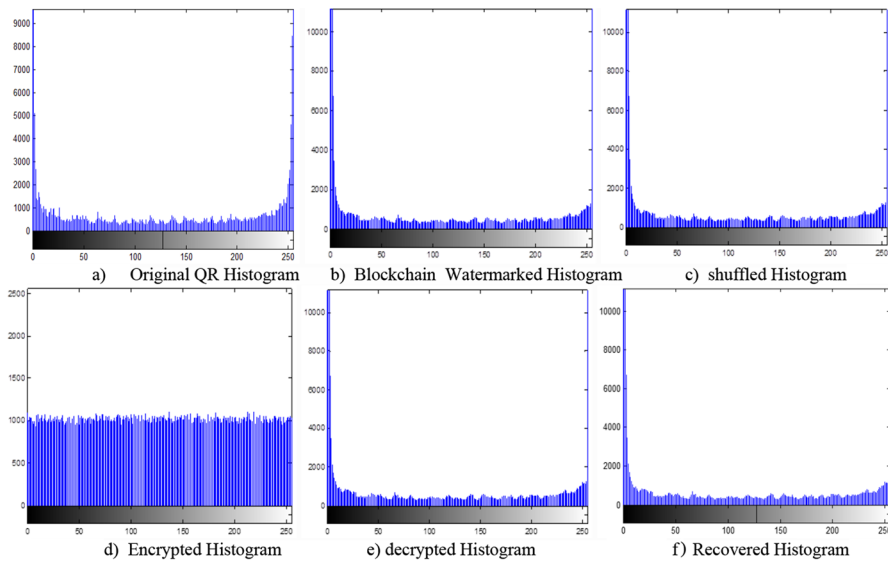
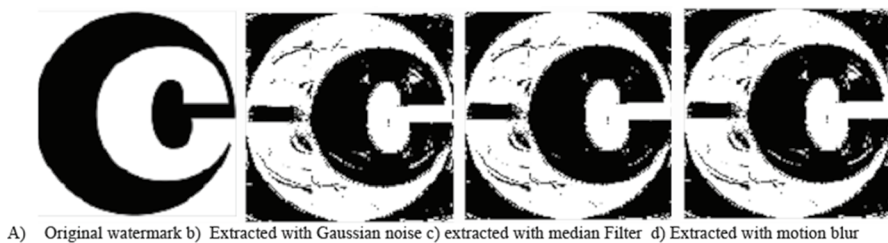


Fig. 17 Sequential results of blockchain-based QR code BCW authentication process



**Fig. 18** Crypto weight histograms for the BCW process



**Fig. 19** Watermark extractions under noisy, filter, and motion blur attacks

```

2022.02.13-20:26:05.069 info [caliper] [round-orchestrator] Finished round 1 (addDrug) in 241.243 seconds
2022.02.13-20:26:05.069 info [caliper] [monitor.js] Stopping all monitors
2022.02.13-20:26:05.169 info [caliper] [report-builder] ### All test results ###
2022.02.13-20:26:05.172 info [caliper] [report-builder]

```

Name	Succ	Fall	Send Rate (TPS)	Max Latency (s)	Min Latency (s)	Avg Latency (s)	Throughput (TPS)
addDrug	100000	0	417.5	0.20	0.00	0.03	417.5

```

2022.02.13-20:26:05.238 info [caliper] [report-builder] Generated report with path /home/neetu/workspace/caliper-workspace/report.html

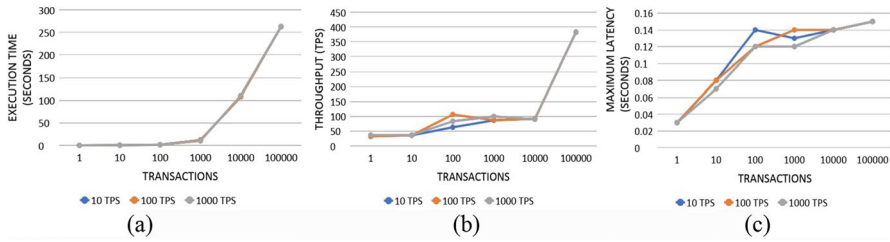
```

**Fig. 20** Performance testing results for 100,000 transactions using the Caliper tool

**Latency:** This is the time between the point of transaction request submission and the network response.

**Resource metrics:** They can be measured by monitoring the CPU, memory, and network I/O consumed by the blockchain system under test.

We evaluated the performance of the proposed system by measuring execution time, throughput, latency, and resource statistics. Our testing involved up to 100,000 transactions and 20 peers. To improve scalability and performance metrics, we have



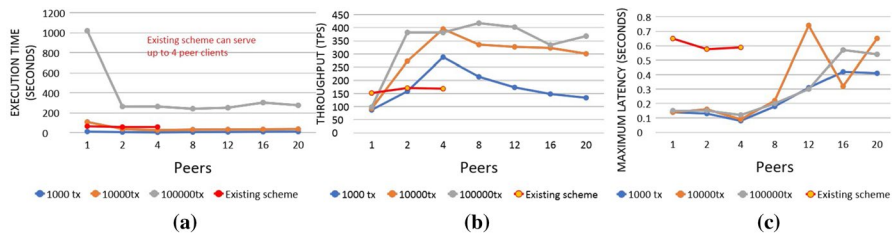
**Fig. 21** Quantitative performance analysis for anti-counterfeiting based on the Hyperledger for different numbers of transactions **a** Execution time, **b** throughput curves, and **c** calculation of maximum latencies

configured the block size to 10 MB and implemented the event sourcing technique in the chaincode. This technique ensures that only the differential changes in asset state are stored on the ledger, optimizing storage efficiency. The highest achieved throughput was 417.5 transactions per second (TPS), which was attained with 100,000 transactions and eight peers. The Caliper report for this result is illustrated in Fig. 20.

We analyzed the performance of the proposed system by conducting tests with varying numbers of transactions, ranging up to 100,000, using a single peer. The metrics evaluated include execution time, throughput, maximum latency, minimum latency, and average latency. The results of these analyses are depicted in Fig. 21.

We conducted simulations with transaction rates (TPS) set at 1, 10, and 100 to analyze the performance of the system. We observed that the execution time is lower for a smaller number of transactions and increases as the number of transactions grows. The throughput is initially low for smaller transaction numbers but rapidly increases for transactions exceeding 10,000. The maximum latency is higher for a smaller number of transactions and decreases as the number of transactions increases. It shows slight variation at different TPS settings for transactions ranging from 10 to 1000. Furthermore, we examined the system's performance by considering different numbers of peers and evaluating the same metrics, as depicted in Fig. 22.

We conducted simulations with transaction numbers set at 1000, 10,000, and 100,000 to analyze the system's performance. Additionally, we compared the



**Fig. 22** Performance analysis for different numbers of peers for anti-counterfeiting based on Hyperledger for different numbers of transactions **a** Execution time, **b** throughput curves, and **c** calculation of maximum latencies

**Table 4** Resource statistics of the proposed system

Name	CPU% (max)	CPU% (avg)	Memory (max) [MB]	Memory (avg) [MB]	Traffic in [MB]	Traffic out [MB]
dev-peer0.manufacturer.pharma-network.com	17.54	5.94	63.9	63.9	1.65	0.543
peer0.manufacturer.pharma-network.com	9.85	5.76	194	194	2.16	3.23
orderer.pharma-network.com	1.55	0.52	58.8	58.7	0.0827	0.141

performance of an existing scheme (Scheme 24) with 10,000 transactions. In our proposed scheme, when there are only a few peers, the execution time is significantly higher. However, with a sufficient number of peers, the execution time becomes stable and low. The execution time is slightly higher for 100,000 simultaneously invoked transactions compared to 1000 and 10,000 transactions. The throughput is low for fewer peers, reaches its highest point at eight peers, and then decreases. The highest throughput was achieved with 100,000 transactions. The latency remains consistently low up to eight peers and then varies with the number of transactions.

To assess resource consumption, we measured CPU-max, CPU-avg, Memory-max, Memory-min, Traffic in, and Traffic out, as shown in Table 4.

The peers in the system exhibited an average CPU utilization of 5.94% with a maximum of 17.54%. The maximum memory utilization was measured at 63.9 MB, and the average memory utilization was 90.5 MB. For the manufacturer peers, the average CPU utilization was approximately 5.76%, with a maximum of 9.85%. The average and maximum memory consumption for the manufacturer nodes were 194 MB. Regarding the orderer node, the maximum CPU utilization was 1.55%, and the average was 0.52%. The maximum memory consumption for the orderer node was 0.827 MB, with an average of 0.141 MB. Overall, the resource consumption of the proposed system is relatively low.

#### 5.4 Performance evaluation of BCW

The watermarking performance is evaluated based on the statistical parameters, including structural similarity and normalized correlation.

**PSNR:** The parameter is defined as the peak signal-to-noise ratio and is mathematically defined as follows:

$$\text{PSNR} = 10 \cdot \log_{10} \left( \frac{\text{MAX}_I^2}{\text{MSE}} \right) \quad (9)$$

**NC:** Normalized correlation (NC), which is what this quantity is called, should be higher and closer to unity in order to adequately represent the robustness of a watermarked standard against various attacks. Mathematically, this parameter is defined as follows;

$$\text{NC} = \sum_{i,j} d(i,j) * d(i-t, j-t) \quad (10)$$

where  $S = m \times n$  is the size of the frame, and  $d$  is the decision variable, 0 (equal) or 1 (not equal).

**SSMI:** It is defined as the structural similarity measure index. It represses the quality of the invisible watermarking.

$$\text{SSIM}(A, B) = \frac{\mu_A \mu_B + c_1}{\mu_A^2 + \mu_B^2 + c_1} + \frac{\sigma_{AB} + c_2}{\sigma_A^2 + \sigma_B^2 + c_2} \quad (11)$$

**Table 5** NC comparison for QR code watermarking for blockchain encryption under various attacks

Images	Variance	Gaussian noise	Speckle noise	Salt and pepper	Motion blur
QR image 1	0.001	0.6610	0.6639	0.6628	0.666
QR image 2		0.6621	0.6632	0.6635	0.662

**Table 6** Qualitative comparison for watermarked image exposed under various attacks

Images	Variance	Gaussian noise	Speckle noise	Salt and pepper	Motion blur
PSNR in dB	0.001	30.54610	33.68161	32.37145	26.86267
NC		0.9990	0.9995	0.9019	0.9975
SSIM		0.8480	0.9396	0.9663	0.8986

where  $\sigma_{AB}$  is the covariance between the  $A$  and  $B$  matrix, and  $c_1$  and  $c_2$  are factors that can be used to stabilize a division with a low denominator.

The quantitative evaluation of BCW is presented in Tables 5 and 6, respectively. The NC comparison for QR code watermarking for attacks applied to blockchain encryption is given in Table 5. It can be observed that the method performs equally for all cases.

Table 6 reflects the actual accuracy of the proposed method and compares the parametric performance of the watermark attacks applied to the BCW image. It is observed from Table 5 that the PSNR average is 30.86546 dB, and the average NC is 0.974475 for all attacks. The NC of 0.9995 is achieved for the most expected speckle noise attack. The best structure similarity of 0.9663 is obtained for the salt and pepper noise.

The results demonstrate achievements, such as a higher average peak signal-to-noise ratio (PSNR) of 30.86546 dB and an average normalized correlation (NC) of 0.974475 across all attacks. A remarkable NC value of 0.9995 further validates the accuracy and effectiveness of our proposed method.

## 5.5 Comparative analysis

In this section, we present a comparison of the performance and security features of our proposed system with existing work. We specifically compare our performance metrics with those of [1, 26–30]. The evaluation in [1] focused on user-based analysis rather than transaction numbers, which limits its usability. Additionally, their resource consumption was relatively high. In contrast, we have extensively tested our design with 100,000 transactions and 20 peer peers, achieving successful results and maintaining reasonable resource consumption. It is worth noting that the testing in [23] was limited to a maximum of 250 transaction rates and was computationally expensive. In contrast, we have successfully tested our proposed system up to 1000 TPS, demonstrating its scalability and improved performance. Regarding (24), their performance evaluation was limited to 10,000 transactions, and their system



was unable to serve more than 4 users within this range. They achieved a throughput of 168.3 TPS and a latency of 58.762 s with 10,000 transactions and four peer peers. In comparison, our suggested design achieved a higher throughput of 395.2 and a significantly lower latency of 0.09 s with the same number of transactions and peers. This demonstrates the superior performance and scalability of our proposed scheme.

The work presented in [28], is less scalable and privacy-preserving due to the use of the public blockchain. The performance evaluation is limited to 600 transactions, with maximum throughput up to 60 TPS and latency up to 1600 ms. Also, the use of RFID reduces the range of product traceability. In contrast, the proposed work utilizes the Hyperledger framework and IoT technology to improve privacy, scalability, and real-time monitoring over longer distances. Moreover, the performance of the proposed work is evaluated up to 100,000 transactions, with maximum throughput up to 417.5 TPS and latency up to 0.15 s. Further, we also evaluated performance metrics such as execution time, throughput, and latency in terms of the number of peers. The work proposed in [29], is also less scalable and privacy-preserving due to the use of the public blockchain. The work presented in [30], is not focused on product traceability. Also, the performance evaluation is limited to 1000 transactions. In comparison with these two Ethereum-based works, the proposed Hyperledger-based design offers more scalability and better performance.

Furthermore, we have compared the security and other features of our proposed scheme with relevant existing works, including [1, 3, 20, 22, 26–30]. A comparison is provided in Table 7, presenting the advantages of our proposed scheme over prior works.

A few of the existing schemes are Ethereum-based but the proposed scheme is Hyperledger Fabric blockchain-based to enhance privacy and scalability. Few of the existing systems lack design algorithms, but we have presented the architecture and algorithms of the proposed system, and within the algorithms, access controls are used to enhance security. A few of the existing systems also lack product traceability, but we have used the location tracker to record the location of the product on the blockchain. In order to validate the identity of the buyer, we have included the buyer

**Table 7** Comparison of proposed scheme with prior works

Security features	[1]	[3]	[20]	[22]	[26]	[27]	[28]	[29]	[30]	Proposed
Hyperledger-based	✓	✓	×	✓	✓	✓	✓	×	×	✓
Privacy	✓	✓	×	✓	✓	✓	✓	×	×	✓
Algorithm	×	✓	×	×	✓	✓	✓	✓	✓	✓
Architecture	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Access control	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Tracking	×	✓	✓	✓	×	✓	✓	✓	✓	✓
Authentication	×	×	×	✓	×	×	✓	✓	✓	✓
Performance testing	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Scalability	×	✓	×	×	✓	✓	✓	✓	✓	✓
Product validation	×	×	×	×	×	×	×	×	×	✓
Blockchain watermarking	×	×	×	×	×	×	×	×	×	✓



authentication mechanism, which is missing in most of the existing work. Most of the existing works evaluated the performance of their designed systems, which were limited to 1000 transactions, but the performance of the proposed system has been evaluated up to 100,000 transactions. Also, we have achieved performance and scalability improvements in terms of throughput, latency, and resource consumption. Moreover, we have included a product validation mechanism that enables buyers to validate the product by scanning the QR code of the product. This was not done in most of the existing works.

Furthermore, we have conducted additional evaluations of the security layer for QR authentication, specifically focusing on BCW and its resilience against watermarking attacks. The results demonstrate achievements, such as a higher average peak signal-to-noise ratio (PSNR) of 30.86546 dB and an average normalized correlation (NC) of 0.974475 across all attacks. A remarkable NC value of 0.9995 further validates the accuracy and effectiveness of our proposed method.

## 6 Conclusions

The paper introduces a blockchain-based multilayer security algorithm design for Hyperledger-based medicine supply chain management using blockchain IoT technology. Its aim is to address the issue of medicine counterfeiting by proposing a decentralized, tamper-proof, transparent, traceable, and validation-driven system. A key feature of the proposed scheme is the concept of a smart location tracker, which verifies the receiver's location and enables tracking capabilities.

Authorized users are empowered to validate medicinal products and access their complete transaction history. The system incorporates various chaincode modules with novel algorithms and access privilege restrictions to manage the supply chain workflow effectively. A validation module with multiple checks is designed, allowing buyers to authenticate medicinal products in real time by scanning their QR codes using a smartphone Dapp. The blockchain technology enables the detection of copied or fake QR codes, ensuring buyer identity validation to ensure only genuine buyers receive the products. The paper includes a demonstration of chaincode algorithms and simulation results of the supply chain.

Furthermore, the paper proposes an additional security mechanism utilizing QR code authentication with blockchain-based encryption and watermarking. A secure, encrypted QR code is proposed for anti-counterfeiting purposes in the supply chain. The performance metrics of the proposed system, including execution time, throughput, latency, and resource statistics, were measured using the Caliper benchmarking tool. The results indicate that the proposed system can successfully handle up to 100,000 transactions with optimal performance. The achieved throughput of 417.5 TPS with 100,000 transactions and eight peers

surpasses existing schemes, demonstrating superior security features and efficient resource utilization, resulting in high throughput and low latency.

**Author contributions** NS helped in conceptualization, methodology, software, writing—original draft preparation, visualization, and investigation. Rajesh Rohilla worked in supervision, validation, and writing—reviewing and editing.

## Declarations

**Conflict of interest** Both authors declare that they have no conflict of interest.

## References

1. Jamil F, Hang L, Kim K, Kim D (2019) A novel medical blockchain model for drug supply chain integrity management in a smart hospital. *Electronics* 8(5):505. <https://doi.org/10.3390/electronics8050505>
2. Uddin M (2021) Blockchain medledger: hyperledger fabric enabled drug traceability system for counterfeit drugs in pharmaceutical industry. *Int J Pharm* 597:120235
3. Liu X, Barenji AV, Li Z, Montreuil B, Huang GQ (2021) Blockchain-based smart tracking and tracing platform for drug supply chain. *Comput Ind Eng* 161:107669
4. Badhotiya GK, Sharma VP, Prakash S, Kalluri V, Singh R (2021) Investigation and assessment of blockchain technology adoption in the pharmaceutical supply chain. *Materials Today: proceedings* 46:10776–10780
5. Yiu NC (2021) Toward blockchain-enabled supply chain anti-counterfeiting and traceability. *Future Internet* 13(4):86
6. Yong B, Shen J, Liu X, Li F, Chen H, Zhou Q (2020) An intelligent blockchain-based system for safe vaccine supply and supervision. *Int J Inf Manag* 52:102024
7. Kshetri N (2021) Blockchain and sustainable supply chain management in developing countries. *Int J Inf Manag* 60:102376
8. Niu B, Mu Z, Cao B, Gao J (2021) Should multinational firms implement blockchain to provide quality verification? *Transp Res Part E: Logistics and Transp Rev* 145:102121
9. Niu B, Dong J, Liu Y (2021) Incentive alignment for blockchain adoption in medicine supply chains. *Transp Res Part E Logist Transp Rev* 152:102276
10. Mettler M (2016) Blockchain technology in healthcare: the revolution starts here. In: 2016 IEEE 18th International Conference on e-Health Networking, Applications and Services, Healthcom 2016, Institute of Electrical and Electronics Engineers Inc. <https://doi.org/10.1109/HealthCom.2016.7749510>
11. Radanović I, Likić R (2018) Opportunities for use of blockchain technology in medicine. *Appl Health Econ Health Policy* 16(5):583–590. <https://doi.org/10.1007/s40258-018-0412-8>
12. Irannezhad M, Shokouhyar S, Ahmadi S, Papageorgiou EI (2021) An integrated fcm-fbwm approach to assess and manage the readiness for blockchain incorporation in the supply chain. *Appl Soft Comput* 112:107832
13. Balci G, Surucu-Balci E (2021) Blockchain adoption in the maritime supply chain: examining barriers and salient stakeholders in containerized international trade. *Transp Res Part E Logist Transp Rev* 156:102539
14. Sharma N, Rohilla R (2022) Blockchain based electronic health record management system for data integrity. In: *Proceedings of International Conference on Computational Intelligence*, Springer, pp 289–297, [https://doi.org/10.1007/978-981-16-3802-2\\_24](https://doi.org/10.1007/978-981-16-3802-2_24)
15. Sharma N, Rohilla R (2020) Blockchain based approach for managing medical practitioner record: a secured design. In: *International Advanced Computing Conference*, Springer, pp 73–82, [https://doi.org/10.1007/978-981-16-0404-1\\_6](https://doi.org/10.1007/978-981-16-0404-1_6)

16. Sharma N, Rohilla R (2023) A novel hyperledger blockchain-enabled decentralized application for drug discovery chain management. *Comput Ind Eng*. <https://doi.org/10.1016/j.cie.2023.109501>
17. Chow SS, Choo K-KR, Han J (2020) Editorial for accountability and privacy issues in blockchain and cryptocurrency
18. Yuen TH (2020) Pachain: private, authenticated & auditable consortium blockchain and its implementation. *Future Gener Comput Syst* 112:913–929
19. Bonnah E, Shiguang J (2020) Decchain: a decentralized security approach in edge computing based on blockchain. *Future Gener Comput Syst* 113:363–379
20. Tseng JH, Liao YC, Chong B, Liao SW (2018) Governance on the drug supply chain via Gcoin blockchain. *Int J Environ Res Public Health*. <https://doi.org/10.3390/ijerph15061055>
21. Çolak M, Kaya I, Özkan B, Budak A, Karasın A (2020) A multi-criteria evaluation model based on hesitant fuzzy sets for blockchain technology in supply chain management. *J Intell Fuzzy Syst* 38(1):935–946. <https://doi.org/10.3233/JIFS-179460>
22. Agrawal TK, Kumar V, Pal R, Wang L, Chen Y (2021) Blockchain-based framework for supply chain traceability: a case example of textile and clothing industry. *Comput Ind Eng* 154:107130. <https://doi.org/10.1016/j.cie.2021.107130>
23. Hulseapple C (2015) Block verify uses blockchains to end counterfeiting and ‘make world more honest. Accessed 12 Apr 2019
24. Mackey TK, Nayyar G (2017) A review of existing and emerging digital technologies to combat the global trade in fake medicines. *Expert Opin Drug Saf* 16(5):587–602. <https://doi.org/10.1080/14740338.2017.1313227>
25. Chang SE, Chen YC, Lu MF (2019) Supply chain re-engineering using blockchain technology: a case of smart contract based tracking process. *Technol Forecast Soc Change* 144:1–11. <https://doi.org/10.1016/j.techfore.2019.03.015>
26. Tanwar S, Parekh K, Evans R (2020) Blockchain-based electronic healthcare record system for healthcare 4.0 applications. *J Inf Secur Appl* 50:102407. <https://doi.org/10.1016/j.jisa.2019.102407>
27. Kumar M, Chand S (2021) Medhypechain: a patient centered interoperability hyperledger-based medical healthcare system: regulation in covid-19 pandemic. *J Netw Comput Appl* 179:102975. <https://doi.org/10.1016/j.jnca.2021.102975>
28. Anita N, Vijayalakshmi M, Mercy Shalinie S (2022) Blockchain-based anonymous anti-counterfeit supply chain framework. *Sādhana* 47(4):208. <https://doi.org/10.1007/s12046-022-01984-2>
29. Anita N, Vijayalakshmi M, Mercy Shalinie S (2022) A lightweight scalable and secure blockchain-based IoT using fuzzy logic. *Wirel Pers Commun* 125(3):2129–2146. <https://doi.org/10.1007/s11277-022-09648-4>
30. Shi S et al (2022) A blockchain-based user authentication scheme with access control for telehealth systems. *Secur Commun Netw*. <https://doi.org/10.1155/2022/6735003>
31. Alsadi M, Arshad J, Ali J, Prince A, Shishank S (2023) TruCert: blockchain-based trustworthy product certification within autonomous automotive supply chains. *Comput Electr Eng* 109:108738. <https://doi.org/10.1016/j.compeleceng.2023.108738>
32. Chauhdary SH, Alkathiri MS, Alqarni MA, Saleem S (2023) An efficient evolutionary deep learning-based attack prediction in supply chain management systems. *Comput Electr Eng* 109:108768. <https://doi.org/10.1016/j.compeleceng.2023.108768>
33. Wang J, Chen J, Ren Y, Sharma PK, Alfarraj O, Tolba A (2022) Data security storage mechanism based on blockchain industrial internet of things. *Comput Ind Eng* 164:107903. <https://doi.org/10.1016/j.cie.2021.107903>
34. Omar IA, Debe M, Jayaraman R, Salah K, Omar M, Arshad J (2022) Blockchain-based supply chain traceability for covid-19 personal protective equipment. *Comput Ind Eng* 167:107995. <https://doi.org/10.1016/j.cie.2022.107995>
35. Xie R, Hong C, Zhu S, Tao D (2015) Anti-counterfeiting digital watermarking algorithm for printed QR barcode. *J Neurocomput*. <https://doi.org/10.1016/j.neucom.2015.04.026>
36. Zheng Z, Zheng H, Ju J, Chen D, Li X, Guo Z, You C, Lin M (2021) A system for identifying an anti-counterfeiting pattern based on the statistical difference in key image regions. *Expert Syst Appl* 183:115410. <https://doi.org/10.1016/j.eswa.2021.115410>
37. Zhu P, Hu J, Zhang Y, Li X (2020) A blockchain based solution for medication anti-counterfeiting and traceability. *IEEE Access* 8

38. Chow YW, Susilo W, Baek J, Kim J (2020) QR code watermarking for digital images. In: You I (ed) *Information Security Applications. WISA 2019. Lecture Notes in Computer Science*, Springer, Cham, vol 11897. [https://doi.org/10.1007/978-3-030-39303-8\\_3](https://doi.org/10.1007/978-3-030-39303-8_3)
39. Cardamone N, d'Amore F (2019) DWT and QR code based watermarking for document DRM. In: Yoo CD, Shi Y-Q, Kim HJ, Piva A, Kim G (eds) *IWDW 2018. LNCS*, Springer, Cham, vol 11378, pp 137–150. [https://doi.org/10.1007/978-3-030-11389-6\\_11](https://doi.org/10.1007/978-3-030-11389-6_11)
40. Xun Y, Li Z, Zhong X, Li S, Su J, Zhang K (2019) Dual anti-counterfeiting of QR code based on information encryption and digital watermarking. In: Zhao P, Ouyang Y, Xu M, Yang L, Ren Y (eds) *Advances in Graphic Communication, Printing and Packaging. Lecture Notes in Electrical Engineering*, Springer, Singapore, vol 543. [https://doi.org/10.1007/978-981-13-3663-8\\_27](https://doi.org/10.1007/978-981-13-3663-8_27)
41. Liu J, Han J, Fu K, Jia J, Zhu D, Zhai G. Application of QR code watermarking and encryption in the protection of data privacy of intelligent mouth opening trainer. In: *IEEE Internet of Things Journal*. <https://doi.org/10.1109/JIOT.2023.3242319>
42. Mannepalli PK, Richhariya V, Gupta SK, Shukla PK, Dutt PK (2021) Block chain based robust image watermarking using edge detection and wavelet transform. *Research square*
43. Rawat P, Shukla PK (2021) Robust digital color image watermarking and encryption algorithms using blockchain over DWT edge coefficient. A book chapter in 1st edition of *Blockchain for information security and privacy*
44. Singh R, Rawat P, Shukla P, Shukla K (2019) Invisible color image watermarking using edge detection and discrete wavelet transform coefficients. *Int J Innov Technol Explor Eng (IJITEE)* 9(1)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

## Authors and Affiliations

**Neetu Sharma<sup>1</sup> · Rajesh Rohilla<sup>1</sup>**

✉ Neetu Sharma  
neetusharma85@gmail.com; neetu\_2k19phdec25@dtu.ac.in  
  
Rajesh Rohilla  
rajesh@dce.ac.in

<sup>1</sup> Delhi Technological University, New Delhi 110042, India



# A Single MOS-Memristor Emulator Circuit

Rahul Kumar Gupta<sup>1,2</sup> · Mahipal Singh Choudhry<sup>1</sup> · Varun Saxena<sup>3</sup> · Sachin Taran<sup>1</sup>

Received: 19 April 2023 / Revised: 20 August 2023 / Accepted: 20 August 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

## Abstract

A floating/grounded memristor emulator having a single MOSFET connected to a first-order RC filter with high operating frequency is presented. The proposed memristor emulator is passive in nature and provides zero static power, and a few pW of dynamic power when input signal is applied. The emulator has been simulated using Cadence Virtuoso Spectre tool. The mathematical analysis substantiates the existence of the proposed memristor emulator and its incremental nature. With a layout area of  $1.586 \mu\text{m}^2$ , the memristor emulator operates up to 80 MHz. The main contribution is the experimental verification of the proposed topology using discrete transistor, ALD1117 Dual P-channel enhancement MOSFET array and discrete elements in the form of an R–C tank circuit. The performance characteristics for its analog and digital applications have been simulated and verified in grounded as well as floating mode. The emulator circuit offers a simplified design and low power consumption compared to other existing memristor emulators.

---

✉ Rahul Kumar Gupta  
[rahulkumargupta\\_2k21phdec08@dtu.ac.in](mailto:rahulkumargupta_2k21phdec08@dtu.ac.in)

Mahipal Singh Choudhry  
[msc\\_1976@yahoo.com](mailto:msc_1976@yahoo.com)

Varun Saxena  
[varunsaxena@jnu.ac.in](mailto:varunsaxena@jnu.ac.in)

Sachin Taran  
[sachintaran@dtu.ac.in](mailto:sachintaran@dtu.ac.in)

<sup>1</sup> Electronics and Communication Engineering, Delhi Technological University, Rithala, New Delhi, Delhi 110047, India

<sup>2</sup> Electronics and Communication Engineering, JSS Academy of Technical Education, C-20/1, Sector-62, Noida 201301, India

<sup>3</sup> School of Engineering, Jawaharlal Nehru University, New Mehrauli Road, Mehrauli, New Delhi 110067, India

**Keywords** Emulator · Floating · Grounded · High frequency · Hysteresis · Memristor · MOSFET

## 1 Introduction

Memristor, first introduced by L. Chua in 1971 and 1976 [17, 18], is now being regarded as an integral part of the fundamental component category. Due to the element's straightforward construction, compact size, low power consumption, and storage capacity, it has been actively used in design of electronic circuits. On account of its non-volatile nature, a memristor has impacted research in the realm of electronic circuits in both the analog and digital domains. Memristor [47] as the fourth essential passive circuit component has been used in a broad variety of contexts, including programmable analog circuits [38], neuromorphic applications circuits [3, 8], chaotic oscillator circuits [36], non-volatile memory [5], asymmetric attractor [53], XOR logic circuits [28], and canonical Chua's circuit [15, 16, 52]. Real-time applications using memristors have had little success on account of the complexities pertaining to the highly intricate and cost intensive fabrication process. Therefore, memristor emulators have emerged as an attractive and viable alternative to showcase the features and applications of memristor-based circuits. In [4], Adhikari et al. provided an illustration of the constricted hysteresis loop and frequency-dependent hysteresis lobe area that are features of the memristor emulator. A number of researchers have proposed memristor emulators that employ various active blocks already in use, including operational transconductance amplifiers (OTA) [7–9, 32], operational amplifiers (OPAMP) [30, 37, 58], current feedback operational amplifiers (CFOA) [1, 2], second-generation current conveyors (CCII) [43, 45], differential-difference current conveyors (DDCC) [54], current conveyor transconductance amplifiers (CCTA) [40], differential voltage current conveyor transconductance amplifiers (DVCCTA) [41], voltage difference transconductance amplifier (VDTA) [48], voltage differencing inverting buffered amplifier (VDIBA) [39], current follower differential input transconductance amplifier (CFDITA) [33], and voltage differencing current conveyor (VDCC) [55]. Memristor based on such active blocks exhibits circuit complexity and a correspondingly high power consumption.

Another attractive strategy has been to design memristor emulators using only MOSFETs with fewer passive circuit components. This alternative approach offers high operating frequency, a significant reduction in circuit complexity, and a relatively low power consumption. Vista and Ranjan in [49] proposed a simple floating memristor emulator operating up to 13 MHz with three NMOS, one capacitor, and one DC current source. Using two BJTs, two diodes, four resistors, and two capacitors, John et al. suggested a non-ideal totally passive memristor in [29] that operates up to a few kHz. Non-ideality is on account of asymmetry in the hysteresis loop and nonzero crossing point. The emulator is fully passive on account of no biasing voltage. In [22], E. Gale introduces two novel additions to the memory conservation theory [23] and describes how to create various non-idealities and illustrates many non-ideal memristor kinds, including filamentary memristor. A weak inversion CMOS-based memristor with a grounded capacitor and two log-domain exponential transconductors is presented by

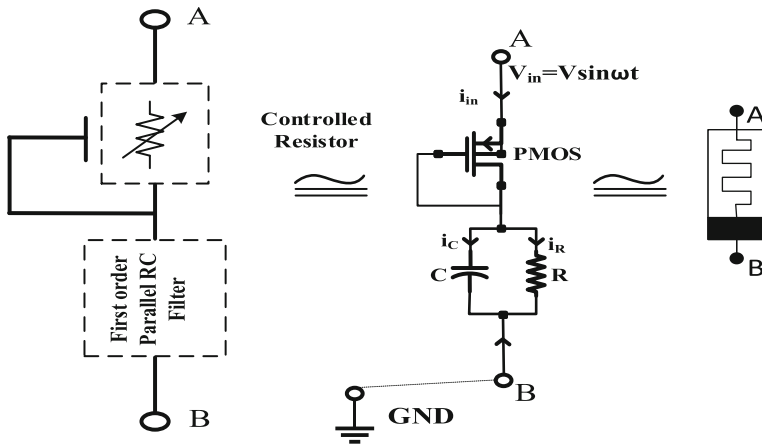
Koymenn and Emmanuel [31]. Babacan et al. in [10] have presented a memristor emulator circuit using only four MOS transistors and also fabricated a ZnO thin-film-based semiconductor memristor device using an external DC biasing with an operating range up to 100 MHz. Babacan in [11] presented a memristor emulator circuit employing just three MOS transistors (NMOS) and a single capacitor, functioning in passive mode up to 100 KHz and requiring no DC biasing. Seven MOS transistors and one grounded capacitor are used in Abdulla's [56] proposed memristor circuit, which has a 50-MHz operating frequency range. One external DC current source and seven MOS transistors—three NMOS, two PMOS, one NMOS CAP, and one zero-voltage threshold (ZVT) NMOS, are used in Vishal's [44] compact CMOS memristor circuit, which has applications in memcomputing. Some researchers [14, 21, 35] have described the behavioral model of a memristor solid-state device and simulate it using SPICE. A dynamic threshold MOSFET technique for memristor emulators has been developed by Pushkar et al. [26, 46] using either two or three MOSFETs without using any passive components or DC biasing. Some of the recent manuscripts have proposed different configurations of four MOSFETs to achieve the memristor emulator characteristics. They, however, exhibit a trade-off between operating frequency and a high power consumption [6, 24, 25, 27]. Recently, some of the works [34, 57, 59] have also proposed memristor emulator with a minimal number of MOSFETs, but they lack in the layout area, power consumption, operating frequency, and technology used. These works are summarized in Table 3. A few but quite interesting realizations of memristors emulators have been achieved without the use of either active blocks or MOSFETs. In [13], the authors have realized a generalized memristor using a full-wave bridge rectifier and a first-order RC filter. In some works, instead of using a first-order RC filter, first-order RL [51] or LCR filter [20] or three-phase bridge rectifier circuit [50] has been used.

This manuscript aims to present a new circuit to realize a grounded memristor emulator. This memristor is voltage controlled and can operate in floating mode as well. The proposed memristor emulator operates at frequencies up to 80 MHz without using any external DC bias and is characterized by zero static power, and a few pW ( $\sim 7.75$  pW) of dynamic power. In this work, a mathematical model of the proposed memristor is substantiated by simulations, hardware implementation, and verification. The key features of the proposed memristor emulators are as follows:

1. It has a single MOSFET connected to an R–C rank circuit.
2. It has an area efficient circuit when compared with other emulators reported in the literature.
3. The simplicity of the proposed emulator facilitates easy integration with other circuits/devices for designing potential applications.
4. Unlike other emulators, the proposed emulator works without an external bias.

A comparative analysis of the key features is available in a supplementary material associated with this work. Numerical analysis of the proposed emulator circuit is conducted using the cadence Virtuoso EDA framework and Generic Process Design Kit. Furthermore, the efficiency of the proposed memristor emulator is validated by implementing various memristor-based applications. It is worth noting that the proposed memristor emulator is realized physically using existing ALD1117 P-channel





**Fig. 1** Block diagram depiction and MOSFET implementation of memristor emulator

MOSFETs, which generates hysteresis loops usually encountered in memristor-based circuits at various frequencies. This further corroborates the correctness of the proposed memristor emulator topology.

A comparison of the proposed memristor emulator with other contemporary emulator circuits is given in Table 2. Section 2 briefly describes the mathematical modeling of the proposed memristor emulator. Section 3 presents the simulation and experimental validation of the memristor emulator. Section 4 showcases different applications, namely synthesis of a Schmitt trigger, chaotic Colpitts oscillator, and XOR logical operation using the proposed memristor emulator. We conclude the work by highlighting the advantages of our memristor emulator design in Sect. 5.

## 2 Design and Theory of Proposed Memristor Emulator

The design of the proposed memristor emulator is centered around a circuit in which a diode-connected load (controlled resistor) [42] is connected in series with a first-order parallel RC filter (see Fig. 1).

The behavior of the diode-connected load under small signal conditions matches that of a two-terminal resistor when the gate and drain are shorted. The impedance of the diode-connected load can be expressed as  $(1/g_m) || r_0$ . Since  $r_0$  is high, this value can be safely approximated to  $1/g_m$  [42].

The circuit uses a PMOS to realize a diode-connected load, and since the drain terminal is shorted to the gate terminal, one can write,  $v_{gs} = v_{ds} = v_c - v_{in}$ . The PMOS will always operate in saturation mode because  $v_{ds} > v_{gs} - v_{tp}$ . The drain current  $I_D$  can then be expressed as

$$I_D = -\mu_p C_{ox} \left( \frac{W}{2L} \right) (v_{gs} - v_{tp})^2 \quad (1)$$

where  $\mu_p$  is the mobility of electrons,  $C_{ox}$  is the oxide capacitance,  $W/L$  is the aspect ratio,  $v_{gs}$  is the gate to source voltage, and  $v_{tp}$  is the threshold voltage. The current across the circuit,  $i_{in}$ , is equal to the drain current and can be expressed as a product of the conductance  $g_m$  and the input voltage,  $v_{in}$ , i.e.,

$$i_{in} = I_D = g_m(v_c, v_{in}) \times v_{in} \quad (2)$$

The conductance  $g_m$  is itself a function of the capacitor voltage  $v_c$  and input voltage  $v_{in}$ . One can write  $I_D = i_C + I_R$  according to Kirchhoff's current law, where  $I_R$  and  $i_C$  are the currents flowing across the resistor and the capacitor, respectively. One can write

$$i_C = I_D - I_R \quad (3)$$

$$C \frac{dv_C}{dt} = f(v_C, v_{in}) = k \times (v_{gs} - v_{tp})^2 - \frac{v_C}{R} \quad (4)$$

where  $k = -\mu_p C_{ox} W/2L$  and  $f$  is a smooth function of the capacitor and input voltages. Now,

$$\frac{dv_C}{dt} = \frac{k(v_{gs} - v_{tp})^2}{C} - \frac{v_C}{RC} \quad (5)$$

$$\frac{dv_C}{dt} = \frac{k(v_C - v_{in} - v_{tp})^2}{C} - \frac{v_C}{RC} \quad (6)$$

Equation (2) and Equation (6) form the basis of our mathematical model of memristor emulator. These two equations are similar to those in [13, 19] wherein instead of a MOSFET, diodes are considered in the emulator circuit.

The total impedance of the proposed emulator circuit is derived in Equations (7–9)

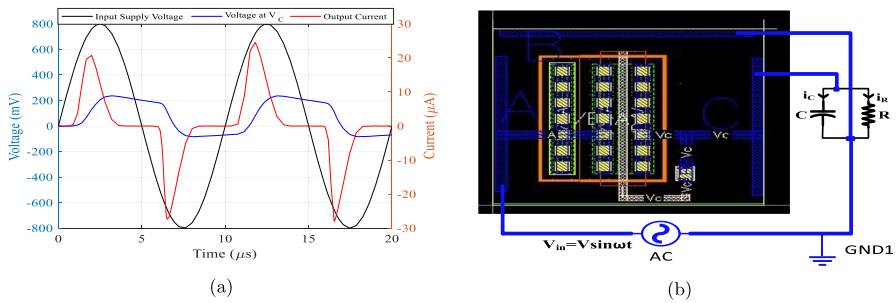
$$M_R = \frac{V_{in}}{I_{in}} = \frac{1}{g_m} + \frac{1}{\frac{1}{X_C} + \frac{1}{R}} \quad (7)$$

where  $g_m = k.(v_{gs} - v_{tp})$ ,  $X_C = 1/sC$  and  $s = j\omega$

$$M_R = \frac{1}{g_m} + \frac{R}{1 + sRC} \quad (8)$$

$$M_R = \frac{1}{k(v_C - v_{in} - v_{tp})} + \frac{R}{1 + sRC} \quad (9)$$

The working of the proposed circuit for one time period, i.e.,  $T = 10 \mu s$ , is illustrated in Fig. 2a. Consider the capacitor of capacitance  $C$  discharged at  $t = 0$ . The capacitor will remain completely discharged so long as magnitude of source to gate voltage,  $v_{sg} \simeq (v_{in} - v_C) < v_{tp}$ . Moreover, the PMOS will also remain OFF until  $|v_{sg}| < v_{tp}$ . Therefore, the current  $I_D$  will be zero. Now, during the interval  $0 \leq t < 2.5 \mu s$ , when the magnitude of voltage  $v_{sg}$  exceeds the threshold voltage  $v_{tp}$ ,



**Fig. 2** **a** Transient behavior of output current with respect to the voltage across the capacitor **b** CMOS layout created for the proposed memristor emulator

the PMOS will begin to conduct and shall behave as a diode-connected load with a drain current  $I_D = g_m(v_{in} - v_C)$ . With a nonzero  $I_D$  flowing in the circuit, the capacitor begins to get charged. As a consequence, the current starts to decrease because the difference  $v_{in} - v_C$  decreases. Now, in the time interval  $2.5 \mu s \leq t < 5 \mu s$ , the input voltage decreases, the capacitor starts discharging and when  $|v_{in} - v_C| < v_{TP}$ , the current flowing the circuit becomes zero again. PMOS again goes to an OFF state. In the negative half cycle, for  $5 \mu s \leq t < 7.5 \mu s$ , the magnitude of the current increases; however, the direction reverses because the capacitor begins to discharge. Finally, in the time interval  $7.5 \mu s \leq t < 10 \mu s$ , the current goes to zero once again when  $|v_{in} - v_C| < v_{TP}$ .

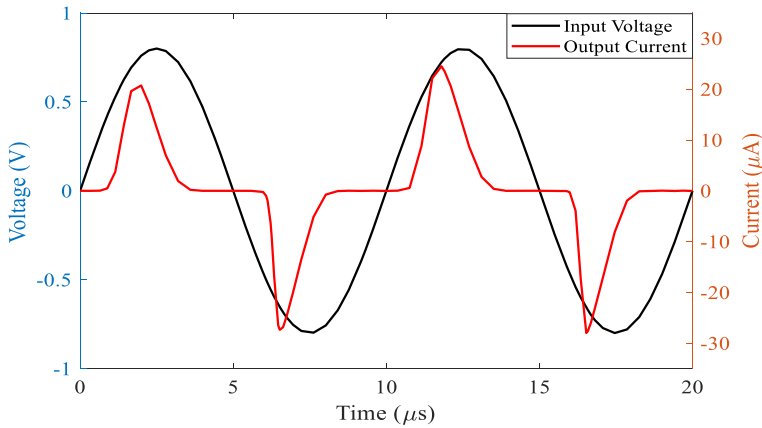
### 3 Model Simulation and Experimental Validation

Using Cadence Virtuoso Spectre tool, the proposed memristor emulator model has been simulated. A resistor of  $1\text{ k}\Omega$ , capacitances of  $1\text{ nF}$ ,  $47\text{ nF}$ ,  $100\text{ nF}$ , and a MOSFET ALD1117 Dual P-channel enhancement MOSFET array are used to verify the memristor emulator circuit experimentally. The physical layout of the proposed emulator circuit of Fig. 1 is shown in Fig. 2b. The layout area of this proposed circuit is  $1.586\text{ }\mu\text{m}^2$ . Since there is no biasing, the static power is zero; however, the dynamic power depends on the input supply and the input current.

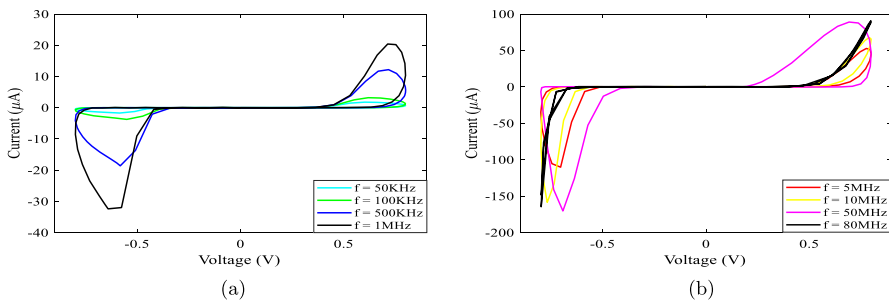
The simulation of an emulator, an R-C tank circuit having a resistor of resistance  $R = 100\text{ k}\Omega$ , and a capacitor of capacitance  $C = 100\text{ pF}$  and PMOS with an aspect ratio of  $0.9\text{ }\mu\text{m}/0.045\text{ }\mu\text{m}$  has been used.

Figure 3 shows the transient current waveform for the proposed memristor emulator when a sinusoidal input voltage having an amplitude of  $0.8 V_{p-p}$  and frequency of 100 kHz is applied. The pinched hysteresis curves for applied input voltage with frequencies ranging from 50 kHz to 1 MHz and 5 MHz to 80 MHz are shown in Fig. 4a and Fig. 4b, respectively.

The symmetric and asymmetric behavior of the proposed memristor depends on the value of the resistance  $R$  (see Fig. 5a). It is evident for the figure that as the value of resistance  $R$  decreases, asymmetry in the hysteresis curve becomes more pronounced.



**Fig. 3** The transient current response to the applied sinusoidal input of  $0.8 V_{p-p}$  and frequency of 100 kHz



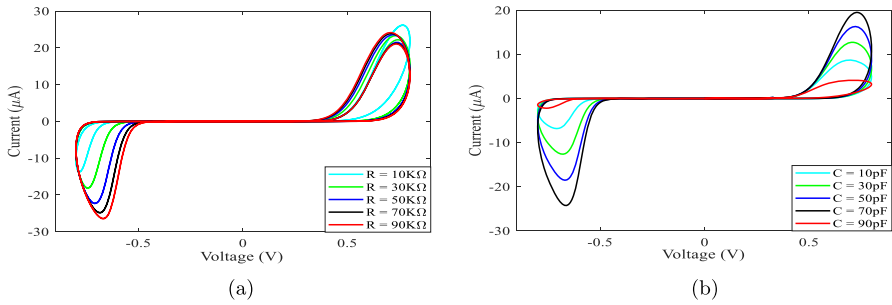
**Fig. 4** The pinched hysteresis curve for frequencies ranging from **a** 50 kHz to 1 MHz and **b** 5 MHz to 80 MHz

The effect of change in capacitance  $C$  on the current is shown in Fig. 5b. It is evident that the symmetry in the hysteresis curves is well defined.

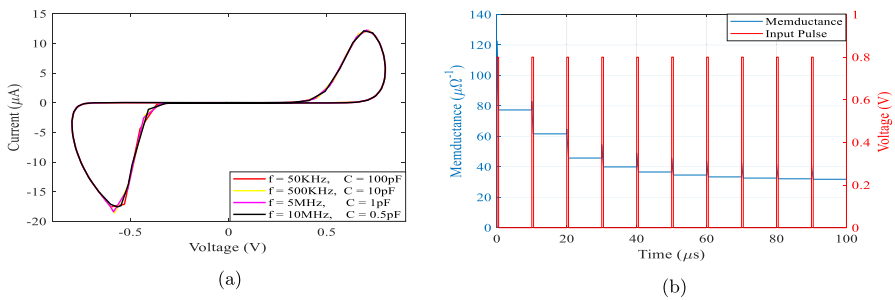
In Fig. 6a, it is evident that the memristor circuit shows similar voltage and current relationship when the product of the operating frequency and the capacitor value is kept constant. This is on account of the fact that the reactance offered by the capacitance remains unchanged.

Figure 6b represents the decremental behavior of the memductance of the proposed memristor emulator when a pulse input of voltage  $V_P = 0.8 V$ , time duration of  $T = 10 \mu s$ , ( $T = T_{on}(500 ns) + T_{off}$ ) is applied. The tank circuit has a capacitor of capacitance  $C = 1 nF$  and resistance  $R = 50 k\Omega$ .

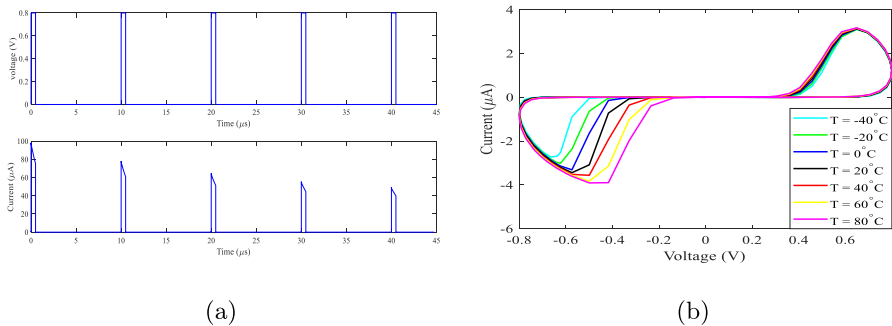
The corresponding staircase output at the capacitor node indicates that the value of memductance is constant even in the absence of input signal exhibiting the non-volatile nature of the proposed memristor emulator. Another way to show the non-volatile nature of the memristor emulator is shown in Fig. 7a. Upon application of the same pulse train, the current decreases from approximately  $98 \mu A$  to  $80 \mu A$  during the first ON cycle of  $500 ns$  and retains the latter current value during the OFF cycle.



**Fig. 5** **a** Hysteresis curve of memristor emulator at different values of resistances (asymmetric). **b** Hysteresis curve at different values of capacitances



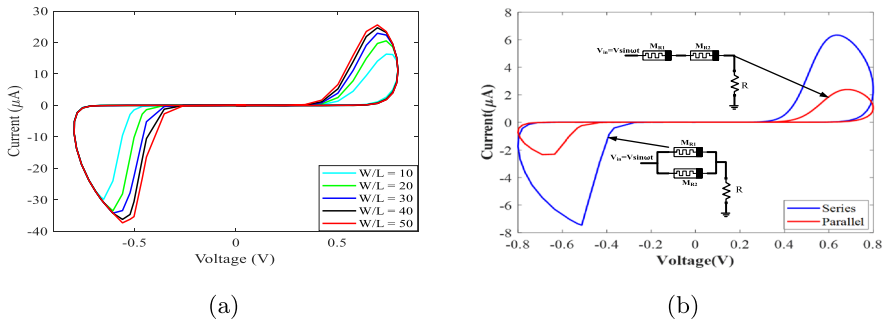
**Fig. 6** **a** V-I curves for memristor emulator at constant products of operating frequency and capacitance of the capacitor. **b** The incremental behavior of the memristor emulator with respect to the memductance. The pulsed input voltage is also given for reference



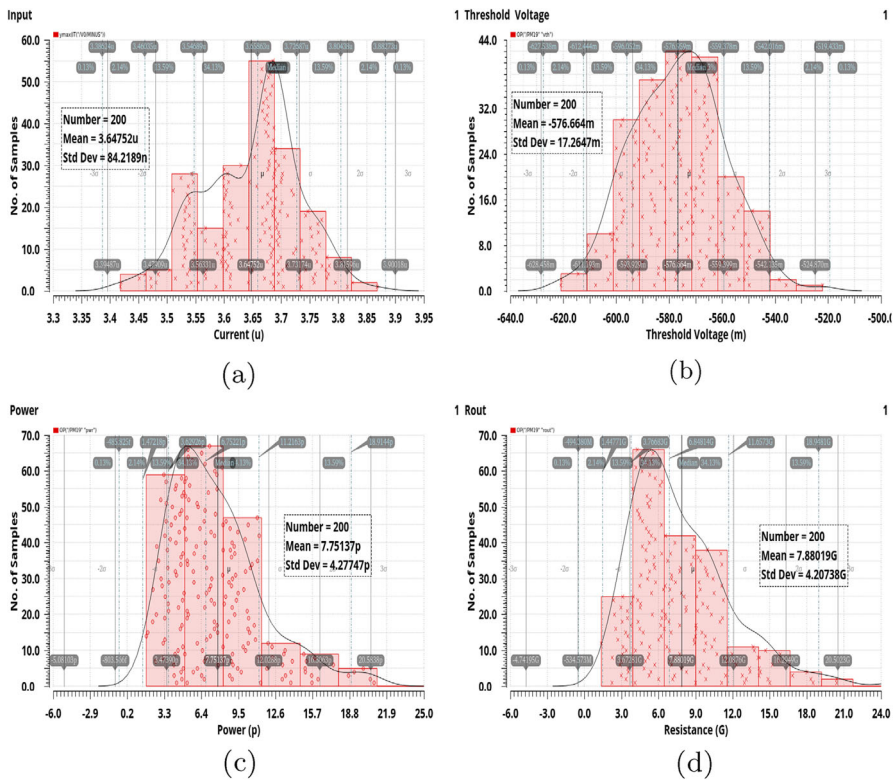
**Fig. 7** **a** Non-volatility behavior of proposed memristor emulator. **b** The pinched hysteresis for different temperatures in  $^{\circ}\text{C}$

This proposed memristor emulator works over a wide range of temperatures ( $-40^{\circ}\text{C}$ – $80^{\circ}\text{C}$ ) as shown in Fig. 7b.

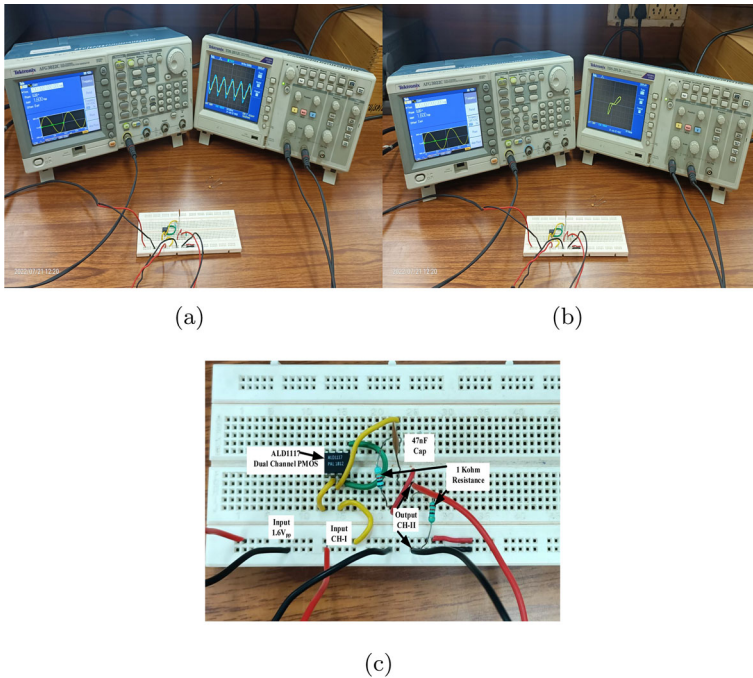
The effect of change in the W/L ratio is shown in Fig. 8a. It is evident that the current flowing through the memristor directly depends on the W/L ratio.



**Fig. 8** **a** The pinched hysteresis curves for varied W/L ratios. **b** The pinched hysteresis curve for different combinations of memristor at 100 kHz



**Fig. 9** Monte Carlo simulation results for memristor emulator parameters. **a** Frequency distribution for current in  $\mu A$ , **b** frequency distribution for threshold voltage in V, **c** frequency distribution for power in pW, **d** frequency distribution for resistance in  $G\Omega$



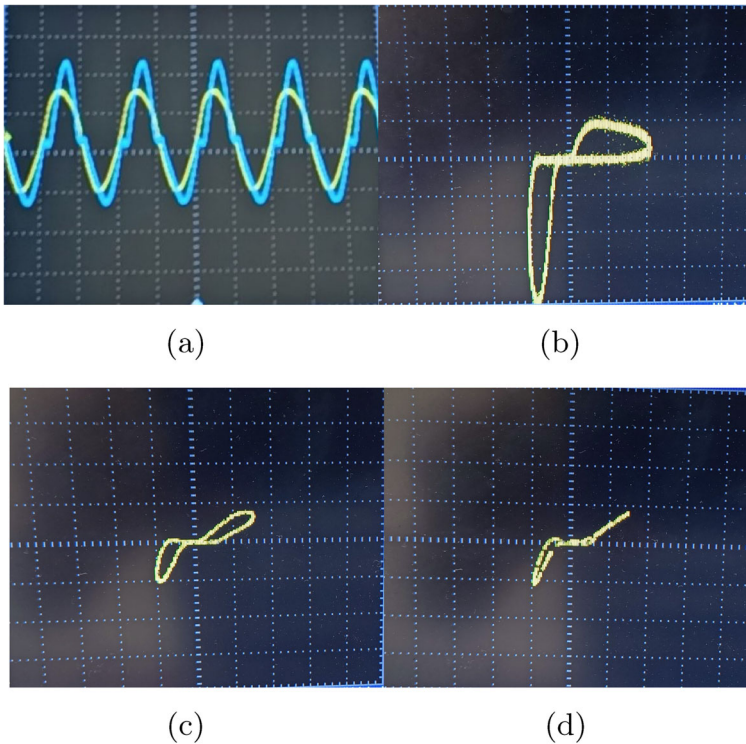
**Fig. 10** **a** Full experimental setup and breadboard circuit implementation, **b** full experimental setup at operating frequency of 100 kHz, and **c** circuit connections on the breadboard

In Fig. 8b, the series combination of memristor emulator shows that the current reduces as the memristance value increases, whereas in the parallel combination of memristor emulators, the current increases due to a reduction in the memristance value.

Monte Carlo analysis has been performed for 200 runs to justify the robustness of the proposed memristor emulator. The memristor emulator parameters such as current, threshold voltage, power and output resistance are chosen for the simulation, and the performance of the proposed memristor emulator is found to be satisfactory. The frequency distribution for current in  $\mu A$  is plotted in Fig. 9a, the frequency distribution for threshold voltage in  $V$  is plotted in Fig. 9b, the frequency distribution for the power in  $pW$  is plotted in Fig. 9c, whereas the frequency distribution for the output resistance is plotted in  $G\Omega$  in Fig. 9d. It is evident that the relevant parameter variation is well within  $3\sigma$  limit about the respective mean  $\mu$ , where  $\sigma$  is the respective standard deviation.

The experimental setup using Tektronix TDS 2012C dual-channel DSO of 100 MHz & 2 GS/s is shown in Fig. 10. The experimental results shown in Figure 11 are for different frequencies and hysteresis curve validating the topology of proposed memristor emulator.





**Fig. 11** Experimental result for pinched hysteresis loop for the proposed memristor emulator. **a** Input and output waveform. **b** Hysteresis loop at 50 kHz. **c** Hysteresis loop at 100 kHz. **d** Hysteresis loop at 500 kHz

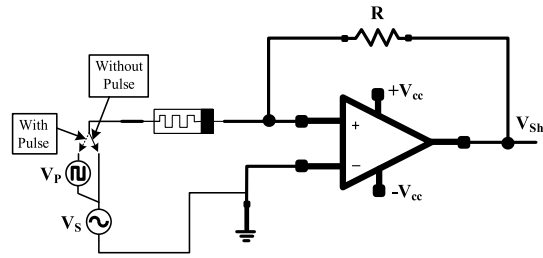
#### 4 Applications of the Proposed Memristor Emulator

In this section, we show some applications of the proposed memristor emulator which underscores the workability and practical usefulness of the emulator. The proposed emulator circuit is used to design a Schmitt trigger, a chaotic Colpitts oscillator, and a XOR operation. The Schmitt trigger circuit shown in Fig. 12a is based on a single operational amplifier (TL082CD).

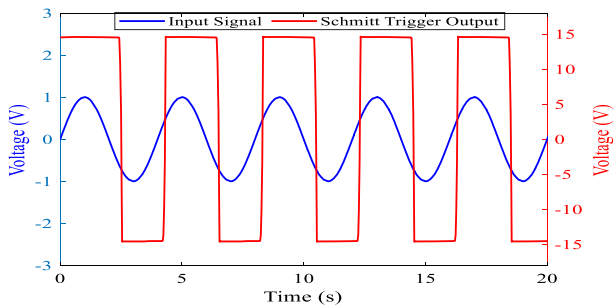
The upper and lower thresholds of the Schmitt trigger are given by

$$V_{TH} = V_{sat} \left( \frac{M_R}{R_1} \right), \quad V_{TL} = -V_{sat} \left( \frac{M_R}{R_1} \right) \quad (10)$$

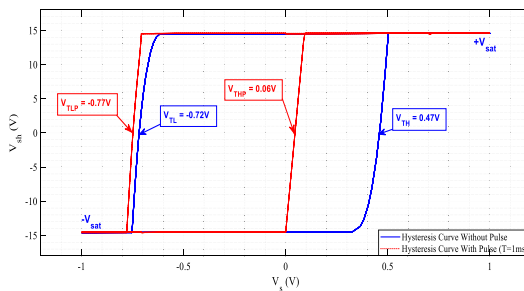
Here,  $V_{sat}$  is the saturation voltage of the operational amplifier in volts,  $M_R$  is the memristance of the memristor emulator, and  $R_1$  is the resistance value of the resistor in the circuit. The supply voltage applied to the operational amplifier is  $\pm 15V$  DC. The resistance value  $R_1 = 50 \text{ M}\Omega$  is chosen with the sinusoidal supply of 250 Hz and 0.8 V ( $V_{p-p}$ ). The input and output voltages of the Schmitt trigger are shown in Fig. 12b. The hysteresis curve changes when we apply a programming pulse of 1 V and 250 Hz along with the input voltage (see Fig. 12c). Applying a sequence



(a)



(b)

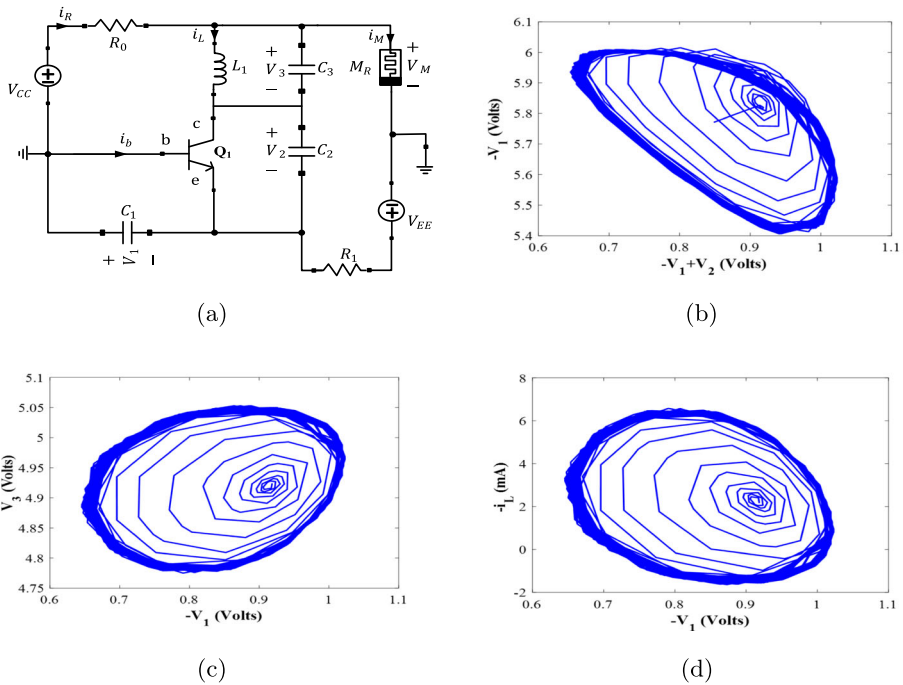


(c)

**Fig. 12** Realization of a Schmitt trigger using proposed memristor emulator. **a** Schmitt trigger circuit. **b** Input and output voltages for the Schmitt trigger circuit. **c** Threshold voltage variation

of programming pulses allows for the modification of the memristance value, which changes the threshold voltage of the Schmitt trigger.

A fourth-order chaotic Colpitts oscillator is proposed in Fig. 13a. A chaotic oscillator circuit finds application in diverse areas such as cryptography, encryption, random signal generators, and specific radar systems. The chaotic Colpitts oscillator designed here using the proposed memristor emulator has a NPN bipolar junction transistor (as



**Fig. 13** **a** Fourth-order Colpitts oscillator. **b–d** Memristive Colpitts oscillator with initial values of the state variables as  $V_1(0) = 0.01$  V,  $V_2(0) = 0.01$  V,  $V_3(0) = 0$  V,  $i_L = 0$  A

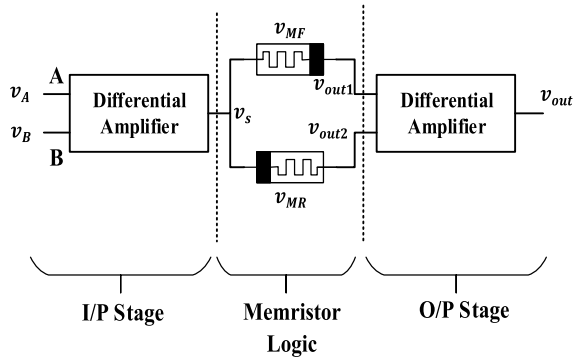
**Table 1** The truth table

$A (v_A)$	$B (v_B)$	$v_A - v_B$	$v_{out} = v_{out1} - v_{out2}$	logic
$V_1$	$V_1$	0	0	0
$V_1$	0	$V_1$	$> 0$	1
0	$V_1$	$-V_1$	$> 0$	1
0	0	0	0	0

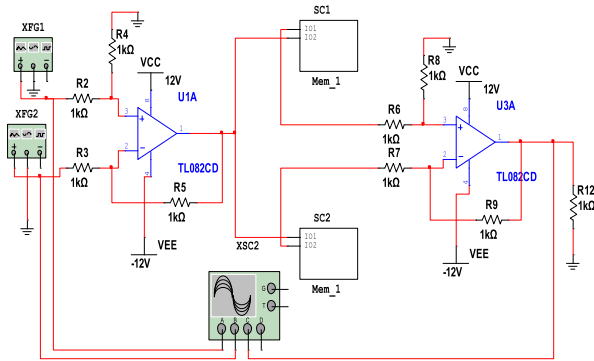
a nonlinear element). The oscillator circuit has capacitors of capacitance  $C_1 = C_2 = 2 \mu F$ , and  $C_3 = 33$  nF, an inductor of inductance  $L = 10$  mH, resistors of resistance  $R_0 = 35 \Omega$ ,  $R_1 = 1.5$  k $\Omega$ , and voltage  $V_{cc} = 5$  V and  $V_{ee} = -5$  V.

For arbitrarily chosen initial values for state variables  $V_1$ ,  $V_2$ ,  $V_3$ , and  $i_L$ , the memristive Colpitts oscillator is chaotic and displays a spiral chaotic attractor, plotted in Fig. 13b, 13c, 13d.

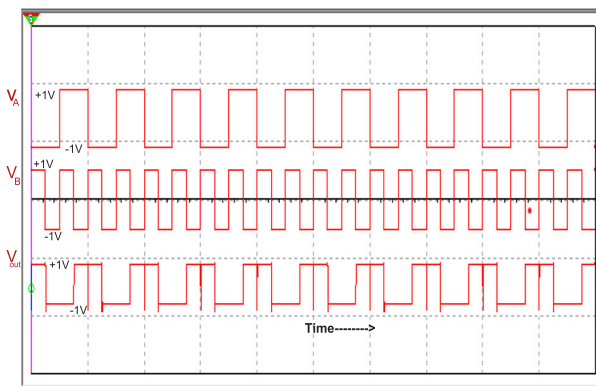
A memristor-based XOR logic gate using differential amplifier is implemented next. The circuit for XOR gate is given in Fig. 14a. The input stage and memristor logic are the part of circuit module, and the output stage is the test module. The differential amplifier circuit is implemented using an operational amplifier TL082 (dual channel, gain bandwidth product of 3 MHz). The truth table for the XOR logic gate is given in Table 1.



(a)



(b)



(c)

**Fig. 14** XOR logic gate using Multisim software. **a** Block diagram of a XOR logic. **b** Multisim circuit realization. **c** The input and output waveforms for the XOR logic gate realization

**Table 2** Design-based comparison of the proposed model with existing memristor emulators

Ref. no.	Publication year	MOSFET count	Existing blocks used	Passive/active component count
[7]	2017	38	OTA-1	R-1, C-1
[32]	2023	17	OTA-1	C-1
[37]	2023	> 20	OPAMP-1	R-1, C- 1
[40]	2017	30	CCTA-1	R-3 C-1
[41]	2017	29	DVCCTA-1	R-3 C-1
[48]	2021	16	VDTA-1	C-1 R-1
[39]	2022	18	VDIBA-2	C-1
[33]	2023	22	CFDITA-1	C-1
[55]	2019	26	VDCC-1 MOSFET-2	C-1
[49]	2019	3	–	C-1
[10]	2018	4	–	–
[56]	2018	7	–	C-1
[24]	2022	4	–	–
[25]	2019	4	–	–
[6]	2022	4	–	C-1
[57]	2023	4	–	C-1
[34]	2023	3	–	C-1
[59]	2023	2	–	–
[12]	2017	6	–	C-1
[13]	2014	–	–	D-4, R-1, C-1
[53]	2020	–	–	D-6, R-1, C-1
[28]	2022	–	–	D-4, R-1, C-1
Proposed Work	–	1	–	R-1, C-1

Based on block diagram in Fig. 14a,

$$v_{out1} = |v_S| - v_{MF}, v_{out2} = |v_S| - v_{MR}, v_{out} = v_{out1} - v_{out2} \quad (11)$$

Figure 14b shows the implementation of the XOR logic gate on Multisim software, whereas Fig. 14c shows the input and output waveforms of the XOR logic gate implemented using the proposed memristor emulator.

Tables 2 and 3 contain a comparative summary of selected previous memristor emulators [6, 7, 10, 12, 13, 24, 25, 28, 32–34, 37, 39–41, 48, 49, 53, 55–57, 59] based on active/passive elements, circuit structure, simulation/experimental validations, technology, operating frequency, and power consumption. The number of MOSFET counts is higher in [7, 32, 33, 37, 39–41, 48, 55]. Some emulators are based on fewer MOS-

**Table 3** Performance-based comparison of the proposed model with existing memristor emulators

Ref. no.	Experiment/simulation	Floating/grounded	Technology	Frequency	Power consumption	Layout area
[7]	Both	Grounded	0.18 $\mu\text{m}$ CMOS	1 kHz	Few nW	NA
[32]	Simulation	Floating	0.18 $\mu\text{m}$ GPDK	1 MHz	411 $\mu\text{W}$	2596 $\mu\text{m}^2$
[37]	Both	Floating	-	30 MHz	Few mW	NA
[40]	Both	Both	0.18 $\mu\text{m}$ TSMC	10 MHz	5.22 mW	NA
[41]	Both	Grounded	0.25 $\mu\text{m}$ TSMC	1 MHz	4.88 mW	NA
[48]	Both	Floating	0.18 $\mu\text{m}$ CMOS	50 MHz	8 $\mu\text{W}$	1065 $\mu\text{m}^2$
[39]	Both	Grounded	0.18 $\mu\text{m}$ TSMC	12.7 MHz	1.34 mW	NA
[33]	Both	Both	0.18 $\mu\text{m}$ GPDK	10 MHz	4.8 $\mu\text{W}$	112 $\mu\text{m}^2$
[55]	Both	Grounded	0.18 $\mu\text{m}$ TSMC	2 MHz	NA	NA
[49]	Both	Floating	0.18 $\mu\text{m}$ TSMC	13 MHz	6.725 nW	2803 $\mu\text{m}^2$
[10]	Both	Grounded	0.18 $\mu\text{m}$ TSMC	100 MHz	NA	366 $\mu\text{m}^2$
[56]	Both	Grounded	0.18 $\mu\text{m}$ TSMC	50 MHz	-	456 $\mu\text{m}^2$
[24]	Both	Floating	0.09 $\mu\text{m}$ GPDK	50 MHz	2.6 $\mu\text{W}$	59.41 $\mu\text{m}^2$
[25]	Simulation	Grounded	0.18 $\mu\text{m}$ TSMC	100 kHz	400 $\mu\text{W}$	196 $\mu\text{m}^2$
[6]	Both	Floating	0.18 $\mu\text{m}$ TSMC	3 MHz	8.24 $\mu\text{W}$	158 $\mu\text{m}^2$
[57]	Both	Grounded	0.18 $\mu\text{m}$ GPDK	10 MHz	20 $\mu\text{W}$	NA
[34]	Both	Grounded	0.09 $\mu\text{m}$ GPDK	5 kHz	175 nW	1154 $\mu\text{m}^2$
[59]	Both	Floating	0.065 $\mu\text{m}$ GPDK	300 MHz	Zero	4.62 $\mu\text{m}^2$
[12]	Simulation	Floating	0.18 $\mu\text{m}$ TSMC	10 Hz	-	NA
[13]	Both	Both	-	10 kHz	-	NA
[53]	Both	Both	-	20 kHz	-	NA
[28]	Both	Both	-	10 kHz	-	NA
Proposed work	Both	Both	0.045 $\mu\text{m}$ GPDK	80 MHz	Zero	1.586 $\mu\text{m}^2$

FETs [6, 10, 12, 24, 25, 34, 49, 56, 57, 59] but operate on low operating frequencies and high power consumption compared to the proposed circuit. Some generalized memristors in [13, 28, 53] are based on passive and active components with low operating frequencies. This work presents a single-MOSFET memristor emulator with an R–C filter that operates on high frequency up to 80 MHz and consumes a meager dynamic power of 7.751 pW. However, because there is no external biasing, the static power is zero, and therefore, the proposed memristor emulator is entirely passive.

## 5 Conclusion

The manuscript proposes a minimally complex grounded memristor model which can operate in floating mode as well. It consists of a single MOSFET and an R–C tank circuit. The proposed memristor emulator circuit is simulated using Cadence Virtuoso Spectre tool. The main contribution is the experimental realization using discrete transistor such as MOSFET ALD1117 Dual P-channel enhancement MOSFET array and discrete elements in the form of an R–C tank circuit. The proposed topology can be either verified using discrete devices or designed using nanometer integrated circuit technology. Furthermore, the simulation results as well as experimental results are consistent with the three fundamental characteristics of a memristor. These characteristics are validated at different operating frequencies and temperatures. The proposed memristor emulator offers a number of advantages over the existing emulators, which include, a simpler and more adaptable design, a smaller number of active and passive components, a high operating frequency range (up to 80 MHz), low power consumption (7.751 pW dynamic and zero static power), lower area utilization ( $1.586 \mu\text{m}^2$ ), and suitability for IC fabrication. It is possible to use this memristor emulator model for both analog and digital applications. This work illustrates applications such as the Schmitt trigger, chaotic Colpitts oscillator, and XOR logical operation to support the use of the proposed memristor design. The main advantages of this emulator are (1) minimal complexity, (2) stable and reliable behavior, (3) high-frequency operation, (4) low power consumption, and (5) less area utilization, which makes a perfect candidate to be used in a wide range of applications, especially at higher frequencies. Disadvantages of this proposed circuit are the presence of an externally connected resistor and a capacitor, a large resistance value of the externally connected resistor of an order of a few 100 k $\Omega$ .

**Acknowledgements** The authors would like to thank the anonymous reviewers for their valuable comments.

**Data Availability** Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

## Declarations

**Conflict of interest** The authors whose names are listed in the manuscript certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or nonfinancial



interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

## References

1. M.T. Abuelma'atti, Z.J. Khalifa, A new memristor emulator and its application in digital modulation. *Analog Integr. Circ. Sig. Process* **80**, 577–584 (2014)
2. M.T. Abuelma'atti, Z.J. Khalifa, A continuous-level memristor emulator and its application in a multivibrator circuit. *Aeu Int J Electron Commun* **69**, 771–775 (2015)
3. G.C. Adam, B.D. Hoskins, M. Prezioso, F. Merrih-Bayat, B. Chakrabarti, D.B. Strukov, 3-d memristor crossbars for analog and neuromorphic computing applications. *IEEE Trans. Electron Devices* **64**(1), 312–318 (2017)
4. S.P. Adhikari, M.P. Sah, M.P.H. Kim, L.O. Chua, Three fingerprints of memristor. *IEEE Trans. Circ. Syst. I Regul. P.* **60**(11), 3008–3021 (2013)
5. H.A.F. Almurib, T.N. Kumar, F. Lombardi, Design and evaluation of a memristor-based look-up table for non-volatile field programmable gate arrays. *IET Circuit Device Syst* **10**(4), 292–300 (2016)
6. Y.R. Ananda, N. Raj, G. Trivedi, A MOS-DTMOS Implementation of Floating Memristor Emulator for High-Frequency Applications. *IEEE Trans Large Scale Integr VLSI Syst* **31**(3), 355–368 (2022)
7. Y. Babacan, A. Yesil, F. Kacar, Memristor emulator with tunable characteristic and its experimental results. *AEU Int. J. Electron. Commun.* **81**, 99–104 (2017)
8. Y. Babacan, F. Kacar, K. Gurkan, A spiking and bursting neuron circuit based on memristor. *Neuro-computing* **203**, 86–91 (2016)
9. Y. Babacan, F. Kacar, Floating memristor emulator with subthreshold region. *Analog Integr. Circ. Sig. Process* **90**, 471–475 (2017)
10. Y. Babacan, A. Yesil, F. Gul, The fabrication and MOSFET-only circuit implementation of semiconductor memristor. *IEEE Trans. Electron Devices* **65**, 1625–1632 (2018)
11. Y. Babacan, Memristor: Three mos transistors and one capacitor. Paper presented at the 21st International Conference on Intelligent Engineering Systems : proceedings, October 20-23, 2017, Larnaca, Cyprus (2017)
12. Y. Babacan, F. Kacar, FCS based memristor emulator with associative learning circuit application. *IU-J Electric Electron Eng* **17**(2), 3433–3437 (2017)
13. B. Bao, J. Yu, F. Hu, Z. Liu, Generalized memristor consisting of diode bridge with first order parallel RC filter. *Int J Bifurc Chaos* **24**(11), 1450143 (2014)
14. D. Batas, H. Fiedler, A memristor spice implementation and a new approach for magnetic flux-controlled memristor modeling. *IEEE Trans. Nanotechnol.* **10**(2), 250–255 (2011)
15. M. Chen, J. Yu, Q. Yu, C. Li, B. Bao, A memristive diode bridge-based canonical chua's circuit. *Entropy* **16**(12), 6464–6476 (2014)
16. M. Chen, M. Li, Q. Yu, B. Bao, Q. Xu, J. Wang, Dynamics of self-excited attractors and hidden attractors in generalized memristor-based chua's circuit. *Nonlinear Dyn.* **81**(1), 215–226 (2015)
17. L. Chua, Memristor-the missing circuit element. *IEEE Trans Circuit Theory* **18**(5), 507–519 (1971)
18. L.O. Chua, S.M. Kang, Memristive devices and systems. *Proc. IEEE* **64**(2), 209–223 (1976)
19. L.O. Chua, The fourth element. *Proc. IEEE* **100**(6), 1920–1927 (2012)
20. F. Corinto, A. Ascoli, Memristive diode bridge with LCR filter. *Electron Lett* **48**(14), 1 (2012)
21. A.S. Elwakil, M.E. Fouda, A.G. Radwan, A simple model of double-loop hysteresis behavior in memristive elements. *IEEE Trans. Circuits Syst. II Express Briefs* **60**(8), 487–491 (2013)
22. E. Gale, The memory-conservation theory of memristance. *UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, 599–604 (2014)
23. E. Gale, Non-ideal memristors for a non-ideal world. *Phys Status Solidi A* **212**(2), 229–238 (2015)
24. M. Ghosh, A. Singh, S.S. Borah, J. Vista, A. Ranjan, S. Kumar, MOSFET-based memristor for high-frequency signal processing. *IEEE Trans. Electron Devices* **69**(5), 2248–2255 (2022)
25. F. Gul, Circuit implementation of nano-scale tio<sub>2</sub> memristor using only metal-oxide-semiconductor transistors. *IEEE Electron Device Lett.* **40**(4), 643–646 (2019)
26. R. K. Gupta, M.S. Chaudhary, S. Taran, V. Saxena, A passive grounded MOSCAP-memristor emulator. In *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECT)*, 1–5 (2022)

27. A. I. Hussein, M. E. Fouda, (2013). A simple MOS realization of current controlled memristor emulator. In 2013 25th International Conference on Microelectronics (ICM) (pp. 1-4). IEEE
28. F. Jiang, F. Yuan, Y. Li, Design and implementation of xor logic circuit based on generalized memristor. *Eur Phys J Spec Top* **231**, 481–491 (2022)
29. J. A. Kalomiros, S. G. Stavrinides, F. Corinto, A two-transistor non-ideal memristor emulator. 2016 5th International Conference on Modern Circuits and Systems Technologies (MOCAST), 1-4 (2016)
30. H. Kim, M.P. Sah, C. Yang, S. Cho, L.O. Chua, Memristor emulator for memristor circuit applications. *IEEE Trans. Circuits Syst. I Regul. Pap.* **59**(10), 2422–2431 (2012)
31. I. Köymen, E. M. Drakakis, Cmos-based nanopower memristor dynamics emulator. In 2014 14th International Workshop on Cellular Nanoscale Networks and their Applications (CNNA), 1-2 (2014)
32. K. Kumar, B.C. Nagar, G. Pradhan, Single ota-based tunable resistorless grounded memristor emulator and its application. *J. Comput. Electron.* **22**(1), 549–559 (2023)
33. A. Kumar, B. Chaturvedi, J. Mohan, Minimal realizations of integrable memristor emulators. *J. Comput. Electron.* **22**(1), 504–518 (2023)
34. P. Kumar, P. Srivastava, R.K. Ranjan, M. Kumngern, New zero power memristor emulator model and its application in memristive neural computation. *IEEE Access* **11**, 5609–5616 (2023)
35. P. Mazumder, S.-M. Kang, R. Waser, Memristors: devices, models, and applications. *Proc. IEEE* **100**(6), 1911–1919 (2012)
36. B. Muthuswamy, P.P. Kokate, Memristor-based chaotic circuits. *IETE Tech. Rev.* **26**(6), 417–429 (2009)
37. P. Nune, S. Mandal, A. Saha, R. Saha, A generic simple model of synaptic memristor with local activity for neuromorphic applications. *J Comput Electron* **22**, 1–14 (2023)
38. Y.V. Pershin, M. Di Ventra, Practical approach to programmable analog circuits with memristors. *IEEE Trans. Circuits Syst. I Regul. Pap.* **57**(8), 1857–1864 (2010)
39. N. Raj, V.K. Verma, R.K. Ranjan, Electronically tunable flux-controlled resistorless memristor emulator. *IEEE Canad J Electric Comput Eng* **45**(3), 311–317 (2022)
40. R.K. Ranjan, N. Rani, R. Pal, S.K. Paul, G. Kanyal, Single CCTA based high frequency floating and grounded type of incremental/decremental memristor emulator and its application. *Microelectron. J.* **60**, 119–128 (2017)
41. R.K. Ranjan, N. Raj, N. Bhuwal, F. Khateb, Single DVCCTA based high frequency incremental/decremental memristor emulator and its application. *AEU-Int J Electron Commun* **82**, 177–190 (2017)
42. B. Rezavi, *Design of analog CMOS integrated circuits*, 2nd edn. (McGraw-Hill, New York, NY, USA, 2017)
43. C. Sánchez-López, J. Mendoza-Lopez, M. Carrasco-Aguilar, C. Muñoz-Montero, A floating analog memristor emulator circuit. *IEEE Trans. Circuits Syst. II Express Briefs* **61**(5), 309–313 (2014)
44. V. Saxena, A compact CMOS memristor emulator circuit and its applications. In 2018 IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS), 190-193 (2018)
45. H. Sozen, U. Cam, Electronically tunable memristor emulator circuit. *Analog Integr. Circ. Sig. Process* **89**, 655–663 (2016)
46. P. Srivastava, R.K. Gupta, R. Sharma, R.K. Ranjan, Mos-only memristor emulator. *Circuits Syst Signal Process.* **39**, 5848–5861 (2020)
47. D.B. Strukov, G.S. Snider, D.R. Stewart, R.S. Williams, The missing memristor found. *Nature* **453**(7191), 80–83 (2008)
48. J. Vista, A. Ranjan, Flux controlled floating memristor employing VDTA: Incremental or decremental operation. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* **40**(2), 364–372 (2020)
49. J. Vista, A. Ranjan, A simple floating mos-memristor for high-frequency applications. *IEEE Trans Large Scale Integr VLSI Syst* **27**(5), 1186–1195 (2019)
50. C. Wu, N. Yang, C. Xu, R. Jia, C. Liu, A novel generalized memristor based on three-phase diode bridge rectifier. Complexity (2019). <https://doi.org/10.1155/2019/1084312>
51. H.G. Wu, B.C. Bao, Q. Xu, First order generalized memristor emulator based on diode bridge and series RL filter. *Acta Electron Sinica* **43**(10), 2129 (2013)
52. Q. Xu, N. Wang, B. Bao, M. Chen, C. Li, A feasible memristive chua's circuit via bridging a generalized memristor. *J Appl Anal Comput* **6**(4), 1152–1163 (2016)
53. Y. Ye, J. Zhou, Q. Xu, M. Chen, H. Wu, Parallel-type asymmetric memristive diode-bridge emulator and its induced asymmetric attractor. *IEEE Access* **8**, 156299–156307 (2020)

54. A. Yesil, Y. Babacan, F. Kaçar, A new ddcc based memristor emulator circuit and its applications. *Microelectron. J.* **45**(3), 282–287 (2014)
55. A. Yesil, Y. Babacan, F. Kacar, Electronically tunable memristor based on vdcc. *AEU-Int J Electron Commun* **107**, 282–290 (2019)
56. A. Yesil, A new grounded memristor emulator based on MOSFET-C. *AEU -Int J Electron Commun* **91**, 143–149 (2018)
57. A. Yesil, Y. Babacan, Tunable memristor employing only four transistors. *AEU-Int. J. Electron. C.* **169**, 154763 (2023)
58. D. Yu, H.H.-C. Iu, A.L. Fitch, Y. Liang, A floating memristor emulator based relaxation oscillator. *IEEE Trans Circuit Syst I Regul P* **61**(10), 2888–2896 (2014)
59. L. Zhou, C. Wang, H. Qin, Q. Wang, A 300 MHz MOS-only memristor emulator. *AEU-Int. J. Electron. C.* **162**, 1434–8411 (2023)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

## Review

# A Spectrum of Solutions: Unveiling Non-Pharmacological Approaches to Manage Autism Spectrum Disorder

Arunima Mondal <sup>1,†</sup>, Rashi Sharma <sup>2,†</sup> , Umme Abiha <sup>3,4</sup>, Faizan Ahmad <sup>5,\*</sup> , Anik Karan <sup>6,\*</sup> , Richard L. Jayaraj <sup>7</sup> and Vaishnavi Sundar <sup>8</sup>

<sup>1</sup> Department of Human Genetics and Molecular Medicine, Central University of Punjab, Ghudda 151401, India

<sup>2</sup> Department of Biotechnology, Delhi Technological University, Bawana, Delhi 110042, India

<sup>3</sup> IDRP, Indian Institute of Technology, Jodhpur 342030, India

<sup>4</sup> All India Institute of Medical Sciences, Jodhpur 342005, India

<sup>5</sup> Department of Medical Elementology and Toxicology, Jamia Hamdard University, Delhi 110062, India

<sup>6</sup> CL Lab LLC, Gaithersburg, MD 20878, USA

<sup>7</sup> Department of Pediatrics, College of Medicine and Health Sciences, United Arab Emirates University, Al Ain 15551, United Arab Emirates

<sup>8</sup> Department of Internal Medicine, University of Nebraska Medical Center, Omaha, NE 68198, USA

\* Correspondence: 1996faizanahmad@gmail.com (F.A.); anik1432@gmail.com (A.K.)

† These authors contributed equally to this work.

**Abstract:** Autism spectrum disorder (ASD) is a developmental disorder that causes difficulty while socializing and communicating and the performance of stereotyped behavior. ASD is thought to have a variety of causes when accompanied by genetic disorders and environmental variables together, resulting in abnormalities in the brain. A steep rise in ASD has been seen regardless of the numerous behavioral and pharmaceutical therapeutic techniques. Therefore, using complementary and alternative therapies to treat autism could be very significant. Thus, this review is completely focused on non-pharmacological therapeutic interventions which include different diets, supplements, antioxidants, hormones, vitamins and minerals to manage ASD. Additionally, we also focus on complementary and alternative medicine (CAM) therapies, herbal remedies, camel milk and cannabidiol. Additionally, we concentrate on how palatable phytonutrients provide a fresh glimmer of hope in this situation. Moreover, in addition to phytochemicals/nutraceuticals, it also focuses on various microbiomes, i.e., gut, oral, and vaginal. Therefore, the current comprehensive review opens a new avenue for managing autistic patients through non-pharmacological intervention.

**Keywords:** neurodevelopmental disorder; ASD; neurotherapeutics; nutritional therapy; microbiome



**Citation:** Mondal, A.; Sharma, R.; Abiha, U.; Ahmad, F.; Karan, A.; Jayaraj, R.L.; Sundar, V. A Spectrum of Solutions: Unveiling Non-Pharmacological Approaches to Manage Autism Spectrum Disorder. *Medicina* **2023**, *59*, 1584. <https://doi.org/10.3390/medicina59091584>

Academic Editor: Rekha Jagadapillai

Received: 16 July 2023

Revised: 22 August 2023

Accepted: 29 August 2023

Published: 31 August 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

In 1908, a Swiss psychiatrist named Eugen Bleuler invented the terminology of “autism”, which originated from the ancient term “autós”, which signifies “self”, to characterize the detachment from reality of patients with schizophrenia [1–3]. Leo Kanner used the phrase in 1943 to describe linguistic and social isolation problems in children who did not have psychosis or other psychological illnesses. Such children struggled to engage and communicate with others, had a specific pattern of behavior, and were uninterested in social affairs [4–8]. One out of every 88 children has developmental difficulties, and this percentage seems to be rising. The frequency of autism in males and females is equivalent to approximately 5:1, affecting about 1.5% of the population [9–15]. The pathophysiology of ASD is not entirely known, and comorbidities, including epilepsy, attention, mood, and language impairments; sleep disturbances; gastrointestinal issues; and intellectual disability, are frequent (70% of cases). ASD is believed to be a developmental defect of brain processes brought on by genetic and neurological reasons, creating social disruption, which results in limited attentiveness and compulsive behaviors [16–20]. An aberrant

gene gets “turned on” during the early stages of fetal development, altering the body. Its expression can be changed without modifying the primary DNA sequence of other genes. The pathogenesis of ASD, which appears to be primarily driven by heterogeneous genetic mutations and variants and modulated by diverse gene–environment interactions, including pregnancy-related factors (such as maternal immune activation, maternal toxins, and perinatal trauma), may be a significant factor in the absence of disease-modifying therapies. Currently, there are few accessible pharmacological and non-pharmacological methods for ASD intervention. Different psychiatric drugs are used as pharmacological interventions, whereas specialized foods, herbal supplements, chiropractic adjustments, art therapy, mindfulness practices, and relaxation techniques are a part of non-pharmacological methods. As there are not any particular behaviors that aid in identifying people with ASD, it does not have a single management strategy. In addition to this, the cost of management of an autistic individual for a lifetime, as estimated in a study conducted in the USA, approximately amounts to around USD 3.6 million, which goes up as the case worsens. Apart from this huge cost, the constant care and support required are beyond estimation and are not a treatment that everyone can afford. Many parents have always turned to alternative therapies to help autistic children [21–38]. For different patients of ASD, specific CAM therapies, which include essential fatty acids, vitamins, an oligoantigenic diet, herbal remedies, and amino acids, are found to give favorable results. ASD nutritional dysfunctions should be considered part of the therapy/management process, as managing autism care is a complex condition for individuals and their families [39–42]. In this review, we focused on non-pharmacological interventions like different diets, supplements, camel milk, hormones, etc., and we also included different microbiomes, i.e., oral, vaginal, and gut microbiomes. This review will reveal a new horizon for the treatment and management of ASD. All non-pharmacological interventions are easily available on the market and are affordable, too. All non-pharmacological interventions can be used in combination with a low dose of pharmacological interventions, i.e., aripiprazole and risperidone. Non-pharmacological treatment has a far better response than relying on only drugs, as these drugs show adverse effects like weight gain, blurred vision, low blood pressure, seizures, low white blood cell count, drowsiness, dizziness, restlessness, dry mouth, constipation, and nausea. All non-pharmacological interventions mentioned in this review will help to manage the symptoms of ASD without showing adverse effects and are discussed in the coming sections of this review.

## 2. Diets for ASD

There is a need for additional management options that can improve outcomes for individuals with ASD. Different dietary supplements for the management of ASD are discussed below.

### 2.1. Elimination Diet for ASD

As the term signifies, some foods are avoided in the diet on the theory that particular ASD symptoms are related to foods that appear to be impacted by dietary hypersensitivities [43]. Such foods create gastrointestinal issues (GI) issues and raise IgG levels as the individual may be sensitive to the foods or their additives [44]. IgE and IgA antibody types have already been linked to immune dysfunction in people with autism. Diets must be closely controlled because removing foods to which an individual is allergic can cause malnutrition, which can worsen the symptoms of the disease by causing anemia. Findings show that adopting an exclusion diet regimen considerably improved the pathogenic alterations in autistic patients [45]. The popular elimination diet (gluten-free casein-free diet, GFCF) removed the proteins included in milk and cereals. The aforementioned diet calls for a decrease in or total removal of all the above-listed proteins [44]. Cows’ milk, cheese, and other dairy products contain casein, which, when removed, can cause a calcium deficiency since it is a crucial nutrient for bone and tooth health. Alternatives such as goat or sheep milk are frequently recommended but might require the body to confront new

allergens [45]. The elimination diet can lead to malnutrition if not carefully monitored, while the specific carbohydrate diet may be challenging to follow and restricts certain foods that are important for overall health. Additionally, some nutritional supplements may interact with medications or have harmful side effects if taken in excessive amounts.

## 2.2. Casein and Gluten for ASD

Gluten, milk, barley, rye, and wheat include casein, which has anti-inflammatory characteristics that regulate immune responses [46,47]. In particular, in persons with ASD, casein and gluten can promote the production of antibodies against IgA and IgG, worsening their immune dysregulation. The small intestinal mucosa works as a luminal barrier, keeping germs out, and such compounds are not permitted to enter the circulatory system. People with ASD, on the other hand, have higher intestinal permeability to such compounds, resulting in inflammation [48–52]. Casein and gluten products need to be consumed based on a clinician's advice, as these diets can cause inflammation at high doses, which can lead to other disorders.

## 2.3. Specific Carbohydrate Diet for ASD

A study by Gottschall, E. (2004) popularized this diet as a method of autism management. This diet's central premise is to prevent the advancement of pathogenic intestinal microflora's by alleviating malabsorption [53,54]. This diet recommends consuming monosaccharides, like those found in fruits, vegetables, and honey, rather than complex polysaccharides because polysaccharides take longer to digest [53]. Difficult polysaccharide digestion disrupts gastrointestinal tract function, resulting in absorption difficulty and the accumulation of left-over food. Intestinal pathogenic flora thrives in this food-accumulated environment [54]. This diet aims to help individuals lose weight, restore normal intestine functions, and minimize intestinal cancer formation. Meat, eggs, natural cheese, vegetables (pepper), cauliflower, onions, cabbage, spinach, homemade yogurt, fruits, nuts (walnuts, almonds), beans, and soaked lentils are all excellent protein sources and are recommended. Complex carbohydrates (e.g., sugar) are prohibited in the specific carbohydrate diet [54,55]. Only those foods that require minimal digestion are allowed. In a study conducted by Żarnowska et al. (2018), this diet was followed by people with Crohn's disease, both colonic and ileocolonic [55]. Symptoms improved after three years of monitoring. These findings may be generalizable across populations of people with ASD. Learning and memory were also highly improved, as were responsive and imaginative language difficulties [55]. Carbohydrate diets need to be consumed based on a clinician's advice and at the recommended dose as complex carbohydrate diets can cause different complications like inflammatory bowel disease (IBD).

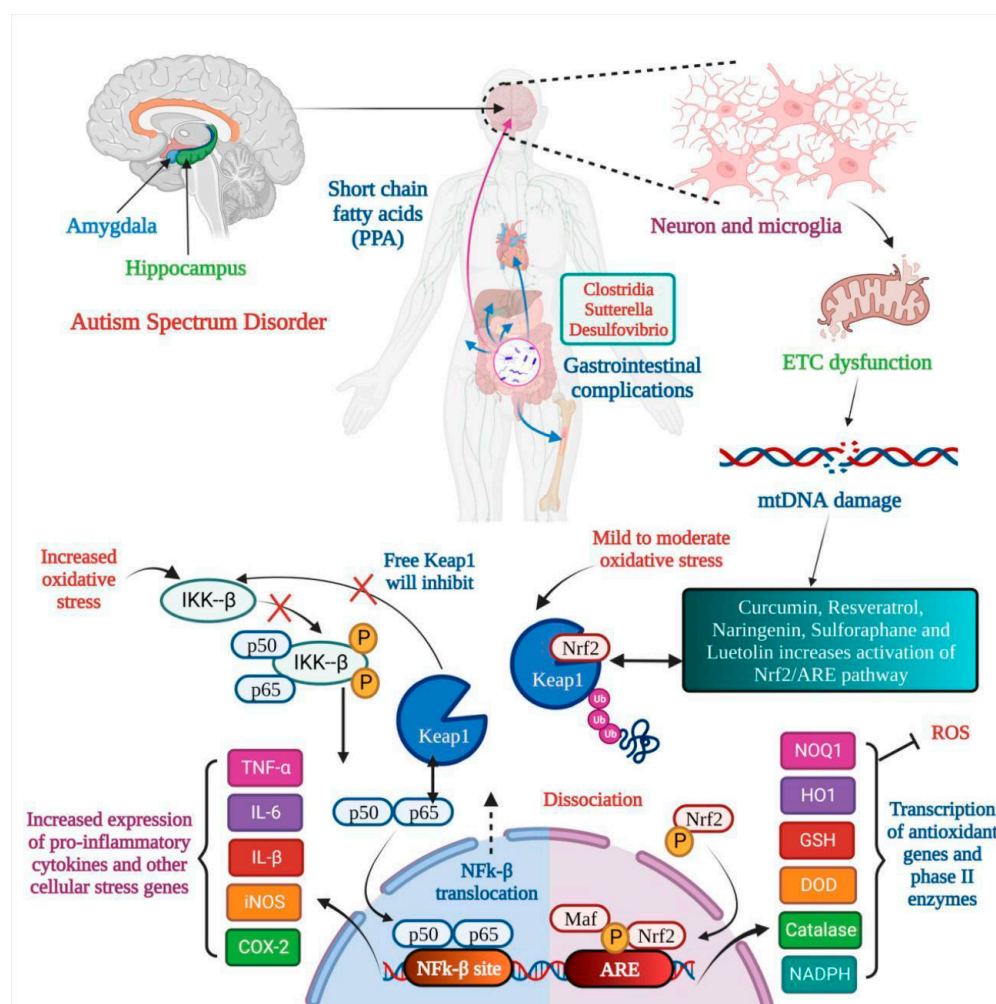
## 2.4. Ketogenic Diet for ASD

The ketogenic diet is a general term for a low-carbohydrate, moderate-protein, and high-fat diet that encourages our body to use ketones instead of glucose for energy. This results in more ketones in the blood, reduced blood glucose, and better functioning of mitochondria [56,57]. This diet has shown potential in treating patients with refractory epilepsy, which is considerably more typical in persons with ASD than those without ASD and other related nervous system problems [58,59]. A study by Kasprowska-Liśkiewicz D. et al. (2017) revealed that their sample exhibited fewer seizures and superior learning and social abilities [60]. El-Rashidi, O. et al. (2017) studies in people with ASD also revealed that the medication produced moderate improvements [61]. A ketogenic diet produces better responses and fewer complications in comparison to the elimination diet, casein and gluten diet, and carbohydrate diet, but still, the dose needs to be set by a clinician/dietician to avoid complications.



### 3. Nutritional Supplements for ASD

Numerous studies have suggested that poor behavioral evaluation test scores are continuously connected to low nutritional fulfillment. Hyperactivity, agitation, and irritability decrease when certain nutrient supplements are administered. Impulsivity and the inability to pay attention both improve dramatically [62]. During the day, a diverse mix of vegetarian and animal proteins is consumed to meet the daily need for amino acids. Amino acids (AAs), which have long been the basic building blocks of our bodies, make up proteins. The body may synthesize certain amino acids, but amino acids should be acquired from protein-rich diets [62]. The effect of nutritional therapy on ASD is shown in Figure 1. According to a study, neuroactive amino acids play a vital role in central brain activities. Neuroactive AAs are crucial in etiology and play a part in treating autistic symptoms [63]. It is also essential to watch for changes in their bodily fluid concentrations and see whether they correspond to early signs. Their availability, metabolism, and receptor functionality must all be considered [63]. They have been connected to the causes and therapies of numerous mental illnesses. More research is needed to see if other amino acids are involved. Discussed below are a few nutritional supplements.



**Figure 1.** Nutritional therapy attunes mitochondrial dysfunction in Autism Spectrum Disorder. In autism, the amygdala and hippocampus are affected, and gastrointestinal complications are seen. Microglia activation leads to electron transport chain (ETC) dysfunction, which results in mtDNA damage, and oxidative stress leads to dysfunction of the Nrf2 pathway and natural products like curcumin, resveratrol, etc., resulting in inactivation of the Nrf2/ARE pathway. Two outcomes are seen: (a) the transcription of antioxidant genes and (b) the increased expression of pro-inflammatory cytokines.



### 3.1. Omega-3 Fatty Acids for ASD

Omega-3 fatty acids are polyunsaturated fatty acids (PUFAs) recognized as -3 fatty acids or n-3 fatty acids. Triglycerides and phospholipids are two natural forms of omega-3 fatty acids. Fat is the most common component of brain nerve cells. Human physiology requires three omega-3 fatty acids, i.e., docosahexaenoic acid (DHA), alpha-linolenic acid (ALA) and eicosapentaenoic acid (EPA). Fish, eggs, and flax seeds are their most common natural sources [64]. PUFAs are essential for human health. The brain can generate neuronal signals in response to new experiences and stimuli. Neuronal plasticity, or the learning environment, is critical in long-term learning. DHA and omega-3 fatty acid levels must be balanced to maintain learning ability and enhance neuronal plasticity through membrane fluidity [64,65]. There is not much evidence to back up omega-3 supplementation's effectiveness in improving the core or linked symptoms of ASD. Three randomized controlled trials (RCTs) comparing omega-3 fatty acids to a placebo revealed no significant differences [64,65]. The placebo group performed significantly better in one trial than the control group. Parent ratings of stereotypy and weariness in children who took omega-3 supplements against those who did not show substantial improvement after six months of treatment compared to the omega-3 group in externalizing behaviors [64,65]. It is advised to consult a physician before consuming omega-3 fatty acids as high doses can cause nausea, loose stools, and stomach upset.

### 3.2. Zinc for ASD

Zinc, a mood mineral, is vital because it is a cofactor for numerous neurotransmitters that affect mood and learning. Low zinc amounts disrupt dopamine production as this neurotransmitter involves learning and emotions such as motivation and pleasure [66]. A lack of zinc affects normal neural activities, including neurotransmission, brain development, and connection; moreover, it indirectly impacts the brain by impairing the immune system and changing the usual gut–brain link. This metal is essential for the neuropeptide social impact. So, to avoid autism, expecting and new mothers take a zinc supplement in their diets [67,68]. High doses of zinc can cause acute gastrointestinal symptoms like abdominal pain, diarrhea, and vomiting, so it is recommended to consult a physician to determine the dose.

### 3.3. Vitamins for ASD

Most vitamins must be present in ideal amounts for healthy brain development. Vitamin D supplementation, in particular, has been demonstrated to help the symptoms of people with autism regress. Vitamins are potent antioxidants that help to protect cellular and mitochondrial function from free radical damage [66,68]. They also function as cofactors in a variety of biological processes. They regulate lipid and protein metabolism and are crucial for DNA synthesis. According to a study by Rollett, A. (1909), reduced folate levels during pregnancy are also related to congenital impairments. It has been connected to hyperactivity in youngsters. Autistic youngsters have also been proven to benefit from vitamin B1. Vitamin C has twin benefits, firstly as an antioxidant and secondly in creating some neurotransmitters [66,67]. Researchers have also determined that a specific gene encoding a particular protein is missing, which is the protein necessary to produce vitamin A. Clinical tests showed that vitamin A treatment enhances language and visual abilities in autistic patients. On the other hand, vitamin A supplementation must be carried out under the supervision of a physician as it can cause liver damage [68].

### 3.4. Iron for ASD

In autistic people, malabsorption of the vitamin inside the gastrointestinal system and their selective eating habits can lead to iron insufficiency. As a result, an iron shortage is reported to negatively affect sleep and neuroprotection. According to specific clinical investigations, cognitive impairment, reduced development, attention issues, and anemia are all related to mood swings in autistic children [69]. Children who have ASD have

been found to have a high prevalence of iron deficiency (ID) and anemia that occurs due to iron deficiency (IDA). There are a small number of studies that link autistic clinical symptoms and iron deficiency indicators. The current research compares the levels of HB, hematocrit, Fe, ferritin, mean corpuscular volume, and red cell distribution width in patients with autism and healthy controls to determine the relationship between the numbers and symptoms. Children with ASD had lower HB levels than children without the disorder. Instead of the intensity of autistic symptoms, IDA in children with ASD may be linked to mental retardation [70]. A high dose of iron can cause iron poisoning, which shows multiple symptoms like nausea, abdominal pain, fever, headache, seizures, etc., so the dose needs to be advised by a physician to avoid complications due to a high dose of iron.

### 3.5. Magnesium (Mg) for ASD

Magnesium (Mg) works synergistically to relieve the clinical signs of autism. When autistic youngsters were given magnesium and vitamin B6, their social interaction and speech increased by 70% [71]. The most recognizable symptoms and indicators of Mg shortage are caused by neuronal and neuromuscular overactivity. In general, the connections between magnesium levels in inverse and direct sites and neurodevelopmental disorders may be a sign of higher excretion of magnesium in children with ASD, which ultimately results in a lower burden of magnesium in the body. The lack of noticeable changes in serum Mg levels may result from homeostatic regulation, which regulates absorption, excretion, and tissue redistribution (particularly in the bones) to maintain circulating Mg levels. Mg has a tremendous impact on neural excitation. Stress-related physical damage is more susceptible to Mg shortage, and Mg supplementation is protective. The neurologic impairment caused by experimental head trauma can be reduced pharmacologically by Mg, probably via the blockage of N-methyl-D-aspartate receptors. Mg salts help around 40% of people with autism when taken with large doses of pyridoxine, perhaps because they impact dopamine metabolism [72]. A high dose of Mg can cause diarrhea, vomiting, depression, low blood pressure, etc., so it is recommended to consult a physician before consuming Mg.

### 3.6. Selenium for ASD

Numerous vital metabolic processes for life depend on selenium. Countless studies have shown that the neuro-endocrine-immune network plays an important role in the interaction between the intestinal microbiota and the brain that impacts autism, and some animal studies have suggested that the gut microbiota may compete with the host for selenium when its accessibility in the organism becomes limited [73]. Selenium at high doses can cause nausea, bad breath, and fever as well as severe problems in the heart, liver, and kidneys, so it is advised to consult a physician before consuming selenium.

## 4. Antioxidants for ASD

Antioxidant supplementation has been demonstrated to improve behavioral symptoms and reduce cognitive loss in patients with autism [74,75]. The following sections discuss several antioxidant substances that have demonstrated potential benefits for individuals with ASD.

### 4.1. Curcumin for ASD

The ingredient in turmeric (*Curcuma longa*), sometimes known as “Indian Solid Gold”, is curcumin. It possesses anti-inflammatory and antioxidant properties and inhibits angiogenesis and cell adhesion. It also inhibits crucial cell signaling pathways, i.e., NF- $\kappa$ B and PI3K, indicating anticarcinogenic capabilities [74–79]. Curcumin’s neuroprotective effects make it useful in treating neurodegenerative illnesses, including Alzheimer’s Disease (AD), Huntington’s Disease (HD), Parkinson’s Disease (PD), and peripheral neuropathy [80–82]. Curcumin targets several cell signaling pathways, and its effects are as follows: increasing

intracellular levels of glutathione, reducing inflammatory components, mitochondrial dysfunction, oxidative/nitrosative stress, and protein aggregation, counteracting the damage caused by heavy metals, and supporting liver detoxification. Its anti-proliferative impact on neurons is the principal method of preventing brain-stimulated microglia and reactive astrocytes from releasing cytokines and other active elements [83,84]. Shu-Juan et al. 2012 [85] showed that curcumin has neurotherapeutic potential by improving autistic behavior and boosting brain-derived neurotrophic factor (BDNF) levels in sodium valproate rat models of autism. For two weeks, 35-day-old rat pups were administered a 10 g/L concentration of curcumin. Their social interactions improved significantly and repetitive behavior decreased, and there were increased BDNF levels in the temporal brain [84–90]. Still, the role of curcumin in autistic phenotypes remains unclear.

#### 4.2. Resveratrol for ASD

Resveratrol is a phenolic acid stilbenoid produced when bacteria and fungi attack plants. It can be found in berries, grapes, and almonds. Resveratrol is effective against oxidative stress and immune function and has the potential to cross the blood–brain barrier (BBB). Resveratrol interacts with a variety of targets, is multifactorial in nature, and inhibits cyclooxygenase (COX), activates sirtuin (silent mating type information regulation homolog 1—SIRT1), induces endothelial nitric oxide synthase (eNOS), and activates peroxisome proliferator-activated receptors (PPARs) [91–95]. Resveratrol allosterically modulates the regulatory target SIRT1. It promotes AMP-activated protein kinase (AMPK) phosphorylation and decreases oxidative damage in F2 hybrid mice [94,95]. Fontes-Dutra et al. 2018 [96] looked at the neurotherapeutic potential of resveratrol (RSV) in a valproic acid (VPA) animal model of autism. The study's primary purpose was to see the neurodevelopmental abnormalities that might be caused by prenatal valproic acid exposure and whether resveratrol could be utilized as a treatment [96]. The effects of resveratrol on sensory behavior were investigated after autism was induced in rats. The location of GABAergic parvalbumin (PVC) neurons in sensory brain areas and the expressions of excitatory and inhibitory synapses were studied [97,98] in rats with an ASD-like phenotype produced by propionic acid (PPA). Resveratrol was administered at 5, 10, and 15 mg/kg [99]. The therapy started the day following the operation and lasted for 28 days. Rats were subjected to behavioral tests between the 7th and the 28th days. Sociability, repetitive conduct, anxiety, unhappiness, and item recognition tests and the Morris water maze test for perseverance were some of the behavioral tests used [100]. They discovered that matrix metalloproteinase-9 (MMP-9) activation caused mitochondrial dysfunction and the production of inflammatory cytokines. Resveratrol restored the core and associated symptoms of autistic phenotypes by suppressing oxidative–nitrosative stress, mitochondrial dysfunction, and TNF- $\alpha$  and MMP-9 expression. Based on their findings, we can say that resveratrol can be a promising therapeutic intervention for the management of ASD [101–103].

#### 4.3. Naringenin for ASD

Flavanone Naringenin (NAR) is abundant in grapefruit, oranges, and tomato skin [104]. Naringenin inhibits human cytochrome P450-metabolizing enzymes of the CYP1A2 isoform [105,106]. Naringenin has an antioxidant effect by inhibiting the NF- $\kappa$ B pathway, which reduces oxidative injury caused by radiation exposure in mice. It also has an antihyperlipidemic effect by preventing the secretion of very low-density lipoproteins (VLDL) [106,107]. BDNF signaling also has antidepressant potential in chronic unpredictable mild stress. Because of its ability to inhibit cell proliferation by binding to estrogen receptors, it has versatile functions in many cancers [108]. It is also beneficial in treating osteoporosis, cancer, and cardiovascular diseases. Naringenin also suppresses neuroinflammation in glial cells by triggering the suppressor of cytokine signaling 3 (SOCS3)-3. Because the NF-B pathway is inactivated, it had a neuroprotective role in a middle cerebral artery occlusion (MCAO) model of ischemic stroke (IS) [109–111]. Naringenin works on ASD in a similar manner as it works on IS. In their study, Bhandari et al. (2018) recently looked at the

neurotherapeutic potential of naringenin, naringenin-loaded glutathione, and Tween-80-coated nanocarriers in treating ASD [112]. They helped to reverse the neuropathology that had developed. PPA administration improved brain uptake by avoiding naringenin's low oral bioavailability and at a low oral dose of 25 mg/kg. As a result, these brain-targeted naringenin nanocarriers can be used in clinics as a neurotherapeutic [113,114]. Based on previous findings, we can suggest that naringenin can be a promising non-pharmacological therapeutic approach to the management of ASD.

#### 4.4. Sulforaphane for ASD

Sulforaphane is a phytochemical belonging to the isothiocyanate group and is chemically recognized as 1-isothiocyanate-4-(methylsulfonyl) butane. The photocatalytic activity of myrosinase produces sulforaphane, once glucoraphanin is metabolized [115,116]. Glucoraphanin is a precursor to sulforaphane and can be found in various vegetables. Broccoli and cauliflower belong to the cruciferous family. According to Yuesheng Zhang [117], sulforaphane is beneficial for reducing oxidative stress in the human body, reducing mitochondrial dysfunction. It is a neuroprotective compound that prevents apoptosis in hippocampal neurons due to oxidative stress and the production of free radicals. It also has anti-diabetic properties [118,119]. It has anticarcinogenic and anti-inflammatory properties, aiding in lowering the infarct volume after ischemic stroke. It works by activating both the nuclear factor erythroid 2-related factor 2 (Nrf2)-dependent and the Nrf2-independent self-contained pathways [120,121]. Astrocytes activate the Nrf2 response, and heat shock protein 27 is upregulated [122]. Singh et al. 2014 [123] investigated the neurotherapeutic effects of sulforaphane in young men aged 13 to 27 with moderate-to-severe ASD. This was a placebo-controlled, randomized, double-blind clinical trial. The patients were given sulforaphane at a 50–150 mol/day dose for 18 weeks, and the results were monitored during a four-week drug-free period [124]. There were 29 ASD patients and 15 control subjects in a placebo-controlled study. After 18 weeks of treatment, their behavior was assessed using behavioral rating scales. According to the findings, treatment with sulforaphane improved the patients' social interaction abilities and reduced the deficits overall [124]. This was demonstrated by an increase in behavioral assessment scores using behavioral rating scales such as the Aberrant Clinical Global Assessment (CGA), Autism Behavior Checklist (ABC), Social Responsiveness Scale (SRS), and Clinical Global Impression—Improvement (CGI-I). As a result, broccoli's sulforaphane can decrease oxidative stress neuroinflammation and prevent DNA damage [124]. Sulforaphane appears to be a safe and effective management option for ASD and other neurological disorders.

#### 4.5. Luteolin for ASD

Luteolin is a natural flavonoid with anti-inflammatory, anti-antioxidant, and neuroprotective properties and can easily penetrate the BBB due to its low lipophilicity. Luteolin has the potential to neutralize ROS and downregulates IL-1 $\beta$ , IL-6, and TNF- $\alpha$ , which might counteract neuroinflammation. It also inhibits the stimulation of astrocytes, as well as microglial activation and proliferation [125]. In a mouse model, luteolin inhibited IL-6 release from activated microglia and reduced maternal IL-6-induced autism-like behavioral deficits related to social interaction [125]. In a recent study by Marianna et al. 2022, chronic luteolin treatment ameliorated hyperactivity, memory, and motor skills in 3dKL5 +/– mice by inhibiting neuroinflammation [126]. Even luteolin shows positive results in clinical models, which is discussed in Table 1. Based on the preclinical and clinical data, luteolin can be an excellent natural medicine for the management of autism.

### 5. Camel Milk for ASD

Camel milk may have recently been used to treat several illnesses, including food allergies, diabetes, hepatitis B, autism, and other autoimmune diseases. In patients suffering from reduced plasma glutathione peroxidase (GSH-Px), superoxide dismutase (SOD) and cysteine were linked to ASD, and the effect of camel milk was documented. It showed

improvement in ASD clinical outcomes [127,128]. Camel milk has more essential nutrients, like Ca, Fe, Mg, Cu, Zn, K, vitamin A, vitamin B2, vitamin C, and vitamin E, than the milk of other herbivorous animals. Moreover, camel milk lacks beta-lactoglobulin and beta-casein, two vital active ingredients that are components of cows' milk and cause milk allergies [129]. Camel milk includes different preventive biomolecules with antimicrobial, anti-viral, and immunologic characteristics. It contains anti-inflammatory protein molecules and antibodies that aid in easing specific primary autistic symptoms [130]. Those antibodies have new structural features with better tissue penetration and hidden epitopes. These characteristics may help avoid infections and provide potential benefits [131]. Furthermore, the nanobody structure of camel milk is highly comparable to the antibodies of immunoglobulins from humans (IgG3) [131]. This implies that the antibodies from camels are similar to the antibodies from humans. Camel milk has a unique composition that allows it to be used in various ways. Improvements in children with autism have been shown to occur. ASD can be treated by increasing superoxide dismutase levels and Plasma GSH, as well as by lowering oxidative stress, which is a component of the etiology of autism [127,128]. According to Gader et al. 2016 [132], camel milk reduces cancer risk. Symptoms of autism have improved significantly, or there has been a significant improvement in basic skills. A study by Al-Awadhi LY et al. (2015) suggests that the antioxidant enzymes and non-enzymatic antioxidant molecules found in camel milk could help to improve typical ASD behaviors [133]. Large-scale dose-focused investigations are necessary to verify the impact of camel milk on oxidative stress parameters and the therapy of autism [133]. Camel milk could be a promising therapeutic intervention for the management of ASD.

## 6. Hormone Therapies for ASD

Children and men with autism have improved social interaction and speech following hormone therapy [134]. Below we discuss in detail studies related to different hormone and the results of hormone therapies using melatonin, oxytocin, and vasopressin.

### 6.1. Melatonin for ASD

Melatonin use for curing sleep disturbances in children with ASD is supported by research. A recent significant randomized, double-blind placebo control (RDBPC) study found that pediatric-appropriate prolonged-release melatonin mini tablets (PedPRM) enhanced bedtime and quality of sleep, with an enhancement in total sleep time and sleep quality [134–138]. Finally, latency and sleep disturbances were reduced. Aside from its sleep-related benefits, a few RDBPC studies have shown that melatonin has been shown to improve communication, rigidity, and mood and reduces anxiety and depression in children with ASD [134–138]. Melatonin could be a promising hormone therapy for the management of ASD.

### 6.2. Oxytocin for ASD

According to new evidence, neuropeptides like oxytocin could be helpful in treating core ASD symptoms. According to a recent meta-analysis, oxytocin does affect children's social cognition and restricted and repetitive behaviors (RRBs) in autism spectrum disorder (ASD). The results of an unreviewed RDBPC study of oxytocin levels in students with autism were presented during an oral exam presentation at the International Autism Conference [139,140]. According to a 2017 International Meeting for Autism Research (IMFAR) study, oxytocin was not superior to other hormones compared to a placebo in reducing social withdrawal. However, it outperformed the placebo in improving social recognition. In contrast to the placebo, oxytocin improved social functioning as evaluated using the Social Responsiveness Scale (SRS) in a recent study of children with ASD RDBPC [139,141]. Based on previous findings, oxytocin could be a promising hormone therapy for the management of ASD.



### 6.3. Vasopressin for ASD

Vasopressin's improved reactions to personal interaction and communication and proof from preclinical studies recommend that Vasopressin receptor 1A (V1AR) antagonists may have pro-social advantages for disorders in which emotional and social functions are core deficits, such as schizophrenia [142]. Antidepressant and anxiolytic properties are also present. The safety and effectiveness of V1a antagonists in ASD have been studied in a few clinical trials [142]. In one study, the vasopressin V1a antagonist RG7713 was given intravenously to adults with high-functioning ASD to enhance social communication. Balovaptan (RG7314), another vasopressin V1a antagonist, has shown promise in improving communication and social interaction in ASD patients. The Food and Drug Administration (FDA) recently granted PB2452 status as one such compound with "Breakthrough Therapy Designation", raising hopes for approval of the first pharmacological method to enhance core social communication deficits in ASD. Vasopressin can cause multiple adverse effects like allergic reactions, stomach pain, nausea, irregular heartbeats, low blood sodium levels, low blood pressure, and pale skin [140–142]. Based on previous findings, we can suggest that vasopressin could be a promising hormone therapy for the management of ASD.

### 7. Herbal Medicine for ASD

A large variety of herbal remedies, which include *Ginkgo Biloba*, *Zingiber officinale* (ginger), *Astragalus membranaceus*, *Centella Asiatica* (Gotu cola), and *Acronis Calamus* (Calamus), may have therapeutic benefits in ASD patients due to their somatic effects such as increasing cerebral blood flow circulation, cognitive function enhancement, soothing or sedative effects, and enhancing the immune system's response [143,144]. A recent comprehensive review found that herbal remedies are safe when used in conjunction with conventional therapy and have positive benefits in controlling the aberrant behaviors and inattentiveness of ASD patients [143,144]. However, such results should be interpreted with caution due to the dearth and ambiguity of available data. Additional limitations on how these treatments can be used have been created due to risks associated with herbal interactions that may arise with other drugs and the questionable sources of herbs. Additional clinical trials are required to advance the research and validate Ayurvedic remedies' potential therapeutic benefits in ASD [143,144].

### 8. Cannabidiols for ASD

Cannabidiol (CBD) is a phytocannabinoid in the cannabis plant *Cannabis sativa* (marijuana). Although marijuana contains hundreds of phytocannabinoids, CBD is the second most prevalent after delta-9-tetrahydrocannabinol (THC), which has psychoactive characteristics [145]. Marijuana has been used for fiber and therapeutic purposes in India, China, and the Middle East for over 8000 years. It was spread to Europe by Napoleon's soldiers arriving from Egypt in the 1800s, and a doctor who campaigned in India subsequently brought it to Britain for medicinal use [145,146]. When used sufficiently, tetrahydrocannabinol (THC), phytocannabinoids' principal psychoactive component in *Cannabis sativa*, can aggravate various neurological conditions. Cannabidiol (CBD) may also help to decrease autistic behavior. This chemical offers therapeutic benefits such as immunomodulation, antioxidant defense, and neuroprotection with few or no adverse effects [147,148]. Sleeplessness, tension, discomfort, and even motor deficits like PD trembling are among the medical diseases that are treated using the non-psychoactive component of marijuana known as CBD. Some of the mechanisms through which BD exerts its neuromodulatory and neuroprotective effects include agonist potentiation, oxidative activity enhancement, 5HT1A transmitter engagement, and anandamide level augmentation. [148–151]. The endocannabinoid system (ECS) contains CB1 in the central nervous system (CNS) and CB2 throughout the body and immune system. The ECS regulates cognition and behavior by modulating synaptic transmission across the CNS. The ECS consists of endocannabinoids, cannabinoid receptors, and enzymes for synthesizing and degrading these endocannabinoids. Due to neurological signaling, endogenous cannabinoids are created and liberated

from the phospholipid bilayer attached to the postsynaptic barrier [148–151]. By activating cannabinoid sensors on the neuromuscular junction and preventing transmitter release from the presynaptic cell, they function backward, signaling molecules across the synaptic gap. Phospholipase C and diacylglycerol lipase (DAGL) are two enzymes involved in synthesizing 2-arachidonoylglycerol (2-AG) [151]. Additionally, it has been revealed that cannabidiol acts as a positive allosteric modulator at GABAA receptors, and observational trials have shown that CBD in the formulation of Epidiolex is indeed an effective analgesic in the treatment of Lennox–Gastaut disorder and Dravet syndrome [152,153]. By controlling the balance of interneurons transmission, CBD’s potential to increase endogenous cannabinoid production levels and promote the GABAergic transfer of information may aid in restoring neuronal function and neuroplasticity [152,153]. In ASD and Fragile X Syndrome (FXS), in which patients do not have seizures, CBD treatment has been proven to help in both animal and human models. Cannabidavarin (CBDV), also a chemical in the Cannabis sativa plant, is now being studied in animal models of ASD. Recently, discovering a THC-free topical CBD has allowed for their investigation in both ASD and FXS [154,155]. Hessler et al. 2019 [156] undertook an accessible trial in Australia utilizing dermal cannabidiol at a dose of 250 mg bi-daily, for twelve weeks, on children with FXS aged 3–17 years old. ZYN002 is a patent-protected clear-gel pharmaceutically manufactured synthetic cannabidiol with a permeation-enhanced formulation for effective transdermal delivery. The significant endpoint, the Stress, Distress, and Emotion (ADAMS) scale, and the secondary measures, including the ABC, exhibited effectiveness [157–160]. In conclusion, a nationwide research experiment including almost 200 children with FXS was carried out, and the significant consequence predictor was only demonstrated by children with FXS who had more than 90% methylation using the ABC and the FX Social avoidance subscale, a scale designed for FXS and adapted from the ABC [159]. Zyn002 is now undergoing a second international experimental investigation to obtain FDA approval for general use, although it still needs this approval. There is a good chance that the upcoming observational trials for autism and FXS will assist some subpopulations, including both diseases, and that marijuana use might increase [156,160]. Based on previous findings, cannabinoids could be a promising therapeutic intervention for the management of ASD. Table 1 represents the names of natural products and clinical data related to autistic children and adults.

**Table 1.** Published clinical data on different nutritional therapies for management of autistic children and adults.

Name of Natural Product	Clinical Model of ASD	Method and Duration	Result	References
Cannabidiol	150 participants (5–21 years of age)	Entire plant cannabis extricate that included cannabidiol and $\Delta^9$ -tetrahydrocannabinol at a 20:1 ratio and distilled cannabidiol and $\Delta^9$ -tetrahydrocannabinol at a 20:1 ratio.	Whole-plant extract showed 49% improvement in behavior with no severe effect, with common adverse effects like decreased appetite and somnolence.	[146]
Cannabidiol	18 autistic patients	Observational study	Cannabidiol and enriched <i>Cannabis sativa</i> extract decreased multiple ASD symptoms, even in epileptic patients	[159]
Luteolin	Children ( $n = 37$ , 4–14 years old)	Children were given luteolin as well as another supplement for four months	50% improvement in eye contact and attention, 25% improvement in social skills, 10% improvement in speech, and 75% improvement in GI	[161]
Luteolin	10-year-old male child	Co-ultra peak-LUT	Improved almost all symptoms	[162]



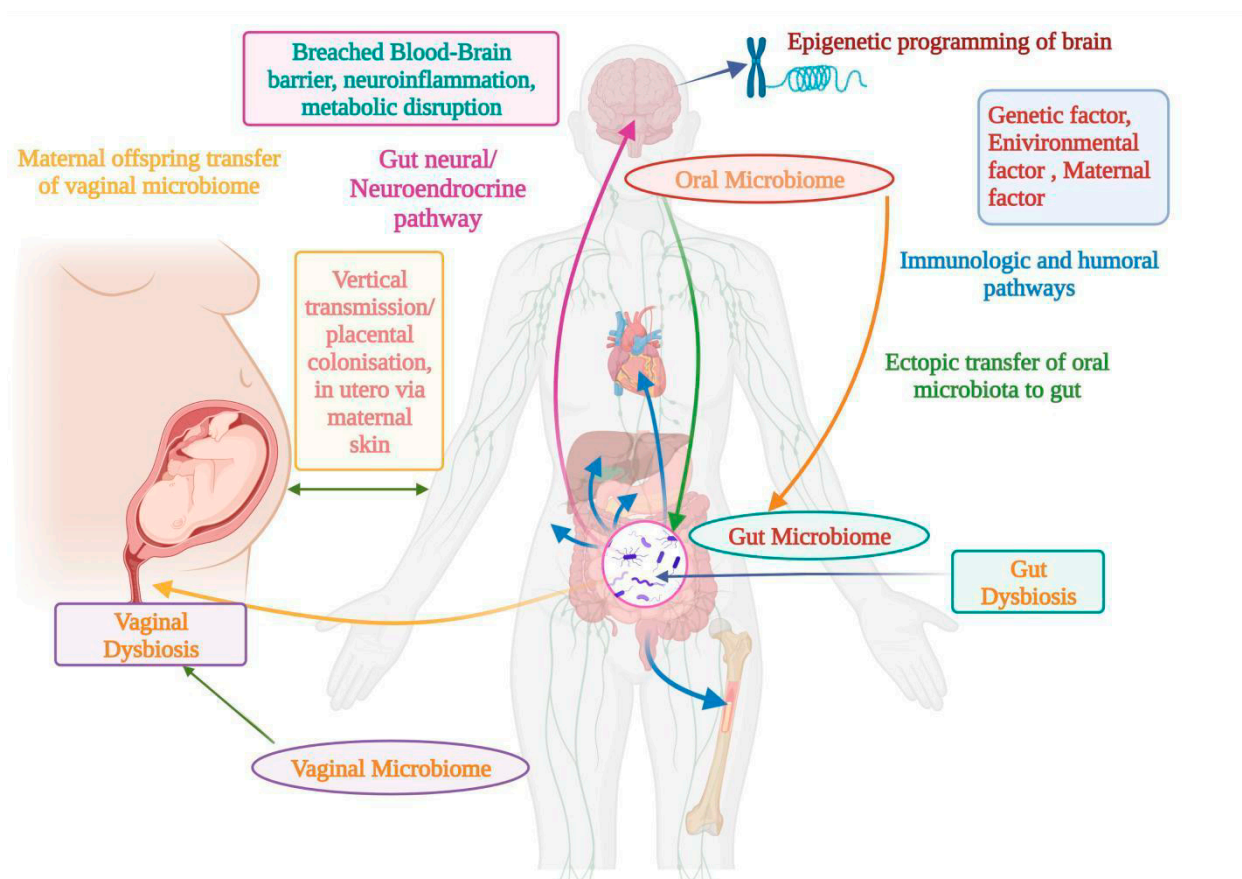
Table 1. Cont.

Name of Natural Product	Clinical Model of ASD	Method and Duration	Result	References
Luteolin	Children	Based on serum levels of IL-6 and TNF	Reduction in IL-6 and TNF levels after 26 weeks of treatment improved behavior	[163]
Luteolin	50 children aged 4–10 years old (42 boys, 8 girls)	Open-label trial, one capsule per 10 kg of weight per day with food	Decreased all clinical signs with no significant harmful effects	[164]
Cannabidiol	34 healthy men (half with ASD), 600 mg cannabidiol taken via oral administration	fMRI response to cannabidiol in ASD	Cannabidiol altered the fractional amplitude of low-frequency fluctuations.	[165]
Cannabidiol	34 healthy men (17 neurotypical and 17 ASD)	A single oral dose of 600 mg cannabidiol or placebo	Modulated glutamate GABA system	[166]
Cannabidiol	188 ASD patients	Medical cannabis treatment	28 patients showed significant improvement, 50 moderate improvement, 6 slight improvement, and 8 no improvement	[167]
<i>Ginkgo biloba</i> extract (Ginko T.D., Tolidaru, Iran)	3 autistic patients	2 × 100 mg, four weeks	Improvement in behavior	[168]
Camel milk	Total of 45 children, three groups of 15 children each	Blood samples for activation-regulated chemokine (TARC) serum level and childhood autism rating scale (CARS) score were taken before and after participants consumed 500 mL of milk per day in their daily diet for two weeks.	Reduced level of TARC and improvement in CAR score	[169]
Gluten-free diet	80 children, two groups (one regular group consisting of 40 children and one gluten-free diet group consisting of 40 children), and 53.9% of children had gastrointestinal abnormalities	The Rome questionnaire was used to examine gastrointestinal symptoms, and the Gilliam Autism Rating Scale 2 (GARS-2) was used to assess psychometric qualities.	Reduction in gastrointestinal symptoms and ASD symptoms	[170]
GFCF diet	37 patients, six months on a regular diet and six months on GFCF	Questionnaires regarding behavior	No change in behavior after consumption of GFCF for 6 months	[171]
GFCF diet	14 children, 3–5 years age	12-week double-blind, placebo-controlled trial study with continuation of the diet, with a 12-week follow-up and dietary supplement delivered via snacks	No change in behavior or other autism symptoms	[172]
Modified ketogenic, gluten-free diet	15 children, 2–17 years of age	Open-label clinical trial for three months	Improvement in autism symptoms	[173]
Vitamin and omega 3	111 children	Trial: Vitamin D (2000 IU/day), omega-3 LCPUFA (722 mg/day EPA and DHA, OM), or both for 12 months.	Vitamin D and omega-3 LCPUFA reduced irritability symptoms	[174]

## 9. Relationship between Microbiome and ASD

Our bodies contain several hundred million microbial colonies that code a hundred times the additional genetic traits of human genetics, including the most recent update predicting a microbial proportion of 1.3 species per individual cell, down from the highly

cited 10:1 and 100:1 ratio [175]. With the massive microbial spread, the human body's microbiome can play therapeutic and pathogenic roles [176,177]. Animal models and human subjects have been used to study microbiomes and ASD. The maternal influence on early intestine development is crucial. Maternal colonization in the offspring is frequently influenced by natural factors [176,177]. The microbiota's perinatal and prenatal periods have an impact on the microbial makeup of the uterus. The olfactory system mediates the spread of harmful oral germs via ectopic translocation. BBB disruption and perivascular and circumventricular space neuroinflammation may be caused due to an oxidative disorder in the brain, implying that the oral microbiota and dysbiosis impact the brain [178,179]. Another possible pathway is believed to be the gastrointestinal system and nervous axis, which regulate the oropharynx. It has a crucial function in ASD pathology. In general, the interaction is positive. A complex mechanism exists between the microflora and the nervous system in autism. Short-chain fatty acids (SCFAs) and BDNFs, among other natural and hereditary factors, may explain, control, and modulate epigenetic pathways. The oropharynx, which plays a vital role in the pathophysiology of the buccal space, is thought to be a mediator in ASD [180–183]. Figure 2 explains the relationships between different microbiomes and ASD.



**Figure 2.** Linkage between ASD and oral, gut, and vaginal microbiomes, along with different factors affecting pathways and their effects on neuroinflammation and metabolic disruption.

### 9.1. Oral Microbiome for ASD

According to the theory, microbial perturbations in the stomach might migrate to the oropharynx and affect the oral microbiome [184]. Children with ASD usually have speech issues and are incredibly picky eaters. Therefore, the oral microbiota was identified as a possible diagnostic sign for ASD [185]. The oropharynx is one of the essential parts of the digestive system. Five sensory motor cranial nerves connect it to the rest of the body and link it to the GI tract. It is thought that the development of autism may be

significantly influenced by a potential exchange mechanism within the brain [186]. The gut–brain axis allows for a connection between the gut and the brain, confirming this theory. Oral bacteria enter the brain, causing responses like inflammation, metabolism disturbance, and spinal cord infection [187]. The olfactory nerve is supposed to work as a sensor in the olfactory tract. The bacterial dispersion to the brain via blood circum-ventricular organs or perivascular spaces is thought to be mediated by a damaged BBB. *Haemophilus parainfluenza*, a Gram-negative bacterium, and its metabolites have been linked to oral diseases. Even routine dental procedures can cause bacteremia, and a portion of these microbes may traverse the BBB. Altered transcript expression has been described in the microglia of ASD individuals, and disrupted microglia function could impair BBB integrity. This could expose the brain to bacterial metabolites, thereby triggering an inflammatory response and altering metabolic activity within the central nervous system. The prolonged disruption of energy metabolism within neurons, oligodendrocytes, and glia could lead to structural changes in the cortex, hippocampus, amygdala, or cerebellum, which have all been documented in ASD individuals to be increased. They could penetrate the BBB, harm the nervous system, and cause ASD. This could expose the brain to bacterial metabolites, thereby triggering an inflammatory response and altering metabolic activity within the central nervous system. Gram-negative, putative periodontal pathogens, are rich in lipopolysaccharides (LPS), which exhibit pro-inflammatory activity. The leakage of LPS through the BBB in ASD individuals could lead to inflammation in the central nervous system (CNS) [188–192]. The microorganisms seen in patients with periodontal disease are not common in healthy people. Numerous health problems have been linked to periodontal disease [193]. The risk of early birth increases by 2–7 times since the organisms that cause intrauterine infections have been found in the mouth rather than the urogenital tract. [193–195]. Placentae have been shown to indicate the mouth microbiota more than the vaginal microbiome, signifying hematogenous spread, especially in underlying periodontal disease and oral intercourse [196]. Such colonization may lead to infection inside the uterus. Eighty-five percent of the oral microbiota is introduced in the initial six months after birth in the newborn–early childhood period, and children’s faces resemble their mothers [196–198].

### 9.2. Gut Microbiome for ASD

The most well-researched aspect of autism is the intestinal flora. Bacteria in a higher organism are found in the gastrointestinal tract, and they play major physiological roles in metabolic activity, digestive health, immune function, and endocrine and neurological function. Any imbalance in the relationship between the intestinal bacteria and the different human cells might result in sickness. Microbes impact the host’s vital biological processes and may be a significant factor in the etiology of many diseases [178,179,199,200]. The importance of gut bacteria in public health has prompted studies to emphasize the significance of identifying these microbes as potential contributors to the development of ASD. *Prevotella*, a promising health biomarker, more abundant in neurotypical people but almost non-existent in autistic people, is another notable Gram-negative bacteria genus [201–203]. *Prevotella* is abundant in people who eat a diet high in phytonutrients, complex carbs, and the oil obtained from fish, and is essential for normal brain development. It metabolizes energy to develop vitamin B1, which helps with the signs of autism [204]. In controlled and prospective clinical studies, two forms of vitamin B12 have been investigated: (1) There is evidence that subcutaneously administered mB12 improves methylation and the clinical symptoms of ASD. Redox metabolism also seems to be linked to improvements in clinical symptoms, biochemistry, and physical medical diseases, particularly in people with unfavorable biochemistry and when paired with folinic acid (also known as leucovorin). (2) A combination of cB12 and mB12 contained in an MVI. *Prevotella* deficiency suggests distinct nutritional autism-related habits in children that alter gut microorganism combinations and influence neurodevelopment, implying that restoring it could be therapeutic [205]. Among the Firmicutes phyla is the Gram-positive genus *Clostridium*,

which is more prevalent in ASD patients. The associated species are *Clostridium botulinum* and *Clostridium histolyticum*. These Gram-positive bacteria produce enterotoxins, which cause diarrhea by damaging intestinal tissue. They may also contribute to the increased cellular uptake of heavy substances like protein obtained from cereals and milk [206]. Moreover, the advantageous bacterial population of a *Bifidobacterium* was discovered in lower numbers among those with autism. Intestinal dysregulation in people with ASD has been confirmed through a rise in potentially pathogenic microbes and a reduction in beneficial bacteria [205,206]. In a different light, ASD patients' apparent lack of gut microbe variety and prosperity increases their susceptibility to a vulnerable gut environment, resulting in GI disturbances, infections, and autistic behaviors [204]. Finally, a symbolic change in the gut's microbial composition disrupts critical physiological processes, which affects the behavioral manifestations of autism and further leads to an utter lack of favorable microbiological byproducts, the release of toxic microbiological lipopolysaccharides, and the pathogenic incursion of the gut lining by immunological cytokines that promote neuroimmune inflammation [205,206].

### 9.3. Vaginal Microbiome for ASD

The mother's characteristics that influence the growth of autism are referred to as vaginal microbiota. Because humans are born germ-free, it is assumed that the first colonization of bacteria in the human stomach occurs during birth and travels through the vaginal opening. However, evidence is emerging that colonization of the uterus may begin much earlier [207–213]. Contamination happens during a cesarean delivery when the infant touches the maternal epidermis and other pathogens. The microbial population that is transferred via vertical transmission impacts the microbial community in the intestine [214–216]. As a result, the substantial disruption to the microorganisms in the mother's genitalia due to high metabolic demand unintentionally impairs neurodevelopment in infants at a critical stage in brain development [217]. The vaginal canal contains more than fifty bacterial species, with *Lactobacillus* being the most common in healthy women [218]. It has been shown that maternal anxiety during the first trimester seems to have an inhibitory effect on genital immunology and the proportion of bacteria responsible for the production of lactose, which causes an imbalance of microbes in the gut. The imbalance of gut microbes is transmitted vertically to children. Vaginosis caused by bacteria is a common ailment. According to research, premature toddlers have a 10-times higher risk of developing ASD than those born at full term. According to research trials and extensive epidemiological studies, many maternal prenatal infectious diseases and the increased signaling molecules produced by immune cells increase the chance of autism in children [219–222].

## 10. Conclusions

ASD has multiple causes, including epigenetic changes and complex genetic mutations, and due to these complications, it is hard to manage ASD. Currently, to manage ASD symptoms, risperidone and aripiprazole are the only two drugs approved by the FDA. There is ambiguity around this illness because there are now just a few alternatives for therapeutic intervention. In this review, we shed light on the non-pharmacological interventions that can help to manage the symptoms of ASD. This review includes all the nutritional approaches, including dietary phytochemicals like curcumin, resveratrol, naringenin, sulforaphane, camel milk, herbal medicines, different hormones, and microbiomes, for managing ASD. These are possible safe and efficient approaches to lessening the burden of the disease and are important because healthcare is expensive, and there is a significant burden on ASD carers. There is an urgent need for multidisciplinary research focusing on drug delivery techniques in the brain for the above-mentioned dietary supplement and microbiome therapy. All non-pharmacological interventions need rigorous clinical trials to bring them to the market, but a few of them can be consumed in their current state with FDA-approved drugs to manage ASD symptoms.

**Author Contributions:** R.S.—writing: reviewing and editing, and figures; F.A.—writing: reviewing and editing, and tables; U.A.—writing: reviewing and editing; A.M.—writing: reviewing and editing; A.K.—writing: reviewing and editing; R.L.J.—writing: reviewing and editing; V.S.—writing: reviewing and editing. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no competing interest.

## Abbreviations

ASD	Autism spectrum disorder
CAM	Complementary and alternative medicine
ROS	Reactive oxygen species
PDDs	Pervasive developmental disorders
AAs	Amino acids
AS	Asperger’s syndrome
NMDA	N-methyl D-aspartate
NO	Nitric oxide
BBB	blood–brain barrier
MDA	Malondialdehyde
DHA	Docosahexaenoic acid
ALA	Alpha-linolenic acid
EPA	Eicosapentaenoic acid
NF- $\kappa$ B	Nuclear factor kappa B
PI3K	Phosphatidylinositol 3-Kinase
RCTs	Randomized controlled trials
BDNF	Brain-derived neurotrophic factor
COX	Cyclooxygenase
PPARs	Peroxisome proliferator-activated receptors
eNOS	Endothelial nitric oxide synthase
SIRT1	Sirtuin (silent mating type information regulation homolog 1)
AMPK	AMP-activated protein kinase
RSV	Resveratrol
VPA	Valproic acid
PVC	Parvalbumin
PPA	Propionic acid
MMP-9	Matrix metalloproteinase-9
SD	Schizotypal disorder
NAR	Naringenin
VLDL	Very low-density lipoproteins
MCAO	Middle cerebral artery occlusion
SOCS3	Suppressor of cytokine signaling 3
NRF2	Nuclear factor erythroid 2–related factor 2
CGA	Clinical Global Assessment
ABC	Autism Behavior Checklist
SRS	Social Responsiveness Scale
GSH-Px	Glutathione peroxidase
SOD	Superoxide dismutase
MPO	Myeloperoxidase
ETC	Electron transport chain
GI	Gastrointestinal
RDBPC	Randomized Double Blind Placebo Control
PedPRM	Pediatric-appropriate prolonged-release melatonin
RRBs	Restricted and repetitive behaviors



IMFAR	International Meeting for Autism Research
SRS	Social Responsiveness Scale
CBD	Cannabidiol
THC	Tetrahydrocannabinol
GABAA	Gamma-Aminobutyric Acid-A
AEA	N-arachidonylethanolamine (anandamide)
ECS	Endocannabinoid system
CNS	Central nervous system
DAGL	Diacylglycerol lipase
2-AG	2-arachidonoylglycerol
FXS	Fragile X Syndrome
CBDV	Cannabidavarin
ADAMS	Anxiety, Depression, and Mood Scale
FDA	Food and Drug Administration
SCFAs	Short-chain fatty acids
Mg	Magnesium
PD	Parkinson's disease

## References

1. Kanner, L. Autistic disturbances of affective contact. *Nerv. Child* **1943**, *2*, 217–250.
2. Asperger, H. Die „Autistischen psychopathen“ im Kindesalter. *Arch. Psychiatr. Nervenkrankh.* **1944**, *117*, 76–136. [[CrossRef](#)]
3. Elsabbagh, M.; Divan, G.; Koh, Y.J.; Kim, Y.S.; Kauchali, S.; Marcín, C.; Montiel-Nava, C.; Patel, V.; Paula, C.S.; Wang, C.; et al. Global prevalence of autism and other pervasive developmental disorders. *Autism Res.* **2012**, *5*, 160–179.
4. Wiśniowiecka-Kowalnik, B.; Nowakowska, B.A. Genetics and epigenetics of autism spectrum disorder—Current evidence in the field. *J. Appl. Genet.* **2019**, *60*, 37–47. [[PubMed](#)]
5. Ahmad, F.; Virmani, A.; Irfan, M.; Rankawat, S.; Pathak, U. Critical appraisals on depressions and psychotic symptoms. *J. Neurobehav. Sci.* **2021**, *8*, 81–88.
6. Baio, J.; Wiggins, L.; Christensen, D.L.; Maenner, M.J.; Daniels, J.; Warren, Z.; Kurzius-Spencer, M.; Zahorodny, W.; Rosenberg, C.R.; White, T.; et al. Prevalence of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2014. *MMWR Surveill. Summ.* **2018**, *67*, 1.
7. Lai, M.C.; Lombardo, M.V.; Baron-Cohen, S. Autism. *Lancet* **2014**, *383*, 896–910.
8. Williams, J.G.; Higgins, J.P.; Brayne, C.E. Systematic review of prevalence studies of autism spectrum disorders. *Arch. Dis. Child.* **2006**, *91*, 8–15. [[CrossRef](#)]
9. Fombonne, E. Epidemiology of autistic disorder and other pervasive developmental disorders. *J. Clin. Psychiatry* **2005**, *66*, 3.
10. Shattuck, P.T. The contribution of diagnostic substitution to the growing administrative prevalence of autism in US special education. *Pediatrics* **2006**, *117*, 1028–1037.
11. Williams, J.; Allison, C.; Scott, F.; Stott, C.; Bolton, P.; Baron-Cohen, S.; Brayne, C. The Childhood Asperger Syndrome Test (CAST): Test-retest reliability. *Autism Int. J. Res. Pract.* **2006**, *10*, 415–427.
12. Wing, L.; Potter, D. The epidemiology of autistic spectrum disorders: Is the prevalence rising? *Ment. Retard. Dev. Disabil. Res. Rev.* **2002**, *8*, 151–161. [[PubMed](#)]
13. Bishop, D.V.; Whitehouse, A.J.; Watt, H.J.; Line, E.A. Autism and diagnostic substitution: Evidence from a study of adults with a history of developmental language disorder. *Dev. Med. Child Neurol.* **2008**, *50*, 341–345. [[PubMed](#)]
14. Croen, L.A.; Grether, J.K.; Hoogstrate, J.; Selvin, S. The changing prevalence of autism in California. *J. Autism Dev. Disord.* **2002**, *32*, 207–215. [[PubMed](#)]
15. Barbaresi, W.J.; Katusic, S.K.; Colligan, R.C.; Weaver, A.L.; Jacobsen, S.J. The incidence of autism in Olmsted County, Minnesota, 1976–1997: Results from a population-based study. *Arch. Pediatr. Adolesc. Med.* **2005**, *159*, 37–44.
16. Hertz-Picciotto, I.; Delwiche, L. The rise in autism and the role of age at diagnosis. *Epidemiology* **2009**, *20*, 84.
17. Mandell, D.S.; Palmer, R. Differences among states in the identification of autistic spectrum disorders. *Arch. Pediatr. Adolesc. Med.* **2005**, *159*, 266–269.
18. Parner, E.T.; Schendel, D.E.; Thorsen, P. Autism prevalence trends over time in Denmark: Changes in prevalence and age at diagnosis. *Arch. Pediatr. Adolesc. Med.* **2008**, *162*, 1150–1156.
19. Loomes, R.; Hull, L.; Mandy, W. What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis. *J. Am. Acad. Child Adolesc. Psychiatry* **2017**, *56*, 466–474.
20. Asherson, P.J.; Curran, S. Approaches to gene mapping in complex disorders and their application in child psychiatry and psychology. *Br. J. Psychiatry* **2001**, *179*, 122–128.
21. Zbiciak, A.; Markiewicz, T. A new extraordinary means of appeal in the Polish criminal procedure: The basic principles of a fair trial and a complaint against a cassatory judgment. *Access Justice East. Eur.* **2023**, *6*, 1–18. [[CrossRef](#)]
22. Fombonne, E.; Zakarian, R.; Bennett, A.; Meng, L.; McLean-Heywood, D. Pervasive developmental disorders in Montreal, Quebec, Canada: Prevalence and links with immunizations. *Pediatrics* **2006**, *118*, e139–e150. [[CrossRef](#)]

23. Jorde, L.B.; Hasstedt, S.J.; Ritvo, E.R.; Mason-Brothers, A.; Freeman, B.J.; Pingree, C.; McMahon, W.M.; Petersen, B.; Jenson, W.R.; Mo, A. Complex segregation analysis of autism. *Am. J. Hum. Genet.* **1991**, *49*, 932–938.
24. Lauritsen, M.B.; Pedersen, C.B.; Mortensen, P.B. Effects of familial risk factors and place of birth on the risk of autism: A nationwide register-based study. *J. Child Psychol. Psychiatry Allied Discip.* **2005**, *46*, 963–971. [[CrossRef](#)] [[PubMed](#)]
25. Muhle, R.; Trentacoste, S.V.; Rapin, I. The genetics of autism. *Pediatrics* **2004**, *113*, e472–e486. [[CrossRef](#)] [[PubMed](#)]
26. Ozonoff, S.; Young, G.S.; Carter, A.; Messinger, D.; Yirmiya, N.; Zwaigenbaum, L.; Bryson, S.; Carver, L.J.; Constantino, J.N.; Dobkins, K.; et al. Recurrence risk for autism spectrum disorders: A Baby Siblings Research Consortium study. *Pediatrics* **2011**, *128*, e488–e495. [[CrossRef](#)]
27. Piven, J.; Gayle, J.; Chase, G.A.; Fink, B.; Landa, R.; Wzorek, M.M.; Folstein, S.E. A family history study of neuropsychiatric disorders in the adult siblings of autistic individuals. *J. Am. Acad. Child Adolesc. Psychiatry* **1990**, *29*, 177–183. [[CrossRef](#)]
28. Risch, N.; Spiker, D.; Lotspeich, L.; Nouri, N.; Hinds, D.; Hallmayer, J.; Kalaydjieva, L.; McCague, P.; Dimiceli, S.; Pitts, T.; et al. A genomic screen of autism: Evidence for a multilocus etiology. *Am. J. Hum. Genet.* **1999**, *65*, 493–507. [[CrossRef](#)] [[PubMed](#)]
29. Schaefer, G.B.; Mendelsohn, N.J.; Professional Practice and Guidelines Committee. Clinical genetics evaluation in identifying the etiology of autism spectrum disorders: 2013 guideline revisions. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **2013**, *15*, 399–407. [[CrossRef](#)] [[PubMed](#)]
30. Constantino, J.N.; Zhang, Y.; Frazier, T.; Abbacchi, A.M.; Law, P. Sibling recurrence and the genetic epidemiology of autism. *Am. J. Psychiatry* **2010**, *167*, 1349–1356. [[CrossRef](#)]
31. Palmer, N.; Beam, A.; Agniel, D.; Eran, A.; Manrai, A.; Spettell, C.; Steinberg, G.; Mandl, K.; Fox, K.; Nelson, S.F.; et al. Association of Sex with Recurrence of Autism Spectrum Disorder Among Siblings. *JAMA Pediatr.* **2017**, *171*, 1107–1112. [[CrossRef](#)] [[PubMed](#)]
32. Sachdeva, P.; Mehdi, I.; Kaith, R.; Ahmad, F.; Anwar, M.S. Potential natural products for the management of autism spectrum disorder. *Ibrain* **2022**, *8*, 365–376. [[CrossRef](#)]
33. Dalton, K.M.; Nacewicz, B.M.; Alexander, A.L.; Davidson, R.J. Gaze-fixation, brain activation, and amygdala volume in unaffected siblings of individuals with autism. *Biol. Psychiatry* **2007**, *61*, 512–520. [[CrossRef](#)]
34. Gamliel, I.; Yirmiya, N.; Jaffe, D.H.; Manor, O.; Sigman, M. Developmental trajectories in siblings of children with autism: Cognition and language from 4 months to 7 years. *J. Autism Dev. Disord.* **2009**, *39*, 1131–1144. [[CrossRef](#)]
35. Gamliel, I.; Yirmiya, N.; Sigman, M. The development of young siblings of children with autism from 4 to 54 months. *J. Autism Dev. Disord.* **2007**, *37*, 171–183. [[CrossRef](#)]
36. Piven, J.; Palmer, P.; Jacobi, D.; Childress, D.; Arndt, S. Broader autism phenotype: Evidence from a family history study of multiple-incidence autism families. *Am. J. Psychiatry* **1997**, *154*, 185–190.
37. Yirmiya, N.; Gamliel, I.; Shaked, M.; Sigman, M. Cognitive and verbal abilities of 24- to 36-month-old siblings of children with autism. *J. Autism Dev. Disord.* **2007**, *37*, 218–229. [[CrossRef](#)]
38. Baron-Cohen, S. Two new theories of autism: Hyper-systemising and assortative mating. *Arch. Dis. Child.* **2006**, *91*, 2–5. [[CrossRef](#)]
39. Ecker, C.; Bookheimer, S.Y.; Murphy, D.G. Neuroimaging in autism spectrum disorder: Brain structure and function across the lifespan. *The Lancet. Neurology* **2015**, *14*, 1121–1134. [[CrossRef](#)]
40. Muhle, R.A.; Reed, H.E.; Stratigos, K.A.; Veenstra-VanderWeele, J. The Emerging Clinical Neuroscience of Autism Spectrum Disorder: A Review. *JAMA Psychiatry* **2018**, *75*, 514–523. [[CrossRef](#)]
41. Lopez-Rangel, E.; Lewis, M.E. Loud and clear evidence for gene silencing by epigenetic mechanisms in autism spectrum and related neurodevelopmental disorders. *Clin. Genet.* **2006**, *69*, 21–22. [[CrossRef](#)] [[PubMed](#)]
42. Samaco, R.C.; Nagarajan, R.P.; Braunschweig, D.; LaSalle, J.M. Multiple pathways regulate MeCP2 expression in normal brain development and exhibit defects in autism-spectrum disorders. *Hum. Mol. Genet.* **2004**, *13*, 629–639. [[CrossRef](#)] [[PubMed](#)]
43. Christison, G.W.; Ivany, K. Elimination diets in autism spectrum disorders: Any wheat amidst the chaff? *J. Dev. Behav. Pediatr. JDBP* **2006**, *27* (Suppl. 2), S162–S171. [[CrossRef](#)] [[PubMed](#)]
44. Buie, T. The relationship of autism and gluten. *Clin. Ther.* **2013**, *35*, 578–583. [[CrossRef](#)] [[PubMed](#)]
45. Jyonouchi, H.; Sun, S.; Itokazu, N. Innate immunity associated with inflammatory responses and cytokine production against common dietary proteins in patients with autism spectrum disorder. *Neuropsychobiology* **2002**, *46*, 76–84. [[CrossRef](#)] [[PubMed](#)]
46. Lau, N.M.; Green, P.H.; Taylor, A.K.; Hellberg, D.; Ajamian, M.; Tan, C.Z.; Kosofsky, B.E.; Higgins, J.J.; Rajadhyaksha, A.M.; Alaedini, A. Markers of Celiac Disease and Gluten Sensitivity in Children with Autism. *PLoS ONE* **2013**, *8*, e66155. [[CrossRef](#)]
47. Quan, L.; Xu, X.; Cui, Y.; Han, H.; Hendren, R.L.; Zhao, L.; You, X. A systematic review and meta-analysis of the benefits of a gluten-free diet and/or casein-free diet for children with autism spectrum disorder. *Nutr. Rev.* **2022**, *80*, 1237–1246. [[CrossRef](#)]
48. Mari-Bauset, S.; Zazpe, I.; Mari-Sanchis, A.; Llopis-González, A.; Morales-Suárez-Varela, M. Evidence of the gluten-free and casein-free diet in autism spectrum disorders: A systematic review. *J. Child Neurol.* **2014**, *29*, 1718–1727. [[CrossRef](#)]
49. Geraghty, M.E.; Bates-Wall, J.; Ratliff-Schaub, K.; Lane, A.E. Nutritional interventions and therapies in autism: A spectrum of what we know: Part 2. *ICAN Infant Child Adolesc. Nutr.* **2010**, *2*, 120–133. [[CrossRef](#)]
50. Horvath, K.; Perman, J.A. Autism and gastrointestinal symptoms. *Curr. Gastroenterol. Rep.* **2002**, *4*, 251–258. [[CrossRef](#)]
51. White, J.F. Intestinal pathophysiology in autism. *Exp. Biol. Med.* **2003**, *228*, 639–649. [[CrossRef](#)] [[PubMed](#)]
52. Mezzelani, A.; Landini, M.; Facchiano, F.; Raggi, M.E.; Villa, L.; Molteni, M.; De Santis, B.; Brera, C.; Caroli, A.M.; Milanese, L.; et al. Environment, dysbiosis, immunity and sex-specific susceptibility: A translational hypothesis for regressive autism pathogenesis. *Nutr. Neurosci.* **2015**, *18*, 145–161. [[CrossRef](#)] [[PubMed](#)]



53. Kawicka, A.; Regulska-Ilow, B. How nutritional status, diet and dietary supplements can affect autism. A review. *Rocz. Państwowego Zakładu Hig.* **2013**, *64*, 1–12.
54. Gottschall, E. Digestion-gut-autism connection: The specific carbohydrate diet. *Med. Veritas* **2004**, *1*, 261–271. [\[CrossRef\]](#)
55. Żarnowska, I.; Chrapko, B.; Gwizda, G.; Nocuń, A.; Mitosek-Szewczyk, K.; Gasior, M. Therapeutic use of carbohydrate-restricted diets in an autistic child; a case report of clinical and 18FDG PET findings. *Metab. Brain Dis.* **2018**, *33*, 1187–1192. [\[CrossRef\]](#) [\[PubMed\]](#)
56. Ruskin, D.N.; Svedova, J.; Cote, J.L.; Sandau, U.; Rho, J.M.; Kawamura, M., Jr.; Boison, D.; Masino, S.A. Ketogenic diet improves core symptoms of autism in BTBR mice. *PLoS ONE* **2013**, *8*, e65021. [\[CrossRef\]](#)
57. Napoli, E.; Dueñas, N.; Giulivi, C. Potential therapeutic use of the ketogenic diet in autism spectrum disorders. *Front. Pediatr.* **2014**, *2*, 69. [\[CrossRef\]](#)
58. Ruskin, D.N.; Murphy, M.I.; Slade, S.L.; Masino, S.A. Ketogenic diet improves behaviors in a maternal immune activation model of autism spectrum disorder. *PLoS ONE* **2017**, *12*, e0171643. [\[CrossRef\]](#)
59. Dai, Y.; Zhao, Y.; Tomi, M.; Shin, B.C.; Thamotharan, S.; Mazarati, A.; Sankar, R.; Wang, E.A.; Cepeda, C.; Levine, M.S.; et al. Sex-Specific Life Course Changes in the Neuro-Metabolic Phenotype of Glut3 Null Heterozygous Mice: Ketogenic Diet Ameliorates Electroencephalographic Seizures and Improves Sociability. *Endocrinology* **2017**, *158*, 936–949. [\[CrossRef\]](#)
60. Kasprowska-Liśkiewicz, D.; Liśkiewicz, A.D.; Nowacka-Chmielewska, M.M.; Nowicka, J.; Małecki, A.; Barski, J.J. The ketogenic diet affects the social behavior of young male rats. *Physiol. Behav.* **2017**, *179*, 168–177. [\[CrossRef\]](#)
61. El-Rashidy, O.; El-Baz, F.; El-Gendy, Y.; Khalaf, R.; Reda, D.; Saad, K. Ketogenic diet versus gluten free casein free diet in autistic children: A case-control study. *Metab. Brain Dis.* **2017**, *32*, 1935–1941. [\[CrossRef\]](#) [\[PubMed\]](#)
62. Hardy, T.M.; Tollefsbol, T.O. Epigenetic diet: Impact on the epigenome and cancer. *Epigenomics* **2011**, *3*, 503–518. [\[CrossRef\]](#) [\[PubMed\]](#)
63. Meeran, S.M.; Ahmed, A.; Tollefsbol, T.O. Epigenetic targets of bioactive dietary components for cancer prevention and therapy. *Clin. Epigenet.* **2010**, *1*, 101–116. [\[CrossRef\]](#) [\[PubMed\]](#)
64. Rollett, A. *Zur Kenntnis der Linolensäure und des Leinöls*; De Gruyter: Berlin, Germany, 1909.
65. Cheng, Y.S.; Tseng, P.T.; Chen, Y.W.; Stubbs, B.; Yang, W.C.; Chen, T.Y.; Wu, C.K.; Lin, P.Y. Supplementation of omega 3 fatty acids may improve hyperactivity, lethargy, and stereotypy in children with autism spectrum disorders: A meta-analysis of randomized controlled trials. *Neuropsychiatr. Dis. Treat.* **2017**, *13*, 2531–2543. [\[CrossRef\]](#) [\[PubMed\]](#)
66. Hagmeyer, S.; Sauer, A.K.; Grabrucker, A.M. Prospects of Zinc Supplementation in Autism Spectrum Disorders and Shankopathies Such as Phelan McDermid Syndrome. *Front. Synaptic Neurosci.* **2018**, *10*, 11. [\[CrossRef\]](#) [\[PubMed\]](#)
67. Parikh, S.; Saneto, R.; Falk, M.J.; Anselm, I.; Cohen, B.H.; Haas, R.; Medicine Society, T.M. A modern approach to the treatment of mitochondrial disease. *Curr. Treat. Options Neurol.* **2009**, *11*, 414–430. [\[CrossRef\]](#)
68. Bou Khalil, R.; Yazbek, J.C. Potential importance of supplementation with zinc for autism spectrum disorder. *L'Encephale* **2021**, *47*, 514–517. [\[CrossRef\]](#)
69. Gunes, S.; Ekinci, O.; Celik, T. Iron deficiency parameters in autism spectrum disorder: Clinical correlates and associated factors. *Ital. J. Pediatr.* **2017**, *43*, 86. [\[CrossRef\]](#)
70. Prakash, P.; Kumari, R.; Sinha, N.; Kumar, S.; Sinha, P. Evaluation of Iron Status in Children with Autism Spectral Disorder: A Case-control Study. *J. Clin. Diagn. Res.* **2021**, *15*, BC01–BC04. [\[CrossRef\]](#)
71. Mousain-Bosc, M.; Roche, M.; Polge, A.; Pradal-Prat, D.; Rapin, J.; Bali, J.P. Improvement of neurobehavioral disorders in children supplemented with magnesium-vitamin B6. *Magnes. Res.* **2006**, *19*, 46–52.
72. Galland, L. Magnesium, stress and neuropsychiatric disorders. *Magnes. Trace Elem.* **1993**, *10*, 287.
73. Błażewicz, A.; Szymańska, I.; Dolliver, W.; Suchocki, P.; Turlo, J.; Makarewicz, A.; Skórzyńska-Dzidusko, K. Are Obese Patients with Autism Spectrum Disorder More Likely to Be Selenium Deficient? Research Findings on Pre- and Post-Pubertal Children. *Nutrients* **2020**, *12*, 3581. [\[CrossRef\]](#)
74. Ak, T.; Gülçin, I. Antioxidant and radical scavenging properties of curcumin. *Chem.-Biol. Interact.* **2008**, *174*, 27–37. [\[CrossRef\]](#)
75. Cole, G.M.; Teter, B.; Frautschy, S.A. Neuroprotective effects of curcumin. *Adv. Exp. Med. Biol.* **2007**, *595*, 197–212. [\[PubMed\]](#)
76. Al-Askar, M.; Bhat, R.S.; Selim, M.; Al-Ayadhi, L.; El-Ansary, A. Postnatal treatment using curcumin supplements to amend the damage in VPA-induced rodent models of autism. *BMC Complement. Altern. Med.* **2017**, *17*, 259. [\[CrossRef\]](#)
77. Bhandari, R.; Kuhad, A. Neuropsychopharmacotherapeutic efficacy of curcumin in experimental paradigm of autism spectrum disorders. *Life Sci.* **2015**, *141*, 156–169. [\[CrossRef\]](#) [\[PubMed\]](#)
78. Panahi, Y.; Badeli, R.; Karami, G.R.; Sahebkar, A. Investigation of the efficacy of adjunctive therapy with bioavailability-boosted curcuminoids in major depressive disorder. *Phytother. Res. PTR* **2015**, *29*, 17–21. [\[CrossRef\]](#)
79. Panahi, Y.; Saadat, A.; Beiraghdar, F.; Sahebkar, A. Adjuvant therapy with bioavailability-boosted curcuminoids suppresses systemic inflammation and improves quality of life in patients with solid tumors: A randomized double-blind placebo-controlled trial. *Phytother. Res. PTR* **2014**, *28*, 1461–1467. [\[CrossRef\]](#)
80. Dong, S.; Zeng, Q.; Mitchell, E.S.; Xiu, J.; Duan, Y.; Li, C.; Tiwari, J.K.; Hu, Y.; Cao, X.; Zhao, Z. Curcumin enhances neurogenesis and cognition in aged rats: Implications for transcriptional interactions related to growth and synaptic plasticity. *PLoS ONE* **2012**, *7*, e31211. [\[CrossRef\]](#)
81. Tizabi, Y.; Hurley, L.L.; Qualls, Z.; Akinfiresoye, L. Relevance of the anti-inflammatory properties of curcumin in neurodegenerative diseases and depression. *Molecules* **2014**, *19*, 20864–20879. [\[CrossRef\]](#)

82. Motterlini, R.; Foresti, R.; Bassi, R.; Green, C.J. Curcumin, an antioxidant and anti-inflammatory agent, induces heme oxygenase-1 and protects endothelial cells against oxidative stress. *Free Radic. Biol. Med.* **2000**, *28*, 1303–1312. [[CrossRef](#)] [[PubMed](#)]
83. Karlstetter, M.; Lippe, E.; Walczak, Y.; Moehle, C.; Aslanidis, A.; Mirza, M.; Langmann, T. Curcumin is a potent modulator of microglial gene expression and migration. *J. Neuroinflamm.* **2011**, *8*, 125. [[CrossRef](#)]
84. Tegenge, M.A.; Rajbhandari, L.; Shrestha, S.; Mithal, A.; Hosmane, S.; Venkatesan, A. Curcumin protects axons from degeneration in the setting of local neuroinflammation. *Exp. Neurol.* **2014**, *253*, 102–110. [[CrossRef](#)] [[PubMed](#)]
85. Bhandari, R.; Paliwal, J.K.; Kuhad, A. Dietary phytochemicals as neurotherapeutics for autism spectrum disorder: Plausible mechanism and evidence. In *Personalized Food Intervention and Therapy for Autism Spectrum Disorder Management*; Springer: Cham, Switzerland, 2020; pp. 615–646.
86. Bassani, T.B.; Turnes, J.M.; Moura, E.; Bonato, J.M.; Cópola-Segovia, V.; Zanata, S.M.; Oliveira, R.; Vital, M. Effects of curcumin on short-term spatial and recognition memory, adult neurogenesis and neuroinflammation in a streptozotocin-induced rat model of dementia of Alzheimer's type. *Behav. Brain Res.* **2017**, *335*, 41–54. [[CrossRef](#)] [[PubMed](#)]
87. Zhang, L.; Fang, Y.; Xu, Y.; Lian, Y.; Xie, N.; Wu, T.; Zhang, H.; Sun, L.; Zhang, R.; Wang, Z. Curcumin Improves Amyloid  $\beta$ -Peptide (1–42) Induced Spatial Memory Deficits through BDNF-ERK Signaling Pathway. *PLoS ONE* **2015**, *10*, e0131525. [[CrossRef](#)]
88. Lee, W.H.; Loo, C.Y.; Bebawy, M.; Luk, F.; Mason, R.S.; Rohanizadeh, R. Curcumin and its derivatives: Their application in neuropharmacology and neuroscience in the 21st century. *Curr. Neuropharmacol.* **2013**, *11*, 338–378. [[CrossRef](#)]
89. Aggarwal, B.B.; Gupta, S.C.; Sung, B. Curcumin: An orally bioavailable blocker of TNF and other pro-inflammatory biomarkers. *Br. J. Pharmacol.* **2013**, *169*, 1672–1692. [[CrossRef](#)]
90. Jacob, A.; Wu, R.; Zhou, M.; Wang, P. Mechanism of the Anti-inflammatory Effect of Curcumin: PPAR-gamma Activation. *PPAR Res.* **2007**, *2007*, 89369. [[CrossRef](#)]
91. Saja, K.; Babu, M.S.; Karunakaran, D.; Sudhakaran, P.R. Anti-inflammatory effect of curcumin involves downregulation of MMP-9 in blood mononuclear cells. *Int. Immunopharmacol.* **2007**, *7*, 1659–1667. [[CrossRef](#)]
92. Frémont, L. Biological effects of resveratrol. *Life Sci.* **2000**, *66*, 663–673. [[CrossRef](#)]
93. Wendeburg, L.; de Oliveira, A.C.; Bhatia, H.S.; Candelario-Jalil, E.; Fiebich, B.L. Resveratrol inhibits prostaglandin formation in IL-1 $\beta$ -stimulated SK-N-SH neuronal cells. *J. Neuroinflamm.* **2009**, *6*, 26. [[CrossRef](#)] [[PubMed](#)]
94. Fullerton, M.D.; Steinberg, G.R. SIRT1 takes a backseat to AMPK in the regulation of insulin sensitivity by resveratrol. *Diabetes* **2010**, *59*, 551–553. [[CrossRef](#)] [[PubMed](#)]
95. Wong, Y.T.; Gruber, J.; Jenner, A.M.; Ng, M.P.; Ruan, R.; Tay, F.E. Elevation of oxidative-damage biomarkers during aging in F2 hybrid mice: Protection by chronic oral intake of resveratrol. *Free Radic. Biol. Med.* **2009**, *46*, 799–809. [[CrossRef](#)]
96. Fontes-Dutra, M.; Santos-Terra, J.; Deckmann, I.; Brum Schwingel, G.; Della-Flora Nunes, G.; Hirsch, M.M.; Bauer-Negrini, G.; Riesgo, R.S.; Bambini-Junior, V.; Hedin-Pereira, C.; et al. Resveratrol Prevents Cellular and Behavioral Sensory Alterations in the Animal Model of Autism Induced by Valproic Acid. *Front. Synaptic Neurosci.* **2018**, *10*, 9. [[CrossRef](#)]
97. Kim, Y.A.; Kim, G.Y.; Park, K.Y.; Choi, Y.H. Resveratrol inhibits nitric oxide and prostaglandin E2 production by lipopolysaccharide-activated C6 microglia. *J. Med. Food* **2007**, *10*, 218–224. [[CrossRef](#)]
98. Bambini-Junior, V.; Zanatta, G.; Della Flora Nunes, G.; Mueller de Melo, G.; Michels, M.; Fontes-Dutra, M.; Nogueira Freire, V.; Riesgo, R.; Gottfried, C. Resveratrol prevents social deficits in animal model of autism induced by valproic acid. *Neurosci. Lett.* **2014**, *583*, 176–181. [[CrossRef](#)]
99. Roulet, F.I.; Lai, J.K.; Foster, J.A. In utero exposure to valproic acid and autism--a current review of clinical and animal studies. *Neurotoxicol. Teratol.* **2013**, *36*, 47–56. [[CrossRef](#)]
100. Moussa, C.; Hebron, M.; Huang, X.; Ahn, J.; Rissman, R.A.; Aisen, P.S.; Turner, R.S. Resveratrol regulates neuro-inflammation and induces adaptive immunity in Alzheimer's disease. *J. Neuroinflamm.* **2017**, *14*, 1. [[CrossRef](#)]
101. McCalley, A.E.; Kaja, S.; Payne, A.J.; Koulen, P. Resveratrol and calcium signaling: Molecular mechanisms and clinical relevance. *Molecules* **2014**, *19*, 7327–7340. [[CrossRef](#)] [[PubMed](#)]
102. Cheng, G.; Zhang, X.; Gao, D.; Jiang, X.; Dong, W. Resveratrol inhibits MMP-9 expression by up-regulating PPAR alpha expression in an oxygen glucose deprivation-exposed neuron model. *Neurosci. Lett.* **2009**, *451*, 105–108. [[CrossRef](#)] [[PubMed](#)]
103. Gao, D.; Zhang, X.; Jiang, X.; Peng, Y.; Huang, W.; Cheng, G.; Song, L. Resveratrol reduces the elevated level of MMP-9 induced by cerebral ischemia-reperfusion in mice. *Life Sci.* **2006**, *78*, 2564–2570. [[CrossRef](#)]
104. Vallverdú-Queralt, A.; Odriozola-Serrano, I.; Oms-Oliu, G.; Lamuela-Raventós, R.M.; Elez-Martínez, P.; Martín-Belloso, O. Changes in the polyphenol profile of tomato juices processed by pulsed electric fields. *J. Agric. Food Chem.* **2012**, *60*, 9667–9672. [[CrossRef](#)] [[PubMed](#)]
105. Birt, D.F.; Hendrich, S.; Wang, W. Dietary agents in cancer prevention: Flavonoids and isoflavonoids. *Pharmacol. Ther.* **2001**, *90*, 157–177. [[CrossRef](#)] [[PubMed](#)]
106. Fuhr, U.; Klittich, K.; Staib, A.H. Inhibitory effect of grapefruit juice and its bitter principal, naringenin, on CYP1A2 dependent metabolism of caffeine in man. *Br. J. Clin. Pharmacol.* **1993**, *35*, 431–436. [[CrossRef](#)]
107. Raza, S.S.; Khan, M.M.; Ahmad, A.; Ashafaq, M.; Islam, F.; Wagner, A.P.; Safhi, M.M.; Islam, F. Neuroprotective effect of naringenin is mediated through suppression of NF- $\kappa$ B signaling pathway in experimental stroke. *Neuroscience* **2013**, *230*, 157–171. [[CrossRef](#)] [[PubMed](#)]
108. Yi, L.T.; Liu, B.B.; Li, J.; Luo, L.; Liu, Q.; Geng, D.; Tang, Y.; Xia, Y.; Wu, D. BDNF signaling is necessary for the antidepressant-like effect of naringenin. *Prog. Neuro-Psychopharmacol. Biol. Psychiatry* **2014**, *48*, 135–141. [[CrossRef](#)]

109. Galluzzo, P.; Ascenzi, P.; Bulzomi, P.; Marino, M. The nutritional flavanone naringenin triggers antiestrogenic effects by regulating estrogen receptor alpha-palmitoylation. *Endocrinology* **2008**, *149*, 2567–2575. [\[CrossRef\]](#) [\[PubMed\]](#)
110. Nahmias, Y.; Goldwasser, J.; Casali, M.; van Poll, D.; Wakita, T.; Chung, R.T.; Yarmush, M.L. Apolipoprotein B-dependent hepatitis C virus secretion is inhibited by the grapefruit flavonoid naringenin. *Hepatology* **2008**, *47*, 1437–1445. [\[CrossRef\]](#) [\[PubMed\]](#)
111. Wu, L.H.; Lin, C.; Lin, H.Y.; Liu, Y.S.; Wu, C.Y.; Tsai, C.F.; Chang, P.C.; Yeh, W.L.; Lu, D.Y. Naringenin Suppresses Neuroinflammatory Responses Through Inducing Suppressor of Cytokine Signaling 3 Expression. *Mol. Neurobiol.* **2016**, *53*, 1080–1091. [\[CrossRef\]](#)
112. Bhandari, R.; Paliwal, J.K.; Kuhad, A. Naringenin and its nanocarriers as potential phytotherapy for autism spectrum disorders. *J. Funct. Foods* **2018**, *47*, 361–375. [\[CrossRef\]](#)
113. Felgines, C.; Texier, O.; Morand, C.; Manach, C.; Scalbert, A.; Régerat, F.; Rémésy, C. Bioavailability of the flavanone naringenin and its glycosides in rats. *Am. J. Physiol. Gastrointest. Liver Physiol.* **2000**, *279*, G1148–G1154. [\[CrossRef\]](#) [\[PubMed\]](#)
114. Kumar, S.; Tikku, A.B. Biochemical and Molecular Mechanisms of Radioprotective Effects of Naringenin, a Phytochemical from Citrus Fruits. *J. Agric. Food Chem.* **2016**, *64*, 1676–1685. [\[CrossRef\]](#) [\[PubMed\]](#)
115. Guerrero-Beltrán, C.E.; Calderón-Oliver, M.; Pedraza-Chaverri, J.; Chirino, Y.I. Protective effect of sulforaphane against oxidative stress: Recent advances. *Exp. Toxicol. Pathol. Off. J. Ges. Toxikol. Pathol.* **2012**, *64*, 503–508. [\[CrossRef\]](#) [\[PubMed\]](#)
116. Negrette-Guzmán, M.; Huerta-Yepez, S.; Tapia, E.; Pedraza-Chaverri, J. Modulation of mitochondrial functions by the indirect antioxidant sulforaphane: A seemingly contradictory dual role and an integrative hypothesis. *Free Radic. Biol. Med.* **2013**, *65*, 1078–1089. [\[CrossRef\]](#) [\[PubMed\]](#)
117. Zhang, Y.; Kensler, T.W.; Cho, C.G.; Posner, G.H.; Talalay, P. Anticarcinogenic activities of sulforaphane and structurally related synthetic norbornyl isothiocyanates. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 3147–3150. [\[CrossRef\]](#) [\[PubMed\]](#)
118. Wang, G.; Fang, H.; Zhen, Y.; Xu, G.; Tian, J.; Zhang, Y.; Zhang, D.; Zhang, G.; Xu, J.; Zhang, Z.; et al. Sulforaphane Prevents Neuronal Apoptosis and Memory Impairment in Diabetic Rats. *Cell. Physiol. Biochem. Int. J. Exp. Cell. Physiol. Biochem. Pharmacol.* **2016**, *39*, 901–907. [\[CrossRef\]](#) [\[PubMed\]](#)
119. Soane, L.; Li, D.W.; Fiskum, G.; Bambrick, L.L. Sulforaphane protects immature hippocampal neurons against death caused by exposure to hemin or to oxygen and glucose deprivation. *J. Neurosci. Res.* **2010**, *88*, 1355–1363. [\[CrossRef\]](#)
120. Zhao, J.; Kobori, N.; Aronowski, J.; Dash, P.K. Sulforaphane reduces infarct volume following focal cerebral ischemia in rodents. *Neurosci. Lett.* **2006**, *393*, 108–112. [\[CrossRef\]](#)
121. Kraft, A.D.; Johnson, D.A.; Johnson, J.A. Nuclear factor E2-related factor 2-dependent antioxidant response element activation by tert-butylhydroquinone and sulforaphane occurring preferentially in astrocytes conditions neurons against oxidative insult. *J. Neurosci. Off. J. Soc. Neurosci.* **2004**, *24*, 1101–1112. [\[CrossRef\]](#)
122. Gan, N.; Wu, Y.C.; Brunet, M.; Garrido, C.; Chung, F.L.; Dai, C.; Mi, L. Sulforaphane activates heat shock response and enhances proteasome activity through up-regulation of Hsp27. *J. Biol. Chem.* **2010**, *285*, 35528–35536. [\[CrossRef\]](#)
123. Singh, K.; Connors, S.L.; Macklin, E.A.; Smith, K.D.; Fahey, J.W.; Talalay, P.; Zimmerman, A.W. Sulforaphane treatment of autism spectrum disorder (ASD). *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 15550–15555. [\[CrossRef\]](#)
124. Bent, S.; Lawton, B.; Warren, T.; Widjaja, F.; Dang, K.; Fahey, J.W.; Cornblatt, B.; Kinchen, J.M.; Delucchi, K.; Hendren, R.L. Identification of urinary metabolites that correlate with clinical improvements in children with autism treated with sulforaphane from broccoli. *Mol. Autism* **2018**, *9*, 35. [\[CrossRef\]](#) [\[PubMed\]](#)
125. Parker-Athill, E.; Luo, D.; Bailey, A.; Giunta, B.; Tian, J.; Shytle, R.D.; Murphy, T.; Legradi, G.; Tan, J. Flavonoids, a prenatal prophylaxis via targeting JAK2/STAT3 signaling to oppose IL-6/MIA associated autism. *J. Neuroimmunol.* **2009**, *217*, 20–27. [\[CrossRef\]](#) [\[PubMed\]](#)
126. Tassinari, M.; Mottotese, N.; Galvani, G.; Ferrara, D.; Gennaccaro, L.; Loi, M.; Medici, G.; Candini, G.; Rimondini, R.; Ciani, E.; et al. Luteolin Treatment Ameliorates Brain Development and Behavioral Performance in a Mouse Model of CDKL5 Deficiency Disorder. *Int. J. Mol. Sci.* **2022**, *23*, 8719. [\[CrossRef\]](#)
127. Kappeler, S.; Farah, Z.; Puhan, Z. Sequence analysis of Camelus dromedarius milk caseins. *J. Dairy Res.* **1998**, *65*, 209–222. [\[CrossRef\]](#) [\[PubMed\]](#)
128. Bashir, S.; Al-Ayadhi, L.Y. Effect of camel milk on thymus and activation-regulated chemokine in autistic children: Double-blind study. *Pediatr. Res.* **2014**, *75*, 559–563. [\[CrossRef\]](#)
129. Shabo, Y.; Barzel, R.; Margoulis, M.; Yagil, R. Camel milk for food allergies in children. *Isr. Med. Assoc. J. IMAJ* **2005**, *7*, 796–798. [\[PubMed\]](#)
130. Zafra, O.; Fraile, S.; Gutiérrez, C.; Haro, A.; Páez-Espino, A.D.; Jiménez, J.I.; de Lorenzo, V. Monitoring biodegradative enzymes with nanobodies raised in Camelus dromedarius with mixtures of catabolic proteins. *Environ. Microbiol.* **2011**, *13*, 960–974. [\[CrossRef\]](#)
131. Tillib, S.V.; Ivanova, T.I.; Vasilev, L.A. Fingerprint-like Analysis of “Nanoantibody” Selection by Phage Display Using Two Helper Phage Variants. *Acta Naturae* **2010**, *2*, 85–93. [\[CrossRef\]](#)
132. Abdel Galil, M.A.G.; Abdulqader, A.A. The unique medicinal properties of camel products: A review of the scientific evidence. *J. Taibah Univ. Med. Sci.* **2016**, *11*, 98–103. [\[CrossRef\]](#)
133. Al-Ayadhi, L.Y.; Halepoto, D.M.; Al-Dress, A.M.; Mitwali, Y.; Zainah, R. Behavioral Benefits of Camel Milk in Subjects with Autism Spectrum Disorder. *J. Coll. Physicians Surg.-Pak. JCPSP* **2015**, *25*, 819–823. [\[PubMed\]](#)



134. Gagnon, K.; Godbout, R. Melatonin and Comorbidities in Children with Autism Spectrum Disorder. *Curr. Dev. Disord. Rep.* **2018**, *5*, 197–206. [[CrossRef](#)] [[PubMed](#)]
135. Cuomo, B.M.; Vaz, S.; Lee, E.; Thompson, C.; Rogerson, J.M.; Falkmer, T. Effectiveness of Sleep-Based Interventions for Children with Autism Spectrum Disorder: A Meta-Synthesis. *Pharmacotherapy* **2017**, *37*, 555–578. [[CrossRef](#)] [[PubMed](#)]
136. Damiani, J.M.; Sweet, B.V.; Sohoni, P. Melatonin: An option for managing sleep disorders in children with autism spectrum disorder. *Am. J. Health-Syst. Pharm. AJHP Off. J. Am. Soc. Health-Syst. Pharm.* **2014**, *71*, 95–101. [[CrossRef](#)]
137. Rossignol, D.A.; Frye, R.E. Melatonin in autism spectrum disorders: A systematic review and meta-analysis. *Dev. Med. Child Neurol.* **2011**, *53*, 783–792. [[CrossRef](#)]
138. Braam, W.; Ehrhart, F.; Maas, A.; Smits, M.G.; Curfs, L. Low maternal melatonin level increases autism spectrum disorder risk in children. *Res. Dev. Disabil.* **2018**, *82*, 79–89. [[CrossRef](#)]
139. Yamasue, H.; Domes, G. Oxytocin and Autism Spectrum Disorders. *Curr. Top. Behav. Neurosci.* **2018**, *35*, 449–465.
140. Wilczyński, K.M.; Zasada, I.; Siwiec, A.; Janas-Kozik, M. Differences in oxytocin and vasopressin levels in individuals suffering from the autism spectrum disorders vs. general population—A systematic review. *Neuropsychiatr. Dis. Treat.* **2019**, *15*, 2613–2620. [[CrossRef](#)]
141. Voinsky, I.; Bennuri, S.C.; Svigals, J.; Frye, R.E.; Rose, S.; Gurwitz, D. Peripheral Blood Mononuclear Cell Oxytocin and Vasopressin Receptor Expression Positively Correlates with Social and Behavioral Function in Children with Autism. *Sci. Rep.* **2019**, *9*, 13443. [[CrossRef](#)]
142. Hendaus, M.A.; Jomha, F.A.; Alhammadi, A.H. Vasopressin in the Amelioration of Social Functioning in Autism Spectrum Disorder. *J. Clin. Med.* **2019**, *8*, 1061. [[CrossRef](#)]
143. Bang, M.; Lee, S.H.; Cho, S.H.; Yu, S.A.; Kim, K.; Lu, H.Y.; Chang, G.T.; Min, S.Y. Herbal Medicine Treatment for Children with Autism Spectrum Disorder: A Systematic Review. *Evid.-Based Complement. Altern. Med. Ecam* **2017**, *2017*, 8614680. [[CrossRef](#)] [[PubMed](#)]
144. Rezapour, S.; Bahmani, M.; Afsordeh, O.; Rafieian, R.; Sheikhan, A. Herbal medicines: A new hope for autism therapy. *J. Herbmед Pharmacol.* **2016**, *5*, 89–91.
145. Patricio, F.; Morales-Andrade, A.A.; Patricio-Martínez, A.; Limón, I.D. Cannabidiol as a Therapeutic Target: Evidence of its Neuroprotective and Neuromodulatory Function in Parkinson's Disease. *Front. Pharmacol.* **2020**, *11*, 595635. [[CrossRef](#)] [[PubMed](#)]
146. Aran, A.; Harel, M.; Cassuto, H.; Polyansky, L.; Schnapp, A.; Wattad, N.; Shmueli, D.; Golan, D.; Castellanos, F.X. Cannabinoid treatment for autism: A proof-of-concept randomized trial. *Mol. Autism* **2021**, *12*, 6. [[CrossRef](#)] [[PubMed](#)]
147. Leweke, F.M.; Piomelli, D.; Pahlisch, F.; Muhl, D.; Gerth, C.W.; Hoyer, C.; Klosterkötter, J.; Hellmich, M.; Koethe, D. Cannabidiol enhances anandamide signaling and alleviates psychotic symptoms of schizophrenia. *Transl. Psychiatry* **2012**, *2*, e94. [[CrossRef](#)]
148. Russo, E.B.; Burnett, A.; Hall, B.; Parker, K.K. Agonistic properties of cannabidiol at 5-HT<sub>1a</sub> receptors. *Neurochem. Res.* **2005**, *30*, 1037–1043. [[CrossRef](#)]
149. Aishworiya, R.; Valica, T.; Hagerman, R.; Restrepo, B. An Update on Psychopharmacological Treatment of Autism Spectrum Disorder. *Neurother. J. Am. Soc. Exp. Neurother.* **2022**, *19*, 248–262. [[CrossRef](#)] [[PubMed](#)]
150. Tsilioni, I.; Taliou, A.; Francis, K.; Theoharides, T.C. Children with autism spectrum disorders, who improved with a luteolin-containing dietary formulation, show reduced serum levels of TNF and IL-6. *Transl. Psychiatry* **2015**, *5*, e647. [[CrossRef](#)]
151. Wei, D.; Dinh, D.; Lee, D.; Li, D.; Anguren, A.; Moreno-Sanz, G.; Gall, C.M.; Piomelli, D. Enhancement of Anandamide-Mediated Endocannabinoid Signaling Corrects Autism-Related Social Impairment. *Cannabis Cannabinoid Res.* **2016**, *1*, 81–89. [[CrossRef](#)] [[PubMed](#)]
152. Bakas, T.; van Nieuwenhuijzen, P.S.; Devenish, S.O.; McGregor, I.S.; Arnold, J.C.; Chebib, M. The direct actions of cannabidiol and 2-arachidonoyl glycerol at GABA<sub>A</sub> receptors. *Pharmacol. Res.* **2017**, *119*, 358–370. [[CrossRef](#)]
153. Thiele, E.A.; Marsh, E.D.; French, J.A.; Mazurkiewicz-Beldzinska, M.; Benbadis, S.R.; Joshi, C.; Lyons, P.D.; Taylor, A.; Roberts, C.; Sommerville, K.; et al. Cannabidiol in patients with seizures associated with Lennox-Gastaut syndrome (GWPCARE4): A randomised, double-blind, placebo-controlled phase 3 trial. *Lancet* **2018**, *391*, 1085–1096. [[CrossRef](#)] [[PubMed](#)]
154. Qin, M.; Zeidler, Z.; Moulton, K.; Krych, L.; Xia, Z.; Smith, C.B. Endocannabinoid-mediated improvement on a test of aversive memory in a mouse model of fragile X syndrome. *Behav. Brain Res.* **2015**, *291*, 164–171. [[CrossRef](#)] [[PubMed](#)]
155. Heussler, H.; Duhig, M.; Hurst, T.; O'Neill, C.; Gutterman, D.; Palumbo, J.M.; Sebree, T. Longer Term Tolerability and Efficacy of ZYN002 Cannabidiol Transdermal Gel in Children and Adolescents with Autism Spectrum Disorder (ASD): An Open-Label Phase 2 Study (BRIGHT [ZYN2-CL-030]). *Adolesc. Psychiatry* **2020**, *12*, 24.
156. Taliou, A.; Zintzaras, E.; Lykouras, L.; Francis, K. An open-label pilot study of a formulation containing the anti-inflammatory flavonoid luteolin and its effects on behavior in children with autism spectrum disorders. *Clin. Ther.* **2013**, *35*, 592–602. [[CrossRef](#)] [[PubMed](#)]
157. Aran, A.; Cassuto, H.; Lubotzky, A.; Wattad, N.; Hazan, E. Brief Report: Cannabidiol-Rich Cannabis in Children with Autism Spectrum Disorder and Severe Behavioral Problems-A Retrospective Feasibility Study. *J. Autism Dev. Disord.* **2019**, *49*, 1284–1288. [[CrossRef](#)]
158. Pretzsch, C.M.; Freyberg, J.; Voinescu, B.; Lythgoe, D.; Horder, J.; Mendez, M.A.; Wichers, R.; Ajram, L.; Ivin, G.; Heasman, M.; et al. Effects of cannabidiol on brain excitation and inhibition systems; a randomised placebo-controlled single dose trial during magnetic resonance spectroscopy in adults with and without autism spectrum disorder. *Neuropsychopharmacology* **2019**, *44*, 1398–1405. [[CrossRef](#)]

159. Fleury-Teixeira, P.; Caixeta, F.V.; Ramires da Silva, L.C.; Brasil-Neto, J.P.; Malcher-Lopes, R. Effects of CBD-enriched cannabis sativa extract on autism spectrum disorder symptoms: An observational study of 18 participants undergoing compassionate use. *Front. Neurol.* **2019**, *10*, 1145. [\[CrossRef\]](#)
160. Bar-Lev Schleider, L.; Mechoulam, R.; Saban, N.; Meiri, G.; Novack, V. Real life experience of medical cannabis treatment in autism: Analysis of safety and efficacy. *Sci. Rep.* **2019**, *9*, 200. [\[CrossRef\]](#)
161. Theoharides, T.C.; Asadi, S.; Panagiotidou, S. A case series of a luteolin formulation (NeuroProtek®) in children with autism spectrum disorders. *Int. J. Immunopathol. Pharmacol.* **2012**, *25*, 317–323. [\[CrossRef\]](#)
162. Bertolino, B.; Crupi, R.; Impellizzeri, D.; Bruschetta, G.; Cordaro, M.; Siracusa, R.; Esposito, E.; Cuzzocrea, S. Beneficial effects of co-ultramicrozonized palmitoylethanolamide/luteolin in a mouse model of autism and in a case report of autism. *CNS Neurosci. Ther.* **2017**, *23*, 87–98. [\[CrossRef\]](#)
163. Jang, S.; Kelley, K.W.; Johnson, R.W. Luteolin reduces IL-6 production in microglia by inhibiting JNK phosphorylation and activation of AP-1. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 7534–7539. [\[CrossRef\]](#) [\[PubMed\]](#)
164. Basu, A.; Das, A.S.; Majumder, M.; Mukhopadhyay, R. Antiatherogenic Roles of Dietary Flavonoids Chrysin, Quercetin, and Luteolin. *J. Cardiovasc. Pharmacol.* **2016**, *68*, 89–96. [\[CrossRef\]](#) [\[PubMed\]](#)
165. Bhattacharyya, S.; Wilson, R.; Appiah-Kusi, E.; O'Neill, A.; Brammer, M.; Perez, J.; Murray, R.; Allen, P.; Bossong, M.G.; McGuire, P. Effect of Cannabidiol on Medial Temporal, Midbrain, and Striatal Dysfunction in People at Clinical High Risk of Psychosis: A Randomized Clinical Trial. *JAMA Psychiatry* **2018**, *75*, 1107–1117. [\[CrossRef\]](#) [\[PubMed\]](#)
166. Bhattacharyya, S.; Falkenberg, I.; Martin-Santos, R.; Atakan, Z.; Crippa, J.A.; Giampietro, V.; Brammer, M.; McGuire, P. Cannabinoid modulation of functional connectivity within regions processing attentional salience. *Neuropsychopharmacol. Off. Publ. Am. Coll. Neuropsychopharmacol.* **2015**, *40*, 1343–1352. [\[CrossRef\]](#)
167. Davidson, B.L.; Gao, G.; Berry-Kravis, E.; Bradbury, A.; Bönnemann, C.; Buxbaum, J.D.; Corcoran, G.R.; Gray, S.J.; Gray-Edwards, H.; Kleiman, R.J.; et al. Gene-based therapeutics for rare genetic neurodevelopmental psychiatric disorders. *Mol. Ther.* **2022**, *30*, 2416–2428. [\[CrossRef\]](#)
168. Niederhofer, H. First preliminary results of an observation of Ginkgo Biloba treating patients with autistic disorder. *Phytother. Res.* **2009**, *23*, 1645–1646. [\[CrossRef\]](#)
169. Al-Ayadhi, L.Y.; Elamin, N.E. Camel milk as a potential therapy as an antioxidant in autism spectrum disorder (ASD). *Evid.-Based Complement. Altern. Med.* **2013**, *2013*, 602834. [\[CrossRef\]](#)
170. Ghalichi, F.; Ghaemmaghami, J.; Malek, A.; Ostadrahimi, A. Effect of gluten free diet on gastrointestinal and behavioral indices for children with autism spectrum disorders: A randomized clinical trial. *World J. Pediatr.* **2016**, *12*, 436–442. [\[CrossRef\]](#)
171. González-Domenech, P.J.; Díaz Atienza, F.; García Pablos, C.; Fernández Soto, M.L.; Martínez-Ortega, J.M.; Gutiérrez-Rojas, L. Influence of a combined gluten-free and casein-free diet on behavior disorders in children and adolescents diagnosed with autism spectrum disorder: A 12-month follow-up clinical trial. *J. Autism Dev. Disord.* **2020**, *50*, 935–948. [\[CrossRef\]](#)
172. Hyman, S.L.; Stewart, P.A.; Foley, J.; Peck, R.; Morris, D.D.; Wang, H.; Smith, T. The gluten-free/casein-free diet: A double-blind challenge trial in children with autism. *J. Autism Dev. Disord.* **2016**, *46*, 205–220. [\[CrossRef\]](#)
173. Lee, R.W.; Corley, M.J.; Pang, A.; Arakaki, G.; Abbott, L.; Nishimoto, M.; Miyamoto, R.; Lee, E.; Yamamoto, S.; Maunakea, A.K.; et al. A modified ketogenic gluten-free diet with MCT improves behavior in children with autism spectrum disorder. *Physiol. Behav.* **2018**, *188*, 205–211. [\[CrossRef\]](#) [\[PubMed\]](#)
174. Mazahery, H. The Role of Vitamin D and Omega-3 Long Chain Polyunsaturated Fatty Acids in Children with Autism Spectrum Disorder. Ph.D. Dissertation, Massey University, Albany, New Zealand, 2018.
175. Sender, R.; Fuchs, S.; Milo, R. Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol.* **2016**, *14*, e1002533. [\[CrossRef\]](#) [\[PubMed\]](#)
176. Qin, J.; Li, R.; Raes, J.; Arumugam, M.; Burgdorf, K.S.; Manichanh, C.; Nielsen, T.; Pons, N.; Levenez, F.; Yamada, T.; et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **2010**, *464*, 59–65. [\[CrossRef\]](#)
177. Proctor, L.M. The national institutes of health human microbiome project. In *Seminars in Fetal and Neonatal Medicine*; WB Saunders: Philadelphia, PA, USA, 2016; Volume 21, pp. 368–372.
178. Umbrello, G.; Esposito, S. Microbiota and neurologic diseases: Potential effects of probiotics. *J. Transl. Med.* **2016**, *14*, 1–11. [\[CrossRef\]](#)
179. Johnson, D.; Letchumanan, V.; Thurairajasingam, S.; Lee, L.H. A revolutionizing approach to autism spectrum disorder using the microbiome. *Nutrients* **2020**, *12*, 1983. [\[CrossRef\]](#) [\[PubMed\]](#)
180. Mazurek, M.O.; Shattuck, P.T.; Wagner, M.; Cooper, B.P. Prevalence and correlates of screen-based media use among youths with autism spectrum disorders. *J. Autism Dev. Disord.* **2012**, *42*, 1757–1767. [\[CrossRef\]](#) [\[PubMed\]](#)
181. Hicks, S.D.; Uhlig, R.; Afshari, P.; Williams, J.; Chroneos, M.; Tierney-Aves, C.; Wagner, K.; Middleton, F.A. Oral microbiome activity in children with autism spectrum disorder. *Autism Res.* **2018**, *11*, 1286–1299. [\[CrossRef\]](#)
182. Kong, X.; Liu, J.; Cetinbas, M.; Sadreyev, R.; Koh, M.; Huang, H.; Adeseye, A.; He, P.; Zhu, J.; Russell, H.; et al. New and preliminary evidence on altered oral and gut microbiota in individuals with autism spectrum disorder (ASD): Implications for ASD diagnosis and subtyping based on microbial biomarkers. *Nutrients* **2019**, *11*, 2128. [\[CrossRef\]](#)
183. Bercik, P.; Park, A.J.; Sinclair, D.; Khoshdel, A.; Lu, J.; Huang, X.; Deng, Y.; Blennerhassett, P.A.; Fahnestock, M.; Moine, D.; et al. The anxiolytic effect of Bifidobacterium longum NCC3001 involves vagal pathways for gut–brain communication. *Neurogastroenterol. Motil.* **2011**, *23*, 1132–1139. [\[CrossRef\]](#)

184. Olsen, I.; Singhrao, S.K. Can oral infection be a risk factor for Alzheimer's disease? *J. Oral Microbiol.* **2015**, *7*, 29143. [\[CrossRef\]](#)
185. Curtin, C.; Hubbard, K.; Anderson, S.E.; Mick, E.; Must, A.; Bandini, L.G. Food selectivity, mealtime behavior problems, spousal stress, and family food choices in children with and without autism spectrum disorder. *J. Autism Dev. Disord.* **2015**, *45*, 3308–3315. [\[CrossRef\]](#) [\[PubMed\]](#)
186. Olsen, I.; Hicks, S.D. Oral microbiota and autism spectrum disorder (ASD). *J. Oral Microbiol.* **2020**, *12*, 1702806. [\[CrossRef\]](#) [\[PubMed\]](#)
187. Segata, N.; Haake, S.K.; Mannon, P.; Lemon, K.P.; Waldron, L.; Gevers, D.; Huttenhower, C.; Izard, J. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol.* **2012**, *13*, R42. [\[CrossRef\]](#) [\[PubMed\]](#)
188. Winter, S.E.; Lopez, C.A.; Bäuml, A.J. The dynamics of gut-associated microbial communities during inflammation. *EMBO Rep.* **2013**, *14*, 319–327. [\[CrossRef\]](#) [\[PubMed\]](#)
189. Hajishengallis, G.; Darveau, R.P.; Curtis, M.A. The keystone-pathogen hypothesis. *Nat. Rev. Microbiol.* **2012**, *10*, 717–725. [\[CrossRef\]](#)
190. Darveau, R.P.; Hajishengallis, G.; Curtis, M.A. Porphyromonas gingivalis as a potential community activist for disease. *J. Dent. Res.* **2012**, *91*, 816–820. [\[CrossRef\]](#)
191. Jaber, M.A. Dental caries experience, oral health status and treatment needs of dental patients with autism. *J. Appl. Oral Sci.* **2011**, *19*, 212–217. [\[CrossRef\]](#)
192. Aas, J.A.; Paster, B.J.; Stokes, L.N.; Olsen, I.; Dewhirst, F.E. Defining the normal bacterial flora of the oral cavity. *J. Clin. Microbiol.* **2005**, *43*, 5721–5732. [\[CrossRef\]](#)
193. Jeffcoat, M.K.; Hauth, J.C.; Geurs, N.C.; Reddy, M.S.; Cliver, S.P.; Hodgkins, P.M.; Goldenberg, R.L. Periodontal disease and preterm birth: Results of a pilot intervention study. *J. Periodontol.* **2003**, *74*, 1214–1218. [\[CrossRef\]](#)
194. Han, Y.W.; Ikegami, A.; Bissada, N.F.; Herbst, M.; Redline, R.W.; Ashmead, G.G. Transmission of an uncultivated Bergeyella strain from the oral cavity to amniotic fluid in a case of preterm birth. *J. Clin. Microbiol.* **2006**, *44*, 1475–1483. [\[CrossRef\]](#)
195. Aagaard, K.; Ganu, R.; Ma, J.; Racusin, D.; Arndt, M.; Riehle, K.; Petrosino, J.; Versalovic, J. 8: Whole metagenomic shotgun sequencing reveals a vibrant placental microbiome harboring metabolic function. *Am. J. Obstet. Gynecol.* **2013**, *208*, S5. [\[CrossRef\]](#)
196. Bearfield, C.; Davenport, E.S.; Sivapathasundaram, V.; Allaker, R.P. Possible association between amniotic fluid micro-organism infection and microflora in the mouth. *BJOG Int. J. Obstet. Gynaecol.* **2002**, *109*, 527–533. [\[CrossRef\]](#) [\[PubMed\]](#)
197. Xiao, J.; Fiscella, K.A.; Gill, S.R. Oral microbiome: Possible harbinger for children's health. *Int. J. Oral Sci.* **2020**, *12*, 1–13. [\[CrossRef\]](#) [\[PubMed\]](#)
198. Xiong, J.; Chen, S.; Pang, N.; Deng, X.; Yang, L.; He, F.; Wu, L.; Chen, C.; Yin, F.; Peng, J. Neurological diseases with autism spectrum disorder: Role of ASD risk genes. *Front. Neurosci.* **2019**, *13*, 349. [\[CrossRef\]](#)
199. Cani, P.D. Human gut microbiome: Hopes, threats and promises. *Gut* **2018**, *67*, 1716–1725. [\[CrossRef\]](#)
200. Anwar, H.; Irfan, S.; Hussain, G.; Faisal, M.N.; Muzaffar, H.; Mustafa, I.; Mukhtar, I.; Malik, S.; Ullah, M.I. Gut microbiome: A new organ system in body. *Parasitol. Microbiol. Res.* **2019**, *1*, 17–21.
201. Principi, N.; Esposito, S. Gut microbiota and central nervous system development. *J. Infect.* **2016**, *73*, 536–546. [\[CrossRef\]](#)
202. Finegold, S.M.; Dowd, S.E.; Gontcharova, V.; Liu, C.; Henley, K.E.; Wolcott, R.D.; Youn, E.; Summanen, P.H.; Granpeesheh, D.; Dixon, D.; et al. Pyrosequencing study of fecal microflora of autistic and control children. *Anaerobe* **2010**, *16*, 444–453. [\[CrossRef\]](#)
203. Zhang, M.; Ma, W.; Zhang, J.; He, Y.; Wang, J. Analysis of gut microbiota profiles and microbe-disease associations in children with autism spectrum disorders in China. *Sci. Rep.* **2018**, *8*, 13918. [\[CrossRef\]](#)
204. Williams, B.L.; Hornig, M.; Buie, T.; Bauman, M.L.; Cho Paik, M.; Wick, I.; Bennett, A.; Jabado, O.; Hirschberg, D.L.; Lipkin, W.I. Impaired carbohydrate digestion and transport and mucosal dysbiosis in the intestines of children with autism and gastrointestinal disturbances. *PLoS ONE* **2011**, *6*, e24585. [\[CrossRef\]](#)
205. Qiao, Y.; Wu, M.; Feng, Y.; Zhou, Z.; Chen, L.; Chen, F. Alterations of oral microbiota distinguish children with autism spectrum disorders from healthy controls. *Sci. Rep.* **2018**, *8*, 1597. [\[CrossRef\]](#) [\[PubMed\]](#)
206. Van Ameringen, M.; Turna, J.; Patterson, B.; Pipe, A.; Mao, R.Q.; Anglin, R.; Surette, M.G. The gut microbiome in psychiatry: A primer for clinicians. *Depress. Anxiety* **2019**, *36*, 1004–1025. [\[CrossRef\]](#) [\[PubMed\]](#)
207. Bronson, S.L.; Bale, T.L. Prenatal stress-induced increases in placental inflammation and offspring hyperactivity are male-specific and ameliorated by maternal antiinflammatory treatment. *Endocrinology* **2014**, *155*, 2635–2646. [\[CrossRef\]](#) [\[PubMed\]](#)
208. Estes, M.L.; McAllister, A.K. Maternal immune activation: Implications for neuropsychiatric disorders. *Science* **2016**, *353*, 772–777. [\[CrossRef\]](#) [\[PubMed\]](#)
209. Connolly, N.; Anixt, J.; Manning, P.; Ping-I Lin, D.; Marsolo, K.A.; Bowers, K. Maternal metabolic risk factors for autism spectrum disorder—An analysis of electronic medical records and linked birth data. *Autism Res.* **2016**, *9*, 829–837. [\[CrossRef\]](#) [\[PubMed\]](#)
210. Wang, Y.; Kasper, L.H. The role of microbiome in central nervous system disorders. *Brain Behav. Immun.* **2014**, *38*, 1–12. [\[CrossRef\]](#)
211. Yassour, M.; Vatanen, T.; Siljander, H.; Hämäläinen, A.M.; Härkönen, T.; Ryhänen, S.J.; Franzosa, E.A.; Vlamakis, H.; Huttenhower, C.; Gevers, D.; et al. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Sci. Transl. Med.* **2016**, *8*, 343ra81. [\[CrossRef\]](#)
212. Korpela, K.; Salonen, A.; Virta, L.J.; Kekkonen, R.A.; Forslund, K.; Bork, P.; De Vos, W.M. Intestinal microbiome is related to lifetime antibiotic use in Finnish pre-school children. *Nat. Commun.* **2016**, *7*, 10410. [\[CrossRef\]](#)

213. Sandler, R.H.; Finegold, S.M.; Bolte, E.R.; Buchanan, C.P.; Maxwell, A.P.; Väisänen, M.L.; Nelson, M.N.; Wexler, H.M. Short-term benefit from oral vancomycin treatment of regressive-onset autism. *J. Child Neurol.* **2000**, *15*, 429–435. [[CrossRef](#)]
214. Collado, M.C.; Rautava, S.; Aakko, J.; Isolauri, E.; Salminen, S. Human gut colonisation may be initiated in utero by distinct microbial communities in the placenta and amniotic fluid. *Sci. Rep.* **2016**, *6*, 23129. [[CrossRef](#)]
215. Jiménez, E.; Marín, M.L.; Martín, R.; Odriozola, J.M.; Olivares, M.; Xaus, J.; Fernández, L.; Rodríguez, J.M. Is meconium from healthy newborns actually sterile? *Res. Microbiol.* **2008**, *159*, 187–193. [[CrossRef](#)] [[PubMed](#)]
216. Fattorusso, A.; Di Genova, L.; Dell’Isola, G.B.; Mencaroni, E.; Esposito, S. Autism spectrum disorders and the gut microbiota. *Nutrients* **2019**, *11*, 521. [[CrossRef](#)] [[PubMed](#)]
217. Jašarević, E.; Howerton, C.L.; Howard, C.D.; Bale, T.L. Alterations in the vaginal microbiome by maternal stress are associated with metabolic reprogramming of the offspring gut and brain. *Endocrinology* **2015**, *156*, 3265–3276. [[CrossRef](#)] [[PubMed](#)]
218. Saunders, S.; Bocking, A.; Challis, J.; Reid, G. Effect of extracorporeal shock wave lithotripsy on bacterial viability-Relationship to the treatment of struvite stones. *Colloids Surf. B Biointerfaces* **2007**, *55*, 138. [[CrossRef](#)]
219. Cribby, S.; Taylor, M.; Reid, G. Vaginal microbiota and the use of probiotics. *Interdiscip. Perspect. Infect. Dis.* **2008**, *2008*, 256490. [[CrossRef](#)]
220. Bokobza, C.; Van Steenwinckel, J.; Mani, S.; Mezger, V.; Fleiss, B.; Gressens, P. Neuroinflammation in preterm babies and autism spectrum disorders. *Pediatr. Res.* **2019**, *85*, 155–165. [[CrossRef](#)]
221. Careaga, M.; Murai, T.; Bauman, M.D. Maternal immune activation and autism spectrum disorder: From rodents to nonhuman and human primates. *Biol. Psychiatry* **2017**, *81*, 391–401. [[CrossRef](#)]
222. Brown, A.S. Epidemiologic studies of exposure to prenatal infection and risk of schizophrenia and autism. *Dev. Neurobiol.* **2012**, *72*, 1272–1276. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



ACCEPTED MANUSCRIPT • OPEN ACCESS

# Analysis of the properties of recycled aggregates concrete with lime and metakaolin

To cite this article before publication: Manvendra Verma *et al* 2023 *Mater. Res. Express* in press <https://doi.org/10.1088/2053-1591/acf983>

## Manuscript version: Accepted Manuscript

Accepted Manuscript is “the version of the article accepted for publication including all changes made as a result of the peer review process, and which may also include the addition to the article by IOP Publishing of a header, an article ID, a cover sheet and/or an ‘Accepted Manuscript’ watermark, but excluding any other editing, typesetting or other changes made by IOP Publishing and/or its licensors”

This Accepted Manuscript is © 2023 The Author(s). Published by IOP Publishing Ltd.



As the Version of Record of this article is going to be / has been published on a gold open access basis under a CC BY 4.0 licence, this Accepted Manuscript is available for reuse under a CC BY 4.0 licence immediately.

Everyone is permitted to use all or part of the original content in this article, provided that they adhere to all the terms of the licence <https://creativecommons.org/licenses/by/4.0>

Although reasonable endeavours have been taken to obtain all necessary permissions from third parties to include their copyrighted content within this article, their full citation and copyright line may not be present in this Accepted Manuscript version. Before using any content from this article, please refer to the Version of Record on IOPscience once published for full citation and copyright details, as permissions may be required. All third party content is fully copyright protected and is not published on a gold open access basis under a CC BY licence, unless that is specifically stated in the figure caption in the Version of Record.

View the [article online](#) for updates and enhancements.

# Analysis of the Properties of Recycled Aggregates Concrete with Lime and Metakaolin

Manvendra Verma<sup>1\*</sup>, Arti Chouksey<sup>2</sup>, Rahul Kumar Meena<sup>3</sup>, Indrajeet Singh<sup>4</sup>

<sup>1\*</sup>Corresponding Author, Department of Civil Engineering, GLA University, Mathura, Uttar Pradesh, India.

[Manvendra.verma@gla.ac.in](mailto:Manvendra.verma@gla.ac.in)

<sup>2</sup> Department of Civil Engineering, Deenbandhu Chhotu Ram University of Science and Technology, Murthal, Haryana, India.

[Arti.civil@dcrustm.org](mailto:Arti.civil@dcrustm.org)

<sup>3</sup> Department of Civil Engineering, Punjab Engineering College, Chandigarh, India.

[rahulmeena@pec.edu.in](mailto:rahulmeena@pec.edu.in)

<sup>4</sup> Department of Civil Engineering, Delhi Technological University, Delhi, India.

[Indracivil1191@gmail.com](mailto:Indracivil1191@gmail.com)

## ABSTRACT

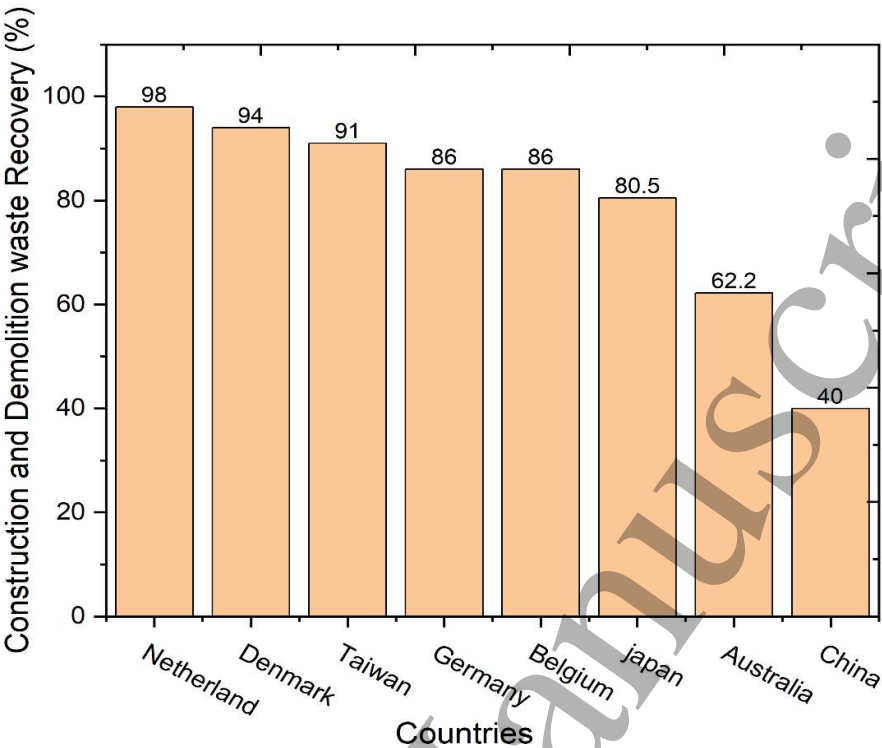
In recent years, the use of alternative materials in cementitious systems has attracted considerable interest due to their potential for augmenting the durability and performance of concrete. This research is investigating the use of three such materials as partial cement replacements in concrete: Recycled Concrete Aggregate (RCA), Limestone, and Metakaolin. RCA is a byproduct of the demolition of concrete structures that can be recycled as aggregate. Incorporating RCA into concrete reduces the environmental impact of waste disposal and reduces the carbon burden. Due to its pozzolanic properties, limestone, a sedimentary rock composed primarily of calcium carbonate, can be used as a substitute for cement. By substituting a portion of cement with limestone, the cement manufacturing process can substantially reduce carbon dioxide emissions. Metakaolin, a thermally treated form of kaolin clay, is yet another alternative material with pozzolanic properties. When used as a partial cement replacement, metakaolin increases the concrete's strength, durability, and chemical resistance. It also contributes to lowering hydration heat and mitigating alkali-silica reactions, thereby enhancing the durability of concrete structures. In this investigation, cement is replaced by limestone powder which is varied from 0% to 50% and the addition of metakaolin of 20% in every mix design. RCA is also incorporated in the mix design as a replacement for coarse aggregate by 20%. In the experimental investigation, various tests were conducted on each mix slump test, density, compressive strength, sulphate attack, mass loss, x-ray diffraction (XRD), and scanning electron microscope (SEM). After the investigation, the compressive strength improved by 15.07%, when metakaolin was added, and when LS was used to replace 10% of the cement, the compressive strength increased by 13.49%. The features of the combinations were negatively impacted when more cement was substituted. Following an investigation of hydration products, filler and dilution effects were found, both of which have the potential to be connected to improved mix quality. A mix that contains 20% metakaolin and 10% limestone powder may be considered the ideal mix owing to its superior strength and sulphate resistance when compared to normal concrete. It consists of less effect on slump value and density, the compressive strength was increased, and minimum mass loss after the sulphate attack. M3 mix best performer among all mix designs. It shows that the mix design with 20% metakaolin and 10% limestone powder is best-suitable for future recommendations.

**Keywords:** Cement replacement; Limestone; Metakaolin; Recycled concrete aggregate; Sustainable construction

## 1. INTRODUCTION

In recent years, there has been considerable interest in the use of alternative materials in cementitious systems due to their potential to improve the durability and performance of concrete [1], [2]. This study examines the use of RCA, Limestone, and Metakaolin as partial cement replacements in concrete. RCA is a byproduct of concrete structure demolition that can be recycled as aggregate. The incorporation of RCA into concrete decreases the environmental impact of waste disposal and the carbon footprint [3], [4]. Limestone, a sedimentary rock comprised predominantly of calcium carbonate, can be substituted for cement due to its pozzolanic properties. The cement manufacturing process can significantly reduce carbon dioxide emissions by substituting a portion of cement with limestone [5]. Metakaolin, a form of kaolin clay that has been thermally treated, is another alternative material with pozzolanic properties. When utilised as a partial cement replacement, metakaolin improves the strength, durability, and chemical resistance of concrete. It also reduces hydration heat and inhibits alkali-silica reactions, enhancing the durability of concrete structures [3], [6]. Recycled Aggregate Concrete (RCA) is a viable replacement for natural aggregate in typical concrete applications [7]–[10]. However, the feasibility of its usage in structural concrete has never been investigated, owing to its structural application limitations [11]. It is recommended that its use be licenced consistently with extra distinctive study and inquiry. The concrete industry's reliance on enormous amounts of natural resources has become a significant obstacle in addressing the enterprise's expanding needs. As the use of these ancient buildings decreases, they are destroyed to make room for new development. Such constructions are demolished for a variety of causes, including international-war damage, natural disasters, and new construction for monetary gain [12]. Globally, about 1 billion tonnes of construction and demolition (C&D) waste and concrete trash are produced annually, according to estimates [13], [14]. As a result, RCA from C&D waste is a great approach to provide a sustainable alternative for the building sector to satisfy its demand [15]. Utilized in structural concrete, RCA offers an abundance of possibilities. The whole production has more than doubled since 2007, when it was 21 billion tonnes, reaching 40 billion tonnes in 2014. The huge increase indicates the impact that the ongoing expansion of development throughout the globe has on the earth's natural resources [16]. Both India and China are the world's major producers of aggregate, with China's contribution accounting for around 38% of the total and India's contribution accounting for approximately 13% of the total [17]. As a result of the growing demand for aggregates, overall production is increasing, which not only contributes to the excessive exploitation of natural resources but also the formation of a wide variety of waste materials as by-products and the quickening of the rate of global warming [18].

Every last scrap of construction and demolition debris C&D waste that is generated in major developing countries is put to good use [13]. According to the findings presented in the studies that were carried out, the countries that have a high C&D recovery rate are displayed in **Figure 1**. These countries typically use recycled concrete aggregate (RCA) in the road base in both the United States and Europe [19]. RCA stands for recycled concrete aggregate. Investigations on the use of RCA as a structural component in newly produced concrete are now only being placed in European countries. These structures include residential, commercial, and social buildings [20]–[23]. Despite this, RCA continues to be used as a filler in road fills, landfills, and embankments all over the globe [24]. Despite this, India is the world's second-largest producer of construction and demolition waste, with an annual output of 530 million tonnes [25]. The vast bulk of the C&D waste that is produced in India is added to one of the most significant sources of solid trash in the nation [26]. It has been calculated that the worldwide production of concrete is around 70 million tonnes, which results in the release of over 65 million tonnes of carbon dioxide and accounts for almost 94% of all emissions of greenhouse gases into the atmosphere [27]. As a consequence, the use of recycled concrete aggregate, also known as RCA, might conserve up to 60 % of the natural aggregate resources, which would lead to sustainable development [28].



**Figure 1 Graph of Construction and demolition waste utilisation in various countries [10].**

High-temperature calcination of kaolin clay results in the transformation of its mineral structure into metakaolin. Metakaolin is dependent on the availability of deposits of high-quality kaolin clay. The United States, Brazil, China, and the United Kingdom are known to have significant reserves of kaolin clay, which can be used for the production of metakaolin [29]. However, metakaolin's regional availability and cost may affect its pervasive application. Incorporating metakaolin into concrete provides numerous benefits. It possesses pozzolanic properties, reacting with the calcium hydroxide produced during cement hydration to form additional cementitious compounds. This reaction increases the concrete's strength, durability, and chemical resistance. Additionally, metakaolin decreases the permeability of concrete, thereby enhancing its resistance to chloride ion ingress and reducing the risk of reinforcement corrosion. In addition, metakaolin reduces the heat of hydration, making it appropriate for applications involving mass concrete [30], [31]. Despite its potential benefits, using metakaolin in concrete presents several obstacles. The variability of metakaolin properties, which can be affected by the purity of the unprocessed kaolin clay, calcination conditions, and refining methods, presents a challenge. To ensure consistent performance, proper quality control measures and standardised testing methods are essential. The high cost of metakaolin is relative to other supplementary cementitious materials is a further obstacle. However, the long-term benefits, such as increased durability and decreased maintenance costs, may outweigh the upfront cost. The use of metakaolin in concrete extends to numerous construction sectors [32]–[34]. Common applications include high-performance concrete, self-consolidating concrete, and concrete with enhanced durability requirements. Depending on the desired performance characteristics, metakaolin can be incorporated as a partial replacement for cement, typically spanning from 5% to 20% by mass. It can also be combined with additional cementitious materials, such as fly ash or silica fume, to produce synergistic results. Metakaolin is a viable supplementary cementitious material for concrete that is readily available. Utilisation of this material can improve the mechanical properties, durability, and efficacy of concrete structures [35]–[38]. Even though availability and cost can differ, proper quality control and source selection of metakaolin can guarantee consistent and reliable performance.

Further research and development are required to optimise its use, standardise its testing procedures, and investigate its potential in various concrete applications [14], [29], [39]–[41].

Limestone powder is a finely-ground byproduct derived from limestone quarries or the refining of limestone during the production of cement. It is predominantly composed of calcium carbonate and has pozzolanic properties [42]. When used as an SCM, powdered limestone reacts with calcium hydroxide, a byproduct of cement hydration, to form additional cementitious compounds. This reaction, known as the pozzolanic reaction, increases the durability and strength of concrete. The use of limestone granules in concrete provides numerous benefits. In the first place, it reduces the carbon footprint of cement production [5]. By replacing a portion of cement with limestone powder, it is possible to reduce the quantity of clinker, the primary component of cement responsible for carbon dioxide emissions. This helps mitigate the environmental impact of the production of concrete [43]. Second, limestone powder improves the workability and cohesiveness of concrete mixtures, resulting in enhanced placement and finishing characteristics. In addition, it decreases the permeability of concrete, thereby enhancing its resistance to chloride ion penetration and extending its service life [44], [45]. It is essential to remember that the efficacy of limestone powder as an SCM is dependent on factors such as its fineness, chemical composition, and dosage. To ensure consistent and dependable performance, proper quality control and testing procedures are essential. Depending on desired performance characteristics, the optimal replacement level of limestone granules in concrete typically ranges from 5% to 15% by mass of cement [46]. Limestone powder has emerged as a valuable SCM in concrete, providing sustainability, strength, and durability benefits. Its use as a substitute for cement decreases carbon emissions and enhances the performance of concrete structures [47], [48]. To optimise its application, establish standardised guidelines, and investigate its compatibility with various cementitious systems, additional research and development are required. There is various way to sustainable construction, in which a geopolymer concrete is in the latest research. GPC have better performance in different sever conditions [49]–[57].

## 2. RESEARCH SIGNIFICANCE

This research is investigating the use of three such materials as partial cement replacements in concrete: Recycled Concrete Aggregate (RCA), Limestone, and Metakaolin. RCA is a byproduct of the demolition of concrete structures that can be recycled as aggregate. Incorporating RCA into concrete reduces the environmental impact of waste disposal and reduces the carbon burden. Due to its pozzolanic properties, limestone, a sedimentary rock composed primarily of calcium carbonate, can be used as a substitute for cement. In this investigation, cement is replaced by limestone powder which is varied from 0% to 50% and the addition of metakaolin of 20% in every mix design. The quality of the RCA is crucial. Proper processing and quality control during production are essential to ensure that the recycled material meets the required standards. In some cases, RCA might have a lower strength compared to natural aggregate, so a gradual replacement, like 20%, helps maintain the overall structural integrity of the concrete mix. RCA is also incorporated in the mix design as a replacement for coarse aggregate by 20%. In the experimental investigation, various tests were conducted on each mix slump test, density, compressive strength, sulphate attack, mass loss, x-ray diffraction (XRD), and scanning electron microscope (SEM). After the experimental and microstructural analysis recommend the best mix design with their better strength and durable properties.

## 3. EXPERIMENTAL PROGRAM

### 3.1. Materials

#### 3.1.1. Lime Stone

The main component of limestone, a sedimentary rock, is calcium carbonate ( $\text{CaCO}_3$ ) in the form of the mineral calcite. With a variety of qualities that make it desirable in the building, architectural, and

other sectors, it is a frequently utilised natural resource. Limestone is a relatively light substance with a density of 2.3 to 2.7 grammes per cubic cm (g/cm<sup>3</sup>). Depending on its composition, porosity, and other elements, limestone's compressive strength may vary significantly. Typically, it falls between 30 and 250 megapascals (MPa). Higher compressive strength limestone is preferred for load-bearing tasks like constructing bridges and other structures. Compared to its compressive strength, limestone typically has a lower tensile strength. Typically, the tensile strength is between 5 and 15 MPa. When building structures that might be subject to tension or bending pressures, it is crucial to take the tensile strength into account. On the Mohs scale, limestone has a hardness of around 3 to 4, making it a rather soft rock. Harder substances like quartz or diamond may readily scratch it. Depending on the kind and grade of limestone, the porosity might change. A rock contains microscopic openings called pores if it is porous. Limestone's durability and strength may be affected by increased porosity since it may allow for the absorption of water and other pollutants. Limestone has different levels of water absorption ability. Its ability to absorb water may range from 0.2% to 10%, depending on the porosity and surface characteristics. The ability of limestone to absorb water when exposed to moisture may alter its resilience and weathering properties. The thermal conductivity of limestone ranges from 1.3 to 4.4 watts per meter-kelvin (W/mK), which is considered to be moderate. Due to this characteristic, it is perfect for applications needing thermal insulation. Calcium carbonate (CaCO<sub>3</sub>) is the main chemical component of limestone, although it may also include other minerals and contaminants. Limestone may have many different colours and shades due to impurities, including white, beige, grey, and even veining. Limestone is a versatile material used in many different applications because of its properties, including the production of cement, dimension stones, architectural components, and soil stabilisation. Its accessibility, hardness, and aesthetic appeal have an impact on its popularity in the building and construction industry. Limestone is purchased from the locally available vendors. Several tests were conducted in the construction engineering, Department of Civil Engineering, GLA University. After the various test's conduction, limestone properties are found that are shown in **Table 1**.

**Table 1 Properties of limestone powder**

S. No.	Property	Value
1.	Colour	light grey - dark grey
2.	pH (of water slurry)	12
3.	Specific Density (g/cm <sup>3</sup> )	2.7
4.	Bulk Density (in loose state) (g/cm <sup>3</sup> )	0.8
5.	Chlorides content (%)	0.003

**3.1.2. Metakaolin**

Kaolin clay is calcined to yield metakaolin, a pozzolanic substance. A regulated heat process, usually between 600 and 800 degrees Celsius, transforms the kaolin clay into a reactive amorphous aluminosilicate material. Metakaolin is similar to other common construction materials in terms of density, ranging from 2.4 to 2.6 grammes per cubic cm (g/cm<sup>3</sup>). The specific composition and ratios of the mélange may affect the compressive strength of materials based on metakaolin. Metakaolin generally has a low compressive strength, however, it may boost concrete's strength development when employed as an addition. The compressive strength of concrete mixes containing metakaolin may be as high as 40 megapascals (MPa). Between 2 and 6 MPa, metakaolin generally has a relatively low tensile strength. However, by improving the material's overall microstructure and bonding qualities, it may improve the tensile strength properties of concrete when utilised as an additional cementitious ingredient. Granular metakaolin doesn't have a quantifiable hardness value. In most of its uses, hardness is not a significant factor. The relatively low porosity of metakaolin indicates that



it does not easily absorb water or other liquids. Due to its low porosity, pozzolanic material is effective in enhancing concrete's durability and chemical resistance. Metakaolin typically has a very low water absorption capacity, ranging from 0.2% to 0.2%. This characteristic of metakaolin, when employed as an addition, decreases the permeability of concrete, increasing its resistance to water penetration and enhancing durability. Metakaolin generally has a thermal conductivity of between 0.7 and 1.5 watts per meter-kelvin (W/mK). In cases where thermal insulation is required, this feature may be helpful. The main component of metakaolin is amorphous aluminosilicate, which is produced when kaolin clay is calcined. By reacting with calcium hydroxide (lime) in the presence of water to produce calcium silicate hydrate (C-S-H) gel and other cementitious compounds, it exhibits pozzolanic capabilities. As a cementitious addition, metakaolin is commonly used in cement and concrete applications. Numerous advantages are brought about by its pozzolanic nature, such as greater strength, lower permeability, improved durability, and better chemical resistance. It is admired for its capacity to support the sustainable use of resources by lowering dependency on traditional cement manufacturing. Metakaolin or calcined clay is purchased from the locally available vendors. Several tests were conducted in the construction engineering, Department of Civil Engineering, GLA University. After the various test's conduction, metakaolin properties are found that are that are shown in Table 2.

**Table 2 Properties of Metakaolin**

S. No.	Property	Value
1.	Specific gravity	2.60
2.	Bulk density (g/cm <sup>3</sup> )	0.36
3.	Physical form	Powder
4.	Colour	Off-White
5.	GE Brightness	81

### 3.1.3. RCA

Concrete, brick or other construction materials that have been destroyed are ground and refined to create recycled coarse aggregates. The characteristics of recycled coarse aggregates might vary depending on the nature of the recycled materials. Depending on the material composition, recycled coarse aggregates may have a range of densities. Its weight per cubic cm (g/cm<sup>3</sup>) is between 2.2 and 2.8, which is equivalent to that of natural aggregates. Compared to natural coarse aggregates, recycled coarse aggregates usually have a lower compressive strength. Depending on the original masonry or concrete's quality and characteristics, the strength may vary. It is essential to take recycled coarse aggregates' compressive strength into account when using them in load-bearing applications. Similar to compressive strength, recycled coarse aggregates often have lower tensile strengths than natural aggregates. The precise tensile strength might change depending on the make-up and characteristics of the original materials. The hardness of recycled coarse aggregates may vary based on the recycled components. The strength and endurance of the masonry or the original concrete might have an impact on hardness. Recycled coarse aggregates may range in porosity from low to high, depending on the original materials' porosity. The aggregates' permeability, durability, and resilience to freeze-thaw cycles are all influenced by porosity. Recycled coarse aggregates often have a higher water absorption capacity than unrecycled coarse aggregates. The porosity and surface properties of recycled materials may affect how well they are assimilated. The workability and durability of concrete and other applications using these aggregates may be impacted by greater water absorption. The thermal conductivity of recycled coarse aggregates is on par with that of natural aggregates. Depending on the particular recycled materials and their unique characteristics, it might change. The original

materials being recycled, such as concrete, brick, or bitumen, determine the chemistry of the recycled coarse aggregates. The characteristics of recycled aggregates may be impacted by any lingering additives or impurities from the original materials. It is crucial to remember that the characteristics of recovered coarse aggregates might vary significantly depending on elements including the calibre of the raw materials, processing methods, and compliance with recycling requirements. Proper quality management and testing are necessary to ensure the appropriateness and effectiveness of recycled coarse aggregates for several applications. RCA is collected from the crusher of municipal corporation of NCT Delhi. Several tests were conducted in the laboratory to find their properties. Various tests results are found of physical properties of recycled coarse aggregates are shown in **Table 3**.

**Table 3 Physical Properties of RCA**

S. No.	Properties	Value
1.	Specific gravity	2.42
2.	Aggregate crushing value (%)	28
3.	Bulk density (kg/m <sup>3</sup> )	1490
4.	Water absorption (%)	4
5.	Soundness (by Sodium sulphate solution) (%)	16.17
6.	Wet aggregate impact value (%)	20
7.	LA abrasion (%)	28

**3.1.4. Cement and Aggregates**

All the concrete materials were collected from the local available materials, and then various test were conducted in the laboratory on the samples. Normal OPC 43 cement is used in the research work. Their physical properties are shown in **Table 4**. Locally available coarse aggregates and stone dust are used in the project. Stone dust is used as fine aggregates. Stone dust physical properties are shown in **Table 5**, whereas coarse aggregates properties are shown in **Table 6**. **Figure 2** shows the gradation curve of stone dust and coarse aggregates.

**Table 4 Cement properties**

S. No.	Test	Result	As per IS 4031-1998
1.	Consistency	30%	30-35
2.	Initial setting time	40 min	Not less than 30min.
3.	Final setting time	1 hr 20 min	Not more than 600min
4.	Specific gravity	3.15	3.10-3.15
5.	Fineness	2.9%	Not exceed 10%
6.	Soundness	2mm	Not exceed 10mm

**Table 5 Fine Aggregate Properties**

S. No.	Test	Results
1.	Zone	Zone II
2.	Grade	Well Graded
3.	Fineness Modulus	2.756 (Medium sand)
4.	Specific Gravity	2.62
5.	Water absorption	1.21 %
6.	Silt Content	6 %
7.	Bulk density	1610 kg/m <sup>3</sup>

**Table 6 Coarse Aggregate Properties**

S. No.	Test	Results
1.	Fineness Modulus	7.29
2.	Specific Gravity	2.79
3.	Water absorption	0.2%
4.	Crushing Value	23%
5.	Impact Value	22%
6.	Flakiness Index	24%
7.	Elongation Index	30%
8.	Abrasion value	8%

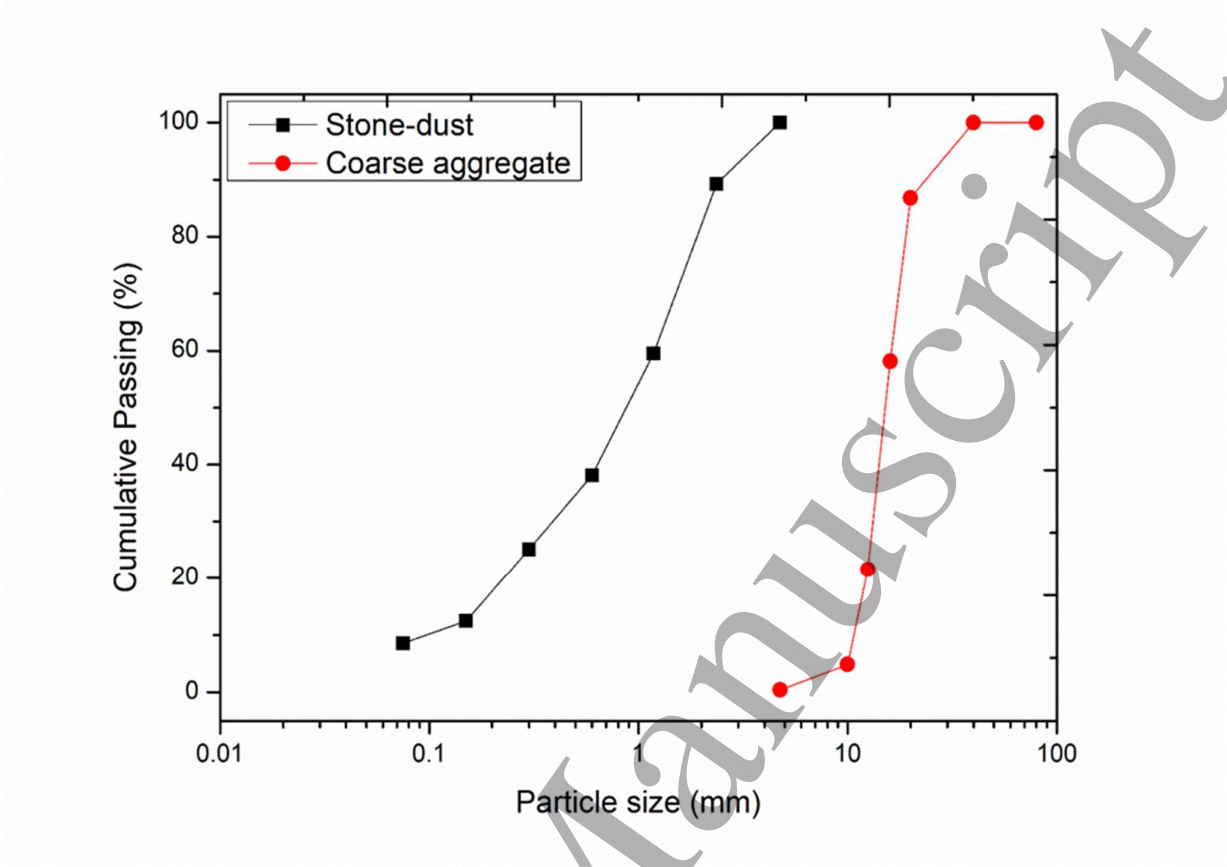


Figure 2 Gradation curve of aggregates

3.2. Synthesis

The quantitative estimate of the sample uses the proportions of all of the ingredients that were used in the fabrication of the sample. The mix identification for each of the various concrete mixes is shown in **Table 7** and **Table 8**. Each sample consists of six cubes measuring 15 cm by 15 cm by 15 cm and six cylinders measuring 10 cm in diameter and 20 cm in length, with an RCA content of 20%. As a result, the total volume that was computed for each sample was equal to 0.038 m<sup>3</sup>. All specimens were cast and cured as per Indian standards

Table 7 Mix Design

Mix Designs	Cement (kg/m <sup>3</sup> )	Metakaolin (kg/m <sup>3</sup> )	Lime Stone Powder (kg/m <sup>3</sup> )	Coarse Aggregate (kg/m <sup>3</sup> )	Fine Aggregate (kg/m <sup>3</sup> )	Super Plasticizer (kg/m <sup>3</sup> )	Water (kg/m <sup>3</sup> )
M1	370	-	-	1269	683	3.7	148
M2	370	74	-	1269	683	3.7	148
M3	333	74	37	1269	683	3.7	148
M4	296	74	74	1269	683	3.7	148
M5	259	74	111	1269	683	3.7	148
M6	222	74	148	1269	683	3.7	148
M7	185	74	185	1269	683	3.7	148

**Table 8 Variation of binding material in mix design**

Mix design	Cement (%)	Metakaolin (%)	Limestone Powder (%)
M1	100	0	0
M2	100	20	0
M3	90	20	10
M4	80	20	20
M5	70	20	30
M6	60	20	40
M7	50	20	50

### 3.3. Test Setups

The various tests conducted in the laboratories in this project are as follows:

#### 3.3.1. Slump Test

The slump test is a method that is frequently used to evaluate the consistency or workability of fresh concrete. It evaluates the flow and deformability of the concrete mixture, which reveals its flexibility and capacity for placement and compacting. assemble the equipment: Amass the necessary equipment, including a measuring scale or ruler, a base plate, a tamping rod, and a slouch cone. Make sure that there are no dirt or leftover concrete on the droop cone. Saturate the droop cone with water and let it drain before starting the test. During this stage, the cone is kept from collecting water from the concrete mix and affecting the test's outcomes. To produce the concrete sample, take a representative sample of recently-poured concrete from the quantity being evaluated. To achieve consistency and homogeneity, the concrete must be thoroughly mixed. Make sure the droop cone is centred and level before placing it on a flat, sturdy base plate. Throughout the test, you should use the base plate to catch any spilled concrete. Concrete should be poured into the slump cone in four equal thicknesses. Use a tamping rod to evenly odd each layer throughout its depth to compact it. Each layer should get around 25 tamping strokes, evenly distributed. After the cone has been filled with concrete, trim any extra using a straightedge or trowel. To level the concrete surface, use the cone's apex. The drooping cone should be gently raised vertically in a continuous upward motion without using any lateral or twisting forces. Lift the cone slowly and steadily to prevent jarring the concrete. Calculate the height difference between the centre of the displaced concrete mass and the cone's original height. The concrete depression is represented by this measurement. With the use of a measuring device or a ruler, calculate the droop in mm. By comparing the measured slump value to the given slump range or project-specific slump criteria, you may interpret the slump result. The consistency and workability of the concrete are indicated by the droop value. A concrete mixture that is more fluid or practicable will have a higher slump, whereas a combination that is firmer or less fluid would have a smaller slump. After the test is over, properly clean the slump cone, tamping rod, and other tools to get rid of any leftover concrete. A frequent and simple method for figuring out how cohesive concrete is the slump test. To make sure that the concrete mixture satisfies the desired workability and compaction criteria, it offers useful information [58].

#### 3.3.2. Density

The density of a concrete cube is a crucial test for determining the strength and quality of the concrete. It is carried out in accordance with defined procedures to deliver consistent and reliable findings.

Typically, concrete cubes are formed when concrete is cast for a particular project. The same quantity of newly mixed concrete that was used during construction is used to create the slabs. Typically, the cubes are 150 mm x 150 mm x 150 mm. For the concrete slabs to become strong, they must undergo a regulated curing process after being cast. The cubes are stored for the customary 28-day curing period in a humid, temperature-controlled chamber. When the curing process is complete, the cubes are removed from the curing chamber and cleaned of any loose particles or dirt on their surface before being weighed. The crystals are then all individually weighed using an accurate scale. The volume of the concrete cube will subsequently be accurately measured in the following phase. The water displacement method is the most popular way to calculate volume, although there are other methods as well. While the cube is submerged, the water level rise in the water-filled container is being watched. The volume of the cube is then calculated by dividing the measured elevation of the water level by the density of the water. One may determine the concrete cube's density by dividing its mass by volume. Usually, the density is expressed in kg/m<sup>3</sup> or lb/ft<sup>3</sup> (kilogrammes per cubic metre or cubic feet, respectively). Multiple slabs made from the same amount of concrete are tested to ensure the measurement's correctness. This cube's values are used to compute the average density value [59].

**3.3.3. Compressive Strength**

The compressive strength of a concrete cube is a crucial factor that defines a material's resistance to compressive pressures. It is a crucial parameter for assessing the reliability and strength of concrete constructions. Preparing the Cube: Concrete cubes are cast using the same quantity of fresh concrete as the building concrete, much like the density measurement process. The cubes are typically 150 mm x 150 mm x 150 mm in size. During the casting process, it is essential to ensure proper compaction in order to remove any voids or air pockets. The concrete slabs are cured in a controlled atmosphere after being cast to give them strength. The cubes are kept in a humidified, temperature-controlled room for the typical curing time of 28 days. The increase of concrete strength depends on proper curing. The cubes are taken out of the curing chamber once the curing process is finished and prepared for testing. For the testing arrangement, the cube must be put on compression testing equipment with properly aligned platens. The platens give the cube a level, flat surface on which the imparted weight may be distributed evenly. The cube receives a steady rate of progressive and uniform compression stress. The load is often applied uniaxially, which means that it is directed along a single axis that is transverse to the cube's top and bottom sides. The weight is steadily raised until the cube breaks or fails. Throughout the testing process, a load cell attached to the testing apparatus measures the compressive strain given to the cube. The cube's deformation or strain is detected concurrently using strain sensors or displacement transducers. The stress-strain behaviour of the cube may be determined in part thanks to these data. The maximum strain that the concrete cube might withstand prior to failing is used to calculate the cube's compressive strength. The failure is often marked by a rapid reduction in applied strain. Divide the maximum load by the cube's cross-sectional area to get the compressive strength. For consistency and accuracy, multiple cubes from the same batch of concrete are examined. Based on the results derived from these cubes, the average compressive strength value is calculated. Typically, the average value of three examined cubes is reported [60].

**3.3.4. Sulphate Attack**

The sulphate attack test gauges how easily concrete will deteriorate if there are sulphate ions present in the soil or water. This test helps determine how long concrete will last in sulphate-rich conditions. To mimic a sulphate-rich environment, a sulphate solution is made. Depending on the testing standard or specific needs, the sulphate solution's concentration and pH level may change. Magnesium and sodium sulphate are two often used solutions. The concrete cubes are dipped into the sulphate solution. There must be no gaps or air spaces between the cubes and the solution, and they must be totally immersed. Depending on the project requirements or testing standard, the cubes are immersed in the sulphate solution for a different amount of time. Typical durations range from a few weeks to a few months. Throughout this period, the cubes are often inspected for evident indications of wear and tear. Periodically, the cubes are removed from the sulphate solution, cleaned, and visually



inspected for any physical alterations or indications of sulphate assault. Surface cracks, expansion, spalling, and discoloration are typical warning signals. Once the specified exposure period has elapsed, the cubes are tested for compressive strength using the previously mentioned method. The compressive strength of the cubes is assessed in comparison to cubes that haven't been treated to sulphate. This comparison aids in determining the extent of the harm caused by sulphate attack. Determining the degree of degradation and any potential impacts on the concrete's long-term resilience [61].

### 3.3.5. XRD

A typical technique for determining the mineralogical makeup of materials, particularly concrete, is X-ray diffraction (XRD). It offers useful details regarding the crystalline phases found in concrete, assisting in assessing its quality, spotting possible issues, and tracking changes over time. For XRD analysis, concrete samples must be produced. Typically, this entails digging cores using drilling equipment or removing tiny pieces of concrete from the building. The samples must be sufficiently homogenous and representative of the desired analysis region. The gathered concrete samples are pulverised into a fine powder and sieved to make sure the X-rays can pass through the substance and produce precise diffraction patterns. You may grind with a mechanical grinder or a mortar and pestle. After being ground into a powder, the mixture is sieved to eliminate any bigger particles that could interfere with the diffraction analysis. After that, a sample holder, such as a glass slide or one made especially for XRD analysis, is put on the sample of ground concrete. To ensure a consistent and stable analysis, the sample must be equitably distributed and securely affixed to the holder. The X-ray diffractometer must be calibrated before executing the XRD analysis. This calibration involves altering the parameters and alignment of the instrument to ensure precise measurements. For this purpose, calibration standards, such as standard reference material, may be used. The mounted concrete sample is inserted in the X-ray diffractometer for X-ray Diffraction analysis. The instrument emanates X-rays that interact with the sample's crystalline phases. Depending on the crystalline structure of the materials present in the concrete, the X-rays are diffracted at particular angles. The X-rays that are diffracted are then detected and recorded. The diffracted X-rays are typically captured as a diffraction pattern, which is a plot of intensity versus diffraction angle. The pattern of diffraction reveals the varieties of crystalline phases present in the concrete. Utilising software, the diffraction pattern is analysed, and compared to known reference patterns, and the crystalline phases are identified. The detected crystalline phases are interpreted and reported about concrete chemistry and mineralogy. The report includes specifics on the types and proportional distributions of the crystalline phases found in the concrete sample. This information may be used to assess the concrete's composition, hydration, and any possible durability concerns. It should be noted that the precise equipment and analysis methods may differ based on the X-ray diffractometer used and the study's goals [62].

### 3.3.6. SEM

SEM is a useful technique for studying the microstructure of concrete at extreme magnification. It gives in-depth details on the internal properties, morphology, and composition of concrete. A small, typical sample of concrete is taken for SEM examination. Carefully cutting or penetrating the concrete's surface will provide the sample. The sample must correctly depict the region of interest and be free of pollutants and dispersed particles. The concrete sample is attached to a specimen fragment or container using an adhesive substance, such as epoxy glue or conductive carbon tape. The mounting procedure's goals are to produce electrical conductivity for imaging and securely fasten the sample to the stem. To improve conductivity and avoid surface charges during SEM imaging, a thin conductive coating is added to the sample. Coating materials that are often utilised include carbon, gold, and gold-palladium alloy. The coating is applied using a sputter coater or another coating tool. The SEM device is calibrated before studying the concrete sample to ensure excellent imaging and accurate readings. The focus, brightness, magnification, and operating distance of the electron beam are only a few of the variables that must be adjusted throughout the calibration process.

The mounted and coated concrete sample is then placed inside the SEM chamber, where the device is run under a vacuum to facilitate imaging. A picture is produced by collecting secondary electrons that are released from the sample's surface when the electron beam scans over it. High-resolution pictures with magnifications ranging from a few to tens of thousands of times may be produced using the SEM. To understand the microstructure of the concrete, the SEM image data that were obtained from the study are analysed. This entails locating different phases, aggregates, cementitious components, fissures, cavities, and any other interesting structures. The data may also provide information about the quality, uniformity, and possible issues of the concrete. A report that contains SEM photos, results of the elemental analysis, and any important conclusions and suggestions details the findings and observations of the SEM investigation. It is essential to note that the specific SEM testing procedures and parameters can vary depending on the instrument employed, the characteristics of the sample, and the objectives of the analysis [63].

**4. RESULTS AND DISCUSSION**

**4.1. Slump Test**

The workability of the mixes that were made by IS: 1199-1959 was evaluated via the use of a slump test, which required the use of a slump cone and a tamping rod, both of which were described in the Indian Standards [58], [64]. The slump test that was performed in the laboratory while the design mixes were being prepared is seen in **Figure 3**. The range of slump values that were recorded for the various blends. **Figure 4** demonstrates quite clearly that there is a direct correlation between the amount of limestone powder in the concrete mixes and a considerable decrease in the workability of the concrete mixes. The dolomite lime stone powder is used for research work by the replacement of cement with the variation of 0 % to 50 %. RCA content is constant 20% of cement. It is fixed on the previous research.



Figure 3 Picture during the slump test

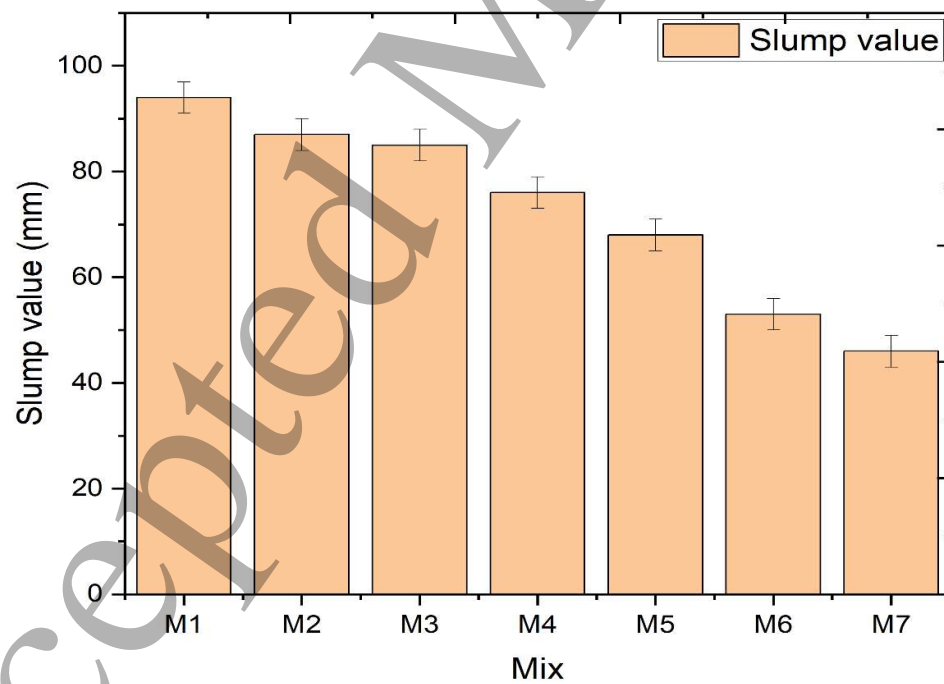
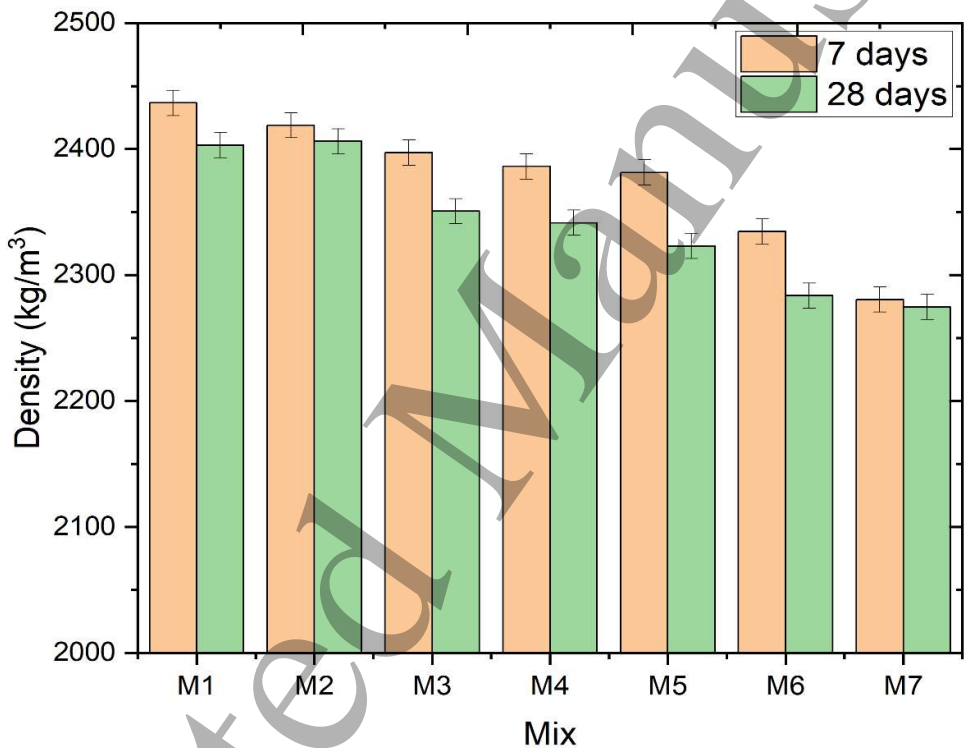


Figure 4 Graph between slump and mix design

#### 4.2. Density

It was necessary to estimate the density of each of the mixtures to be able to examine the disparities in weight that existed between the various mixtures [65]. The fluctuation in the density of concrete mixes as they aged from 7 days to 28 days is shown in **Figure 5**, which shows the differences between the two ages. When compared to the density of design mixes after 7 days, the density of design mixes after 28 days has a lower value. The hydration process of the mixtures is responsible for the loss of water that has occurred as a result of the procedure. The continuous decreases of density were occurring due to the addition of calcined clay and limestone powder. This resulted in a lower-density variation pattern. The use of powdered limestone instead of cement and the incorporation of metakaolin into the mixtures are both responsible for this result. The specific gravity of metakaolin and limestone powder is around two-thirds of cement, and the end product is noticeably lighter concrete [31]. Concrete that is often produced from mixtures that include additional cementitious ingredients that have a lower specific gravity can be described as having a lighter weight.



**Figure 5 Graph between density and mixed design**

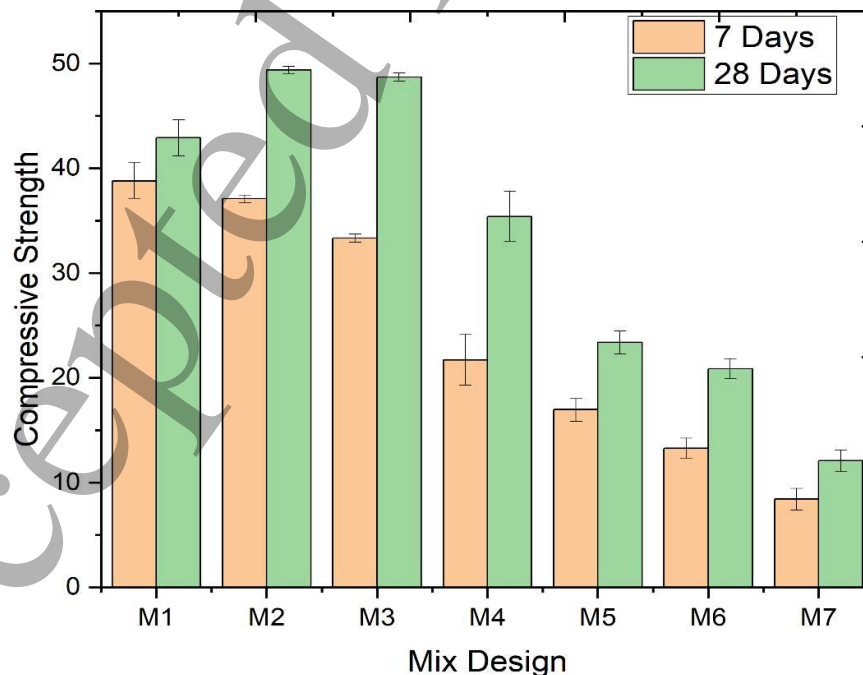
**4.3. Compressive Strength**

To ensure compliance with the standards given in IS 516-1959, concrete cubes with a variety of various mix designs have been cast and examined [60]. Following the findings of the literature research, a mixing sequence was developed. After being cast, the mixtures were left to cure for a total of twenty-eight days. After that, the compressive strength was measured after 7 days, 14 days, and then 28 days after the first measurement. The results of the compressive strength test of the concrete mixtures are shown in **Figure 6**. The compressive strength of the mixtures at the age of 28 days is shown in **Figure 6**. Compressive strength is calculated by the Eq. (1) which includes the load applied up to failure of specimens and cross-sectional area.

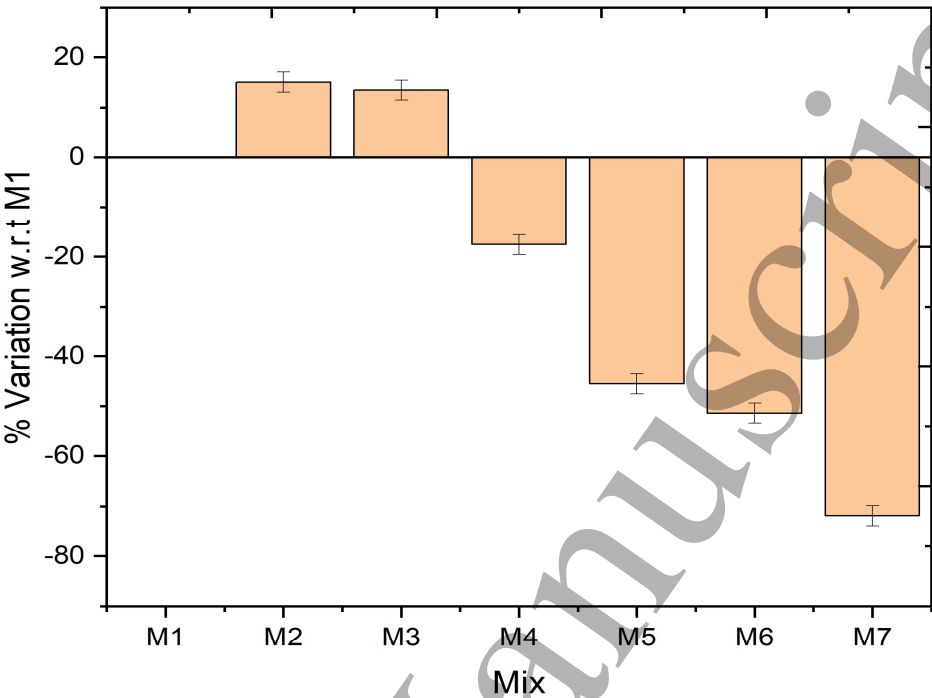
$$\text{Compressive Strength} = \frac{\text{Applied Load or Force}}{\text{Cross-section Area}} \dots\dots\dots(1)$$

Based on the findings shown in **Figure 6**, it was discovered that the compressive strength of the mixes rose when 20% of the cement was replaced within 28 days. With the addition of 10% limestone powder, there was a discernible weakening in the material's compressive strength. The compressive strength of the material decreased as a consequence of the continued addition of powdered limestone as a substitute. OPC was added to geopolymer concrete prepared from recycled concrete aggregate (RCA) to increase its mechanical and transport qualities [66]. The addition of 30% MK enhanced the compressive strength, porosity, and water absorption of recycled aggregate geopolymer concretes, with values of 134%, 69%, and 89%, respectively, compared to concrete without MK [16]. The effect of aggregate type and aggregate content on setting time was shown to be minor. Increased RCA content may cause an increase in the initial mixing water due to the additional water required by RA, which may extend the setting time somewhat [14]. the recyclability of geopolymer concrete aggregates in commonly used Portland cement concrete as a potential end-of-life waste management alternative for geopolymer concrete. At 20% RCA replacement, the compressive strength, modulus of elasticity, and flexural strength of RCA decreased by only around 14%, 1%, and 3%, respectively. This emphasises the need of making no or minor changes to the PCC mix at 20% RCA substitution for usage in structural applications [67].

The approval requirements of Clause 16.1 (a) of IS 456:2000 have been satisfied by the results of the compressive strength test that were achieved [68]. Nevertheless, the design mixes M5, M6, and M7 are not satisfy clause 16. (b) condition of IS 456:2000 standard. Hence, they are not suitable as concrete design mixes. The use of recycled concrete aggregate as the source may be credited with contributing to the considerably improved strength of the mixtures (RCA). It is possible to determine the high-strength concrete by the use of RCA aggregates in the mix design. On the other hand, this does not have any effect on the observed variance in compressive strength. **Figure 7** illustrates the percentage difference between each of the concrete mixes and M1, which serves as the reference Mix. In the recent research, predict the strength by using machine learning techniques without destructive tests [25], [69]–[74].



**Figure 6 Graph between compressive strength and mixed design**



**Figure 7 Residual compressive strength of various mix**

The source of the recycled concrete aggregate (RCA) used in mix M1 is high-strength concrete, which contributes to the mix's high compressive strength. After that, there is a high rise in the compressive strength of the M2 mix, which may be due to the addition of metakaolin that increases the reaction between them. The equation for pozzolanic reactions looks like this when written out:



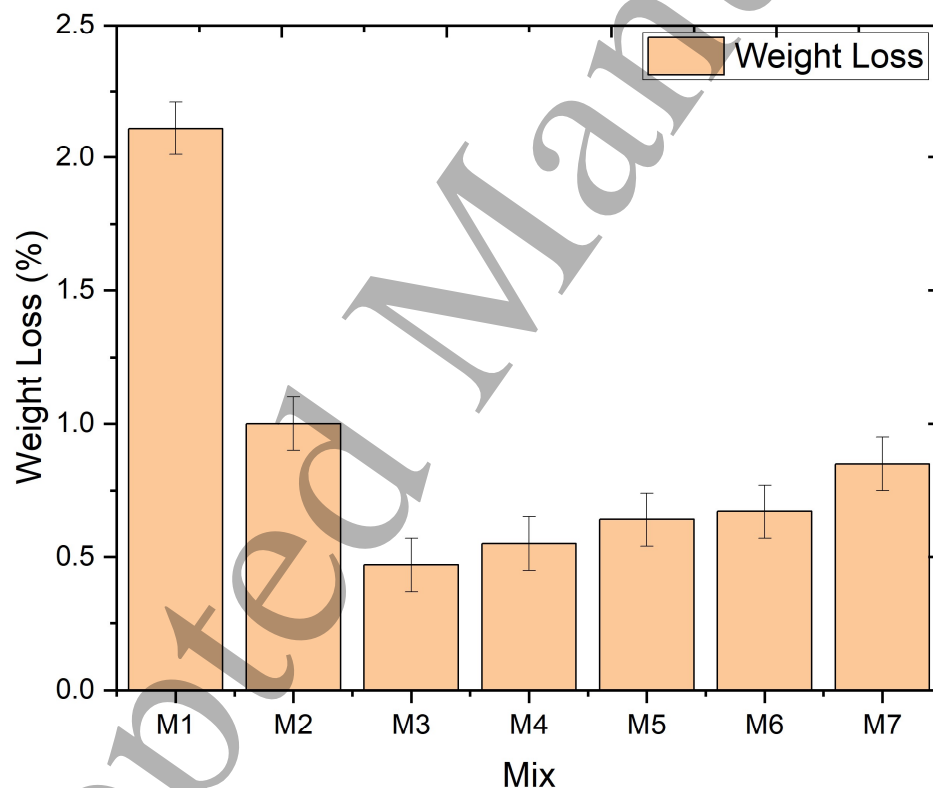
The incorporation of metakaolin caused the strength development in the mix design. This effect includes the interaction of metakaolin with calcium hydroxide (OH)<sub>2</sub> and results in the production of more CSH gel, which contributes to increased tensile strength. When limestone powder (at a replacement rate of 10%) is added to the design mix M3, there is a 13.49 % reduction in the material's compressive strength. This could have been caused by the filling effect of the limestone powder and the decrease in hydration products. Following this, there is a tendency for the compressive strength to fall to a significant degree with each rise in the proportion of cement that is replaced by limestone powder. This behaviour may be explained by the diluting effect brought about by the limestone powder. Because limestone powder has such a low accessible alumina concentration, the vast majority of it does not participate in the processes that are taking place. As a result, the replacement of the cement content results in the development of a smaller number of hydration products.

**4.4. Sulphate attack**

The different concrete mixtures were submerged for a total of 28 days in a solution containing 2.5% (v/v) magnesium sulphate, and the mixtures' resistance to sulphate attack was evaluated [61]. The concrete mixtures were tested to ascertain the amount of weight loss and the variance in compressive strength of concrete cylinders. According to the sulphate attack test procedure described in the standard code, this experiment was done by adding sulphate to water and making a solution, then the concrete mix specimens were kept in the solution for a specific time. **Figure 8** illustrates how the various kinds of concrete mixtures shed their respective amounts of weight in distinctive ways.



The findings shown in **Figure 8** make it abundantly evident that increasing the percentage of limestone powder in concrete mixes resulted in a general weakening of the mixture's compressive strength. Despite this, the percentage loss in compressive strength went down while concurrently experiencing a large decline. **Figure 8** shows that the M3 mix concrete specimens got the minimum mass loss after being exposed to sulphate attack. The degeneration of the hydration products in the absence of limestone powder may be related to the variance in the loss of strength as well as the percentage loss of weight [75]. Along with the addition of limestone powder, a rise in the effective water-cement ratio leads to an increase in the amount of limestone powder that is packed into the cement pores [54]. Because of its density and the fact that it does not allow water to get through, the concrete matrix that this product is more resistant to sulphate attack [76]. The stages of Ca-rich gel in the combined slag/fly ash system are decalcified by magnesium, which causes the binder system to break down and gypsum to precipitate. Dimensional instability and mechanical performance loss occur from the weakly cohesive and expanding nature of magnesium sulphate attack products. On the other hand, immersion of Na<sub>2</sub>SO<sub>4</sub> geopolymer pastes did not result in any visible binder breakdown or the conversion of the Binder phase components into sulphate-containing precipitates [54].



**Figure 8 Graph between weight loss of various mix**

#### 4.5. XRD (X-Ray Diffraction)

Because of the differences in how their examined qualities behaved, XRD tests were conducted on four different concrete mixes (i.e., M1, M2, M3, and M7) [62]. Cement was the sole binder in the reference design mix M1; the supplemental cementitious material in design mix M2 was metakaolin;



and the supplementary cementitious materials in design mix M3 and M7 were metakaolin and limestone powder, respectively. Because the compressive strength of M2 and M3 was found to be greater than that of the reference mix M1, they were both finalised in preparation for the XRD test. The design mix M7 was chosen for the XRD examination because it had the lowest compressive strength but the largest proportion of cement that had been replaced with limestone powder. The mineralogical characterisation of the reaction product of alkali activation of the BFAGC exhibits crystalline quartz (Q) and gmelinite (Gm). A partial crystalline phase was formed as a result of the interaction between PFA, POFA, and the sodium silicate activator. The major phase discovered was a crystalline N-A-S-H phase resembling albite ( $\text{NaAlSi}_3\text{O}_8$ ; PDF03-0451) [77], [78].

The XRD patterns of the mixes that were chosen for the XRD study are shown in **Figure 9**. These patterns include M1, M2, M3, and M7. The data were graphed with the help of the programme Origin, which is specifically designed for making graphs [55]. **Figure 9** shows a similar pattern of peaks in the XRD graph of different mix designs. This suggests that the concrete mixes all produced very comparable hydration products. The design mix M2 (consisting of 100% cement and 20% limestone powder) was found to have the greatest peak intensities, as was seen by the researchers. After that, a marginal lessening in the strength of the peaks was observed in the mix M3, and subsequently in the reference mix M1, respectively. The design mix M7 (which consisted of 50% cement, 20% metakaolin, and 50% limestone powder) was found to have the lowest peaks. This lower strength of the peaks across all of the mixes may be ascribed to the decreased hydration product production that occurred [79].

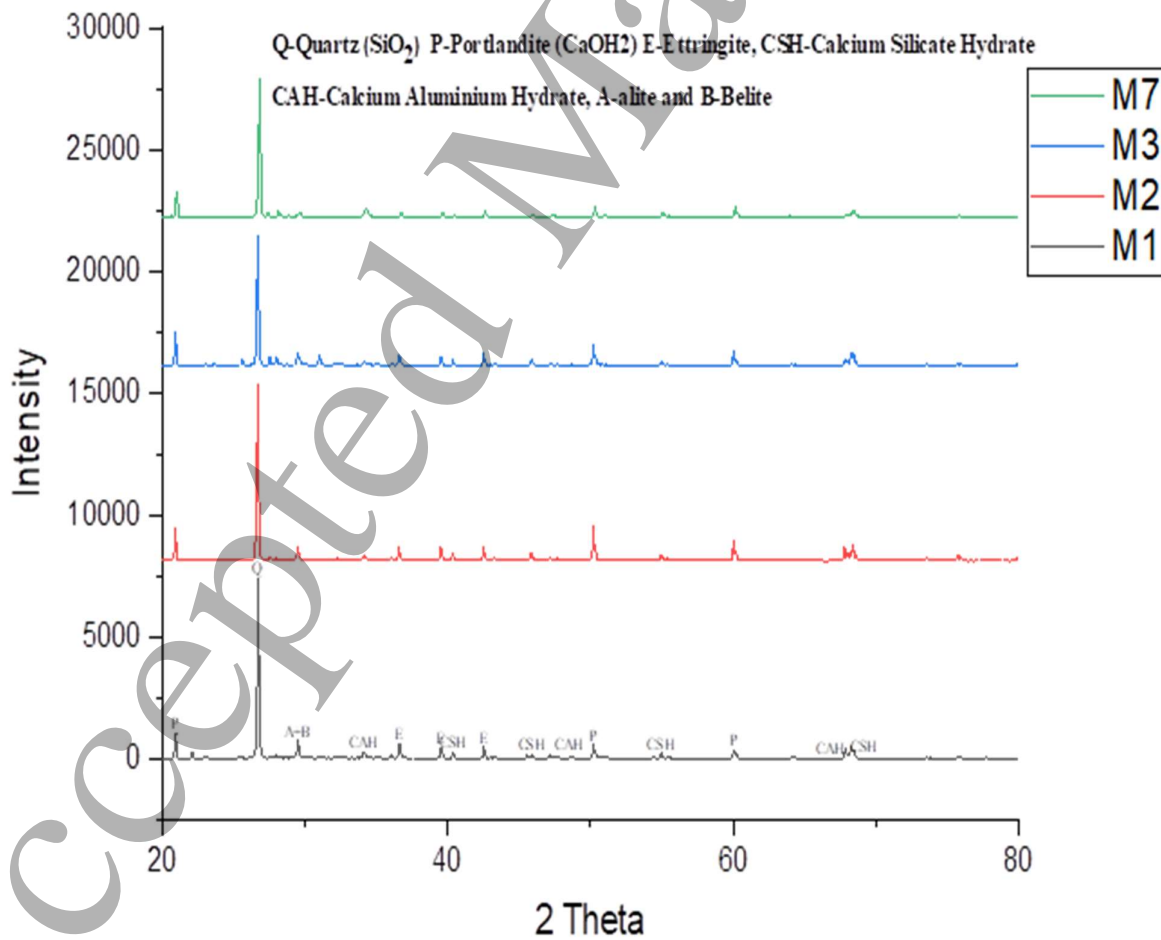
It was observed that the quartz ( $\text{SiO}_2$ ) peaks in mix M2 were more pronounced than those in mix M1. The large proportion of metakaolin and fine aggregate in the combination may have contributed to this. The peaks of Portlandite tending to be prominent in mix M2, which in turn is a sign of the consumption of Portlandite in mix M2, indicates the consumption of Portlandite in the pozzolanic activity of metakaolin, which results in the synthesis of strength-providing CSH gel [53]. The pozzolanic reaction of metakaolin is thought to be responsible for the fact that the peaks of CSH gel are often at their highest in M2. On the other hand, when limestone powder is used, the peaks of CSH gel tend to decline with M3. This might be because limestone powder has a filler effect, which decreases the development of CSH gel and could be the cause of the phenomenon [80]. The stable crystallisation of ettringite may be inferred from the fact that the ettringite peak is more prominent in mixes M2 and M3, respectively. The peaks of Alite, Belite, and CSH gel are at their lowest in mix M7, which may be due to the nucleation impact of limestone powder. As a result of the significant amount of cement that is being replaced, limestone powder contributes very little to the pozzolanic reaction [81]. This results in a decrease in the amount of cement clinker and hydration products [82]. As a consequence, one may get the following conclusion: a combination of M2 and M3 had superior crystal structures as well as greater hydration products, which led to increased compressive strength [83].

**4.6. Scanning Electron Microscope (SEM) Analysis**

Utilizing a scanning electron microscope, the effect of the added cementitious material on mixtures M1, M2, M3, and M7 was investigated. This material consisted of limestone powder and metakaolin. The incorporation of limestone powder and metakaolin has resulted in observable alterations to the microstructures of the material [84]. Previous research has revealed that the ITZ in traditional natural aggregate concrete is the weakest link in terms of mechanical performance and durability, and its creation is related to the wall effect induced by natural aggregates [7].

The existence of hydration products in CSH gel is seen in **Figure 10**. This hydration product is accountable for CSH gel's increased compressive strength. On the other hand, it can be seen that the Interfacial Transition Zone (ITZ) has a weak microstructure, and the voids are also extremely obvious to the naked eye. It has been seen that the ettringite needle formations are dispersed all over the place. **Figure 11** and **Figure 12** both show ettringite needles that are much shorter and thicker than those

seen in **Figure 10**. However, when metakaolin was used in the process, the crystallisation of the ettringite turned out to be quite stable. It has also been discovered that the CSH gel has a rather high density. It has been found that the Interfacial Transition Zone (ITZ) has improved, and the void size has been greatly reduced, which together contribute to an improvement in the final microstructure of the concrete mix. The reaction of calcined clay and its fine structure may be responsible for this, since they may have filled the holes that were left by the cement particles. The concrete mix M7 shown in **Figure 13** has the lowest compressive strength of all the mixes. Because of the decreased cement concentration, the ettringite crystals are hardly discernible to the naked eye. The CSH gel has a propensity for being dispersed, and the Interfacial Transition Zone (ITZ) has a propensity for becoming better when a denser composition is used. This may be due to the diluting effect created by the significant amount of cement that was replaced with powdered limestone. Because of this, the quantity of cement clinker is decreased, which in turn leads to a lower level of hydration products and, as a consequence, a lower level of compressive strength. On the other hand, M7 was the mix that proved to be the most sulfate-resistant because of its thick microstructure. The insertion of RCA particles above 20% wt enhances the elastic modulus of the RCA matrix substantially. This inclusion, however, diminishes tensile strength and overall strain. SEM images indicated a significant reduction in porosity throughout the extrusion to injection moulding process, which explains the lack of water absorption. However, breaking after moulding would be to blame for the reduction in strength and strain [7].



**Figure 9 XRD graph of various mix**

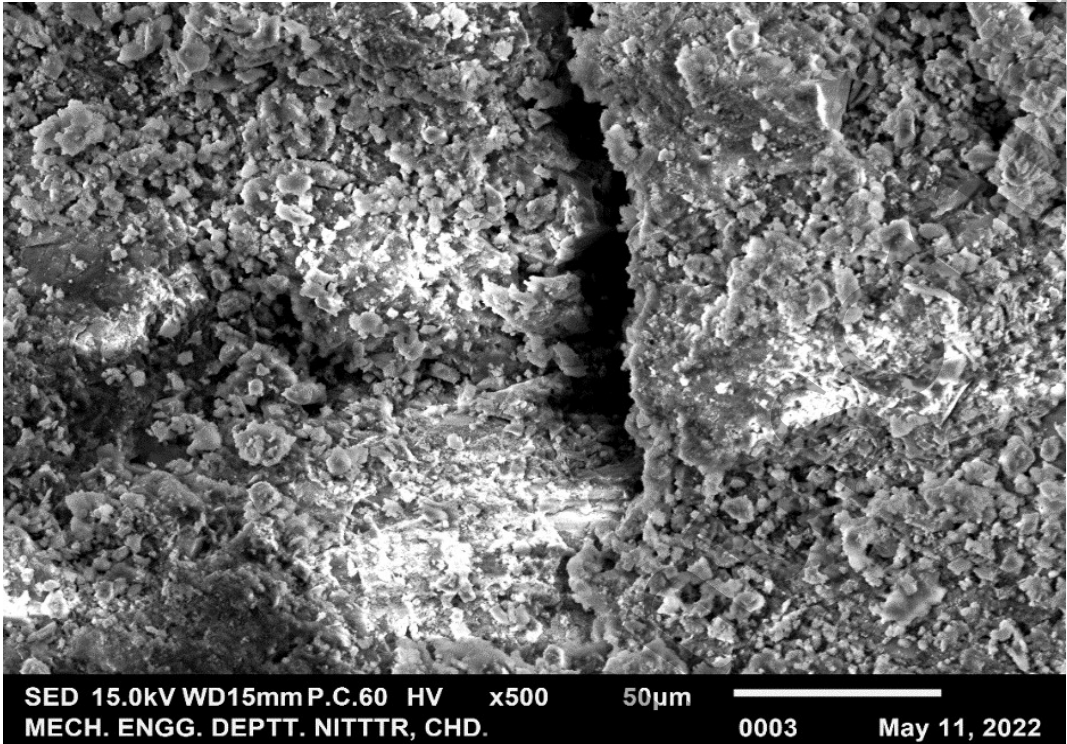


Figure 10 M1 mix SEM image

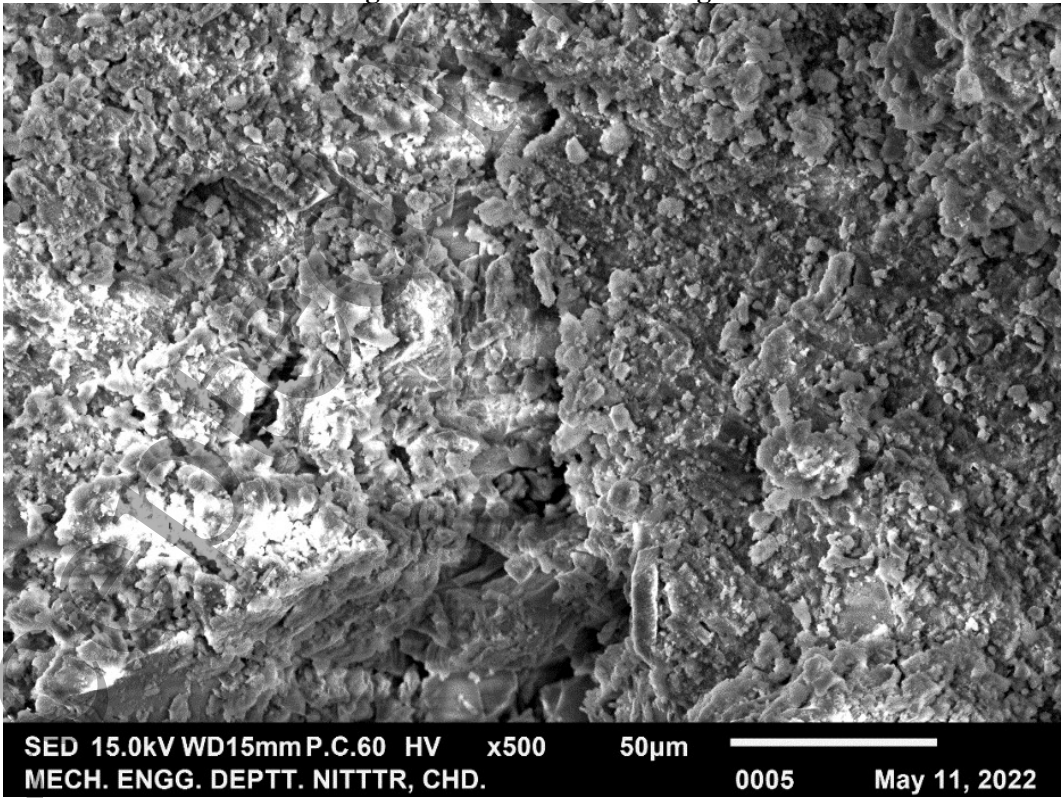


Figure 11 M2 mix SEM image

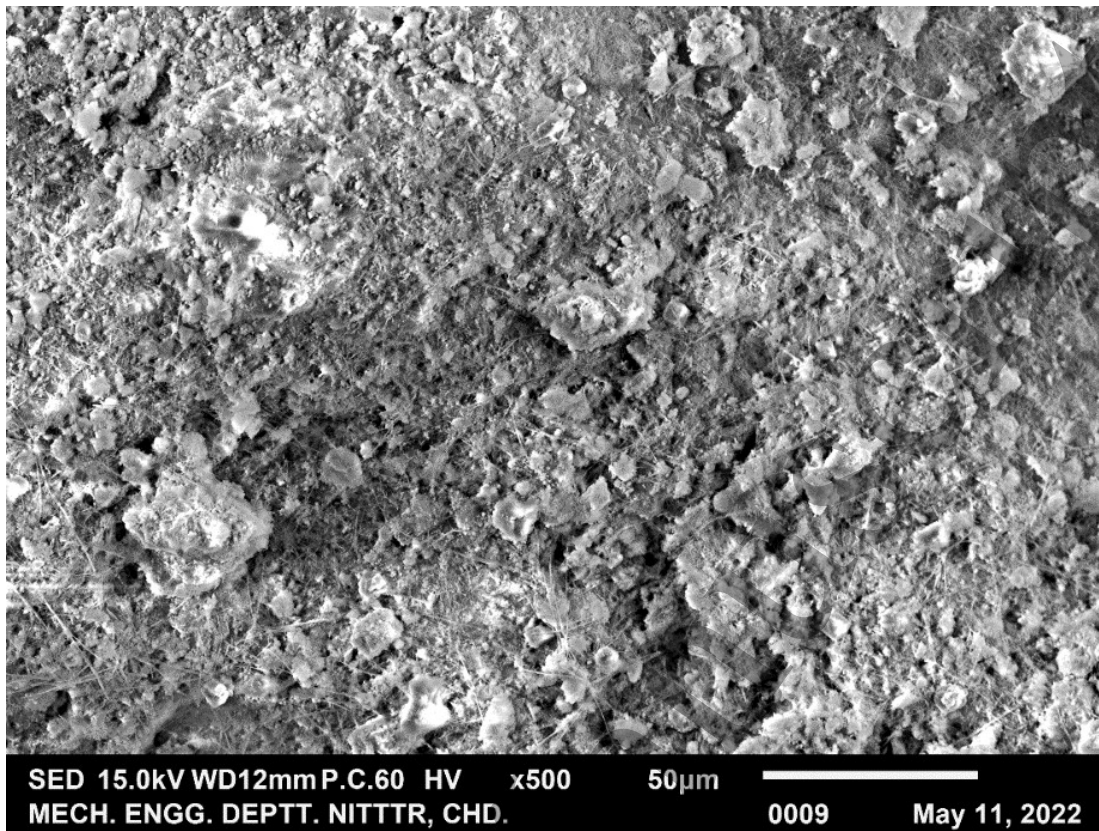


Figure 12 M3 mix SEM image

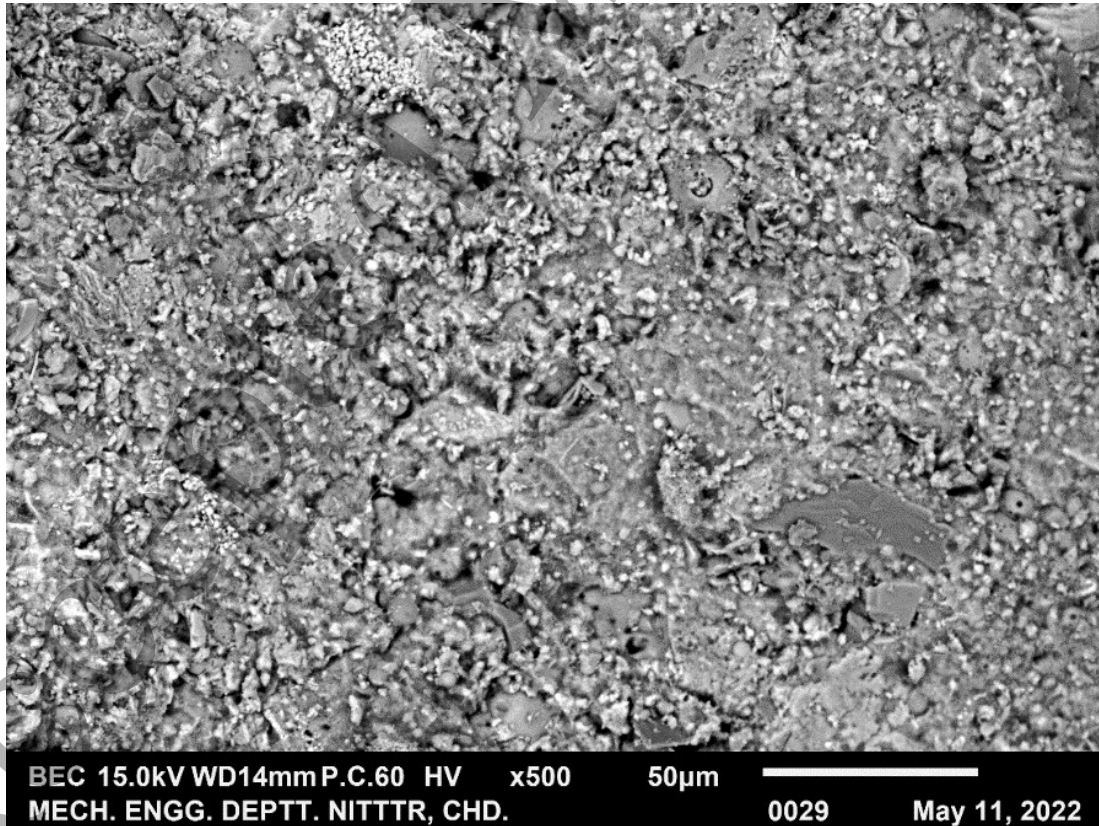


Figure 13 M7 Mix SEM image

## 5. CONCLUSION

This study looked at ways to enhance the quality and microstructure of concrete by using recycled concrete aggregate (RCA) with additional cementitious material, which in turn lowers the negative effect that it has on the environment and its natural resources. In addition, the outcomes that were achieved were the outcomes that had been sought.

- The inclusion of metakaolin and powdered limestone tends to reduce the workability of concrete mixtures. This may be explained by the fact that there is a greater variety in the limestone powder that is utilised in the mixes, which results in a higher need for water. When more cement is replaced with powdered limestone, there is often a correlation between this change and a drop in the density of the concrete mixtures. As a direct consequence of this, the concrete ended up being on the lighter side.
- Adding calcined clay as SCM at the same water-cement ratio has the potential to boost the compressive strength of the concrete. Because lime has a diluting effect, the amount of strength-providing hydration products in concrete is decreased when more than 20% of the cement is replaced by limestone powder. This can be attributed to the fact that adding more than 20% more cement leads to a reduction in the compressive strength of the concrete. Substituting cement with powdered limestone has the potential to improve the sulfate-resisting characteristics of concrete, and this potential benefit is likely to be realized. The concrete mix that had the most limestone powder was determined to have the lowest percentage loss in compressive strength, while the concrete mix that included 20% metakaolin had the greatest percentage loss in compressive strength.
- The resistance to sulphate attack was significantly decreased when cement was the sole binder used in the construction. The findings of the X-ray Powder Diffraction (XRD) test unequivocally demonstrate that the concrete mixes are very comparable to one another. The fact that the intensity of the peaks of the hydration compounds is getting weaker shows that the hydration process is taking longer, which means that fewer hydration products are being made. This is evidenced by the fact that the intensity of the peaks of the hydration compounds is getting weaker. The pictures captured by the scanning electron microscope (SEM) reveal that the microstructure of the concrete has been significantly improved. The microstructure of the concrete mix, which is rather thick, is likely to blame for the filler effect. When metakaolin and limestone powder is changed by 20%, up to a proportion of 10%, the ITZ tends to increase the strength of concrete specimens. Contrary to this, the nucleation effect increases the strength of concrete mixes and is related to the placement of the CSH gel. As a result, we can conclude that the concrete mix design M3 has superior strength and durability and that the optimal quantity of limestone powder to include in a mix that also included 20% metakaolin as a cementitious ingredient was 10%.

### Funding Statement

There is no funding for this research.

### Conflict of Interest

There are no conflicts of interest or competing interests in this article.

### Author Contribution

All authors have participated in (a) conception and design, analysis and interpretation of the data; (b) drafting the article or revising it critically for valuable intellectual content; and (c) approval of the final version.

### Availability of data and material

The available data had been used and discussed in the manuscript.



## Compliance with ethical standards

This manuscript has not been submitted to, nor is it under review at, another journal or other publishing venue.

## Consent to Participate

As a corresponding author or on behalf of all authors of the research paper, I consent to participate.

## Consent to publication

All authors of the research paper consent to the publication.

## Acknowledgement

This work is supported by the Civil Engineering Department, GLA University.

## REFERENCES

- [1] Nihal Arioglu; Z. Canan Girgin; and Ergin Arioglu, "Evaluation of Ratio between Splitting Tensile Strength and Compressive Strength for Concretes up to 120 MPa and its Application in Strength Criterion," *ACI Mater. J.*, no. January-February, pp. 18–24, 2006.
- [2] Y. H. M. Amran, R. Alyousef, H. Alabduljabbar, and M. El-Zeadani, "Clean production and properties of geopolymer concrete; A review," *J. Clean. Prod.*, vol. 251, 2020, doi: 10.1016/j.jclepro.2019.119679.
- [3] S. D. Vemu Venkata Praveen Kumar, Naga Prasad, "Influence of metakaolin on strength and durability characteristics of ground granulated blast furnace slag based geopolymer concrete," *Struct. Concr.*, no. November, pp. 1–11, 2019, doi: 10.1002/suco.201900415.
- [4] F. U. A. Shaikh, F. Uddin, and A. Shaikh, "Mechanical and durability properties of fly ash geopolymer concrete containing recycled coarse aggregates," *Int. J. Sustain. Built Environ.*, vol. 5, no. 2, pp. 277–287, 2016, doi: 10.1016/j.ijsbe.2016.05.009.
- [5] C. Selvamony, M. S. Ravikumar, S. U. Kannan, and S. B. Gnanappa, "Investigations on self-compacted self-curing concrete using limestone powder and clinkers," *J. Eng. Appl. Sci.*, vol. 5, no. 3, pp. 1–6, 2010.
- [6] P. Rovnanik, "Effect of curing temperature on the development of hard structure of metakaolin-based geopolymer," *Constr. Build. Mater.*, vol. 24, pp. 1176–1183, 2010, doi: 10.1016/j.conbuildmat.2009.12.023.
- [7] F. J. H. T. Vieira Ramos, R. H. M. Reis, I. Grafova, A. Grafov, and S. N. Monteiro, "Eco-friendly recycled polypropylene matrix composites incorporated with geopolymer concrete waste particles," *J. Mater. Res. Technol.*, vol. 9, no. 3, pp. 3084–3090, 2020, doi: 10.1016/j.jmrt.2020.01.054.
- [8] P. Nuaklong, P. Jongvivatsakul, T. Pothisiri, V. Sata, and P. Chindaprasirt, "Influence of rice husk ash on mechanical properties and fire resistance of recycled aggregate high-calcium fly ash geopolymer concrete," *J. Clean. Prod.*, vol. 252, p. 119797, 2020, doi: 10.1016/j.jclepro.2019.119797.
- [9] H. Nazarpour and M. Jamali, "Mechanical and freezing cycles properties of geopolymer concrete with recycled aggregate," *Struct. Concr.*, vol. 21, no. 3, pp. 1004–1012, 2020, doi: 10.1002/suco.201900317.

- [10] J. Turk, Z. Cotič, A. Mladenović, and A. Šajna, "Environmental evaluation of green concretes versus conventional concrete by means of LCA," *Waste Manag.*, vol. 45, pp. 194–205, 2015, doi: 10.1016/j.wasman.2015.06.035.
- [11] A. Bhogayata, S. V. Dave, and N. K. Arora, "Utilization of expanded clay aggregates in sustainable lightweight geopolymer concrete," *J. Mater. Cycles Waste Manag.*, pp. 1–13, 2020, doi: 10.1007/s10163-020-01066-7.
- [12] M. A. Khan, S. A. Memon, F. Farooq, M. F. Javed, F. Aslam, and R. Alyousef, "Compressive Strength of Fly-Ash-Based Geopolymer Concrete by Gene Expression Programming and Random Forest," *Adv. Civ. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/6618407.
- [13] B. B. Mukharjee and S. V. Barai, "Influence of Nano-Silica on the properties of recycled aggregate concrete," *Constr. Build. Mater.*, vol. 55, pp. 29–37, 2014, doi: 10.1016/j.conbuildmat.2014.01.003.
- [14] J. Xie, W. Chen, J. Wang, C. Fang, B. Zhang, and F. Liu, "Coupling effects of recycled aggregate and GGBS/metakaolin on physicochemical properties of geopolymer concrete," *Constr. Build. Mater.*, vol. 226, pp. 345–359, 2019, doi: 10.1016/j.conbuildmat.2019.07.311.
- [15] A. Nazari, A. Bagheri, J. Sanjayan, P. N. J. A. Yadav, and H. Tariq, "A Comparative Study of Void Distribution Pattern on the Strength Development between OPC-Based and Geopolymer Concrete," *Adv. Mater. Sci. Eng.*, vol. 2019, 2019, doi: 10.1155/2019/1412757.
- [16] P. Nuaklong, V. Sata, and P. Chindaprasirt, "Properties of metakaolin-high calcium fly ash geopolymer concrete containing recycled aggregate from crushed concrete specimens," *Constr. Build. Mater.*, vol. 161, pp. 365–373, 2018, doi: 10.1016/j.conbuildmat.2017.11.152.
- [17] S. M. A. Kabir *et al.*, "Performance evaluation and some durability characteristics of environmental friendly palm oil clinker based geopolymer concrete," *J. Clean. Prod.*, vol. 161, no. 2017, pp. 477–492, 2017, doi: 10.1016/j.jclepro.2017.05.002.
- [18] K. H. Younis and S. M. Mustafa, "Feasibility of Using Nanoparticles of SiO<sub>2</sub> to Improve the Performance of Recycled Aggregate Concrete," *Adv. Mater. Sci. Eng.*, vol. 2018, pp. 3–5, 2018, doi: 10.1155/2018/1512830.
- [19] Q. Wang, W. Ahmad, A. Ahmad, F. Aslam, A. Mohamed, and N. I. Vatin, "Application of Soft Computing Techniques to Predict the Strength of Geopolymer Composites," *Polymers (Basel)*, vol. 14, p. 1074, 2022.
- [20] A. Gupta, N. Gupta, and K. K. Saxena, "Mechanical and durability characteristics assessment of geopolymer composite (Gpc) at varying silica fume content," *J. Compos. Sci.*, vol. 5, no. 9, 2021, doi: 10.3390/JCS5090237.
- [21] P. Gupta, N. Gupta, K. K. Saxena, and S. Goyal, "A novel hybrid soft computing model using stacking with ensemble method for estimation of compressive strength of geopolymer composite," *Adv. Mater. Process. Technol.*, vol. 00, no. 00, pp. 1–16, 2021, doi: 10.1080/2374068X.2021.1945271.
- [22] R. Tomar, K. Kishore, H. Singh Parihar, and N. Gupta, "A comprehensive study of waste coconut shell aggregate as raw material in concrete," *Mater. Today Proc.*, vol. 44, pp. 437–443, 2021, doi: 10.1016/j.matpr.2020.09.754.
- [23] A. Gupta, N. Gupta, and K. K. Saxena, "Experimental study of the mechanical and durability properties of Slag and Calcined Clay based geopolymer composite," *Adv. Mater. Process. Technol.*, vol. 00, no. 00, pp. 1–15, 2021, doi: 10.1080/2374068X.2021.1948709.
- [24] P. K. Mehta and P. J. M. Monteiro, *Concrete Microstructure, Properties, and Materials*, vol.



4. 2014.

- [25] K. Upreti *et al.*, “Prediction of Mechanical Strength by Using an Artificial Neural Network and Random Forest Algorithm,” *J. Nanomater.*, vol. 2022, pp. 1–12, 2022, doi: DOI: 10.1155/2022/7791582.
- [26] P. P. Abhilash, D. K. Nayak, B. Sangoju, R. Kumar, and V. Kumar, “Effect of nano-silica in concrete; a review,” *Constr. Build. Mater.*, vol. 278, p. 122347, 2021, doi: 10.1016/j.conbuildmat.2021.122347.
- [27] M. Atiq Orakzai, “Hybrid effect of nano-alumina and nano-titanium dioxide on Mechanical properties of concrete,” *Case Stud. Constr. Mater.*, vol. 14, p. e00483, 2021, doi: 10.1016/j.cscm.2020.e00483.
- [28] S. M. Mustakim *et al.*, “Improvement in Fresh, Mechanical and Microstructural Properties of Fly Ash- Blast Furnace Slag Based Geopolymer Concrete By Addition of Nano and Micro Silica,” *Silicon*, 2020, doi: 10.1007/s12633-020-00593-0.
- [29] B. Sabir, S. Wild, and J. Bai, “Metakaolin and calcined clays as pozzolans for concrete: A review,” *Cem. Concr. Compos.*, vol. 23, no. 6, pp. 441–454, 2001, doi: 10.1016/S0958-9465(00)00092-5.
- [30] H. Y. Zhang, V. Kodur, B. Wu, L. Cao, and F. Wang, “Thermal behavior and mechanical properties of geopolymer mortar after exposure to elevated temperatures,” *Constr. Build. Mater.*, vol. 109, pp. 17–24, 2016, doi: 10.1016/j.conbuildmat.2016.01.043.
- [31] O. Vogt, C. Ballschmiede, N. Ukrainczyk, and E. Koenders, “Evaluation of sulfuric acid-induced degradation of potassium silicate activated metakaolin geopolymers by semi-quantitative sem-edx analysis,” *Materials (Basel)*, vol. 13, no. 20, pp. 1–23, 2020, doi: 10.3390/ma13204522.
- [32] D. L. Y. Kong, J. G. Sanjayan, and K. Sagoe-Crentsil, “Comparative performance of geopolymers made with metakaolin and fly ash after exposure to elevated temperatures,” *Cem. Concr. Res.*, vol. 37, pp. 1583–1589, 2007, doi: 10.1016/j.cemconres.2007.08.021.
- [33] A. Allahverdi, E. N. Kani, and M. Yazdanipour, “Effect of blast furnace slag natural pozzolan-based geopolymer cement,” *Ceram. – Silikáty*, vol. 55, no. 1, pp. 68–78, 2011, doi: 10.1016/j.clay.2016.07.020.
- [34] K. tuo Wang, Y. He, X. ling Song, and X. min Cui, “Effects of the metakaolin-based geopolymer on high-temperature performances of geopolymer/PVC composite materials,” *Appl. Clay Sci.*, vol. 114, pp. 586–592, 2015, doi: 10.1016/j.clay.2015.07.008.
- [35] A. M. Aguirre-Guerrero, R. A. Robayo-salazar, R. M. De Gutiérrez, and R. M. de Gutiérrez, “Applied Clay Science A novel geopolymer application: Coatings to protect reinforced concrete against corrosion,” *Appl. Clay Sci.*, vol. 135, pp. 437–446, 2017, doi: 10.1016/j.clay.2016.10.029.
- [36] K. Chandrasekhar Reddy, “Investigation of Mechanical and Microstructural Properties of Fiber-Reinforced Geopolymer Concrete with GGBFS and Metakaolin: Novel Raw Material for Geopolymerisation,” *Silicon*, 2020, doi: 10.1007/s12633-020-00780-z.
- [37] M. L. Kumar and V. Revathi, “Microstructural Properties of Alkali-Activated Metakaolin and Bottom Ash Geopolymer,” *Arab. J. Sci. Eng.*, vol. 45, no. 5, pp. 4235–4246, 2020, doi: 10.1007/s13369-020-04417-6.
- [38] I. Vegas, J. Urreta, M. Frías, and R. García, “Freeze-thaw resistance of blended cements containing calcined paper sludge,” *Constr. Build. Mater.*, vol. 23, no. 8, pp. 2862–2868, 2009,

doi: 10.1016/j.conbuildmat.2009.02.034.

- [39] Y. Aygörmmez, O. Canpolat, M. M. Al-mashhadani, and M. Uysal, "Elevated temperature, freezing-thawing and wetting-drying effects on polypropylene fiber reinforced metakaolin based geopolymer composites," *Constr. Build. Mater.*, vol. 235, no. 11752, pp. 1–32, 2020, doi: 10.1016/j.conbuildmat.2019.117502.
- [40] A. Celik, K. Yilmaz, O. Canpolat, M. M. Al-mashhadani, Y. Aygörmmez, and M. Uysal, "High-temperature behavior and mechanical characteristics of boron waste additive metakaolin based geopolymer composites reinforced with synthetic fibers," *Constr. Build. Mater.*, vol. 187, pp. 1190–1203, 2018, doi: 10.1016/j.conbuildmat.2018.08.062.
- [41] M. Alghannam, A. Albidah, H. Abbas, and Y. Al-Salloum, "Influence of Critical Parameters of Mix Proportions on Properties of MK-Based Geopolymer Concrete," *Arab. J. Sci. Eng.*, vol. 46, no. 5, pp. 4399–4408, 2021, doi: 10.1007/s13369-020-04970-0.
- [42] R. Abbas, M. A. Khareby, H. Y. Ghorab, and N. Elkhoshkhany, "Preparation of geopolymer concrete using Egyptian kaolin clay and the study of its environmental effects and economic cost," *Clean Technol. Environ. Policy*, vol. 22, no. 3, pp. 669–687, 2020, doi: 10.1007/s10098-020-01811-4.
- [43] S. Oyeibisi, A. Ede, F. Olutoge, and D. Omole, "Geopolymer concrete incorporating agro-industrial wastes: Effects on mechanical properties, microstructural behaviour and mineralogical phases," *Constr. Build. Mater.*, vol. 256, p. 119390, 2020, doi: 10.1016/j.conbuildmat.2020.119390.
- [44] F. Zunino and K. Scrivener, "Cement and Concrete Research The reaction between metakaolin and limestone and its effect in porosity refinement and mechanical properties," *Cem. Concr. Res.*, vol. 140, p. 106307, 2021, doi: 10.1016/j.cemconres.2020.106307.
- [45] M. Antoni, J. Rossen, F. Martirena, and K. Scrivener, "Cement and Concrete Research Cement substitution by a combination of metakaolin and limestone," *Cem. Concr. Res.*, vol. 42, no. 12, pp. 1579–1589, 2012, doi: 10.1016/j.cemconres.2012.09.006.
- [46] S. Ishak, H. S. Lee, J. K. Singh, M. A. M. Ariffin, N. H. A. S. Lim, and H. M. Yang, "Performance of fly ash geopolymer concrete incorporating bamboo ash at elevated temperature," *Materials (Basel)*, vol. 12, no. 20, pp. 1–17, 2019, doi: 10.3390/ma12203404.
- [47] L. Hu and Z. He, "A fresh perspective on effect of metakaolin and limestone powder on sulfate resistance of cement-based materials," *Constr. Build. Mater.*, vol. 262, p. 119847, 2020, doi: 10.1016/j.conbuildmat.2020.119847.
- [48] K. Hassannezhad, Y. Akyol, M. C. Dursun, C. W. Ow-yang, and M. A. Gulgun, "Effect of Metakaolin and Lime on Strength Development of Blended Cement Paste," *Constr. Mater.*, pp. 297–313, 2022.
- [49] M. Verma and N. Dev, "Sodium hydroxide effect on the mechanical properties of flyash-slag based geopolymer concrete," *Struct. Concr.*, vol. 22, no. S1, pp. E368–E379, 2021, doi: 10.1002/suco.202000068.
- [50] M. Verma and N. Dev, "Effect of ground granulated blast furnace slag and fly ash ratio and the curing conditions on the mechanical properties of geopolymer concrete," *Struct. Concr.*, vol. 23, no. 4, pp. 2015–2029, 2022, doi: 10.1002/suco.202000536.
- [51] M. Verma and N. Dev, "Effect of Liquid to Binder Ratio and Curing Temperature on the Engineering Properties of the Geopolymer Concrete," *Silicon*, vol. 14, no. 4, pp. 1743–1757, 2022, doi: 10.1007/s12633-021-00985-w.

- [52] M. Verma and N. Dev, "Effect of SNF-Based Superplasticizer on Physical, Mechanical and Thermal Properties of the Geopolymer Concrete," *Silicon*, vol. 14, no. 3, pp. 965–975, 2022, doi: 10.1007/s12633-020-00840-4.
- [53] A. Chouksey, M. Verma, N. Dev, I. Rahman, and K. Upreti, "An investigation on the effect of curing conditions on the mechanical and microstructural properties of the geopolymer concrete," *Mater. Res. Express*, vol. 9, no. 5, p. 55003, 2022, doi: 10.1088/2053-1591/ac6be0.
- [54] R. Kumar, M. Verma, and N. Dev, "Investigation on the Effect of Seawater Condition, Sulphate Attack, Acid Attack, Freeze–Thaw Condition, and Wetting–Drying on the Geopolymer Concrete," *Iran. J. Sci. Technol. Trans. Civ. Eng. Civ. Eng.*, vol. 46, no. 4, pp. 2823–2853, 2022, doi: 10.1007/s40996-021-00767-9.
- [55] R. Kumar, M. Verma, N. Dev, and N. Lamba, "Influence of chloride and sulfate solution on the long-term durability of modified rubberized concrete," *J. Appl. Polym. Sci.*, no. 139, pp. 1–15, 2022, doi: DOI: 10.1002/app.52880.
- [56] M. Verma and N. Dev, "Effect of Superplasticiser on Physical, Chemical and Mechanical Properties of the Geopolymer Concrete," in *Second ASCE India Conference on "Challenges of Resilient and Sustainable Infrastructure Development in Emerging Economies" (CRSIDE2020)*, 2020, pp. 1183–1189.
- [57] M. Verma, N. Dev, I. Rahman, M. Nigam, M. Ahmed, and J. Mallick, "Geopolymer Concrete: A Material for Sustainable Development in Indian Construction Industries," *Crystals*, vol. 12, no. 2022, p. 514, 2022, doi: 10.3390/cryst12040514.
- [58] IS 7320 1974, "Specification for concrete slump test apparatus," *Bur. Indian Stand.*, no. 2008, 2008, doi: 10.1136/archdischild-2011-300172.
- [59] IS: 1199 - 1959, "Methods of Sampling Sampling and Analysis of Concrete," 2018.
- [60] IS 516 1959, "Methods of Tests for Strength of Concrete," 2004.
- [61] ASTM C1898 20, "Standard Test Methods for Determining the Chemical Resistance of Concrete Products to Acid Attack," *ASTM Int.*, no. Stage II, pp. 9–10, 2020, doi: 10.1520/D1898-20.1.
- [62] ASTM D934 13, "Identification of Crystalline Compounds in Water-Formed Deposits by X-Ray Diffraction," 2014. doi: 10.1520/D0934-08.2.
- [63] ASTM E 1508-98, "Standard Guide for Quantitative Analysis by Energy-Dispersive Spectroscopy," 2003.
- [64] IS 7325 1974, "Specification for Apparatus for Determining Constituents of Fresh Concrete," 1999.
- [65] ASTM C138/C138M 17a, "Standard Test Method for Density ( Unit Weight ), Yield, and Air Content ( Gravimetric ) of Concrete," *ASTM Int.*, pp. 1–6, 2019, doi: 10.1520/C0138.
- [66] P. Nuaklong, V. Sata, A. Wongsu, K. Srinavin, and P. Chindaprasirt, "Recycled aggregate high calcium fly ash geopolymer concrete with inclusion of OPC and nano-SiO<sub>2</sub>," *Constr. Build. Mater.*, vol. 174, pp. 244–252, 2018, doi: 10.1016/j.conbuildmat.2018.04.123.
- [67] S. Mesgari, A. Akbarnezhad, and J. Z. Xiao, "Recycled geopolymer aggregates as coarse aggregates for Portland cement concrete and geopolymer concrete: Effects on mechanical properties," *Constr. Build. Mater.*, vol. 236, p. 117571, 2020, doi: 10.1016/j.conbuildmat.2019.117571.
- [68] IS 456 2000, "Plain and Reinforced Concrete Code of Practice," 2000.

- [69] K. Upreti and M. Verma, "Prediction of compressive strength of high-volume fly ash concrete using artificial neural network," *J. Eng. Res. Appl.*, vol. 1, no. December, pp. 24–32, 2022, doi: 10.55953/JERA.2022.2104.
- [70] U. Sharma, N. Gupta, and M. Verma, "Prediction of Compressive Strength of Geopolymer Concrete using Artificial Neural Network," *Asian J. Civ. Eng.*, pp. 1–14, 2023, doi: 10.1007/s42107-023-00678-2.
- [71] M. Verma, "Prediction of compressive strength of geopolymer concrete using random forest machine and deep learning," *Asian J. Civ. Eng.*, pp. 1–10, 2023, doi: 10.1007/s42107-023-00670-w.
- [72] M. Verma, "Prediction of compressive strength of geopolymer concrete by using ANN and GPR," *Asian J. Civ. Eng.*, pp. 1–9, 2023, doi: 10.1007/s42107-023-00676-4.
- [73] M. Verma *et al.*, "Prediction of Compressive Strength of Green Concrete by Artificial Neural Network," in *ICACIS 2022*, 2023, pp. 622–632. doi: /10.1007/978-3-031-25088-0\_55.
- [74] M. Verma, K. Upreti, M. R. Khan, M. S. Alam, S. Ghosh, and P. Singh, "Prediction of Compressive Strength of Geopolymer Concrete by Using Random Forest Algorithm," in *ICACIS 2022*, 2023, pp. 170–179. doi: 10.1007/978-3-031-25088-0\_14.
- [75] H. K. Kim, J. H. Jeon, and H. K. Lee, "Workability, and mechanical, acoustic and thermal properties of lightweight aggregate concrete with a high volume of entrained air," *Constr. Build. Mater.*, vol. 29, pp. 193–200, 2012, doi: 10.1016/j.conbuildmat.2011.08.067.
- [76] F. N. Okoye, J. Durgaprasad, and N. B. Singh, "Effect of silica fume on the mechanical properties of fly ash based-geopolymer concrete," *Ceram. Int.*, vol. 42, no. 2, pp. 3000–3006, 2016, doi: 10.1016/j.ceramint.2015.10.084.
- [77] P. Duxson, J. L. Provis, G. C. Lukey, J. S. J. van Deventer, J. S. J. Van Deventer, and J. S. J. van Deventer, "The role of inorganic polymer technology in the development of 'green concrete'," *Cem. Concr. Res.*, vol. 37, pp. 1590–1597, 2007, doi: 10.1016/j.cemconres.2007.08.018.
- [78] I. Ismail *et al.*, "Cement & Concrete Composites Modification of phase evolution in alkali-activated blast furnace slag by the incorporation of fly ash," *Cem. Concr. Compos.*, vol. 45, pp. 125–135, 2014, doi: 10.1016/j.cemconcomp.2013.09.006.
- [79] S. A. El-Ghany Abo El-Enein, E. A. El-Aziz Kishar, S. R. Refaey Zedan, and R. Abu-Elwafa Mohamed, "Effect of nano-SiO<sub>2</sub> (NS) on dolomite concrete towards alkali-silica reaction," *HBRC J.*, vol. 14, no. 2, pp. 165–170, 2018, doi: 10.1016/j.hbrj.2016.08.004.
- [80] C. A. Rosas-Casarez *et al.*, "Experimental study of XRD, FTIR and TGA techniques in geopolymeric materials," *Int. J. Adv. Comput. Sci. Its Appl.*, vol. 4, no. 4, pp. 25–30., 2014, [Online]. Available: [https://www.researchgate.net/profile/Jose\\_Gomez-Soberon/publication/274079395\\_Experimental\\_study\\_of\\_XRD\\_FTIR\\_and\\_TGA\\_techniques\\_in\\_geopolymeric\\_materials/links/55154d890cf2f7d80a32bf4c.pdf](https://www.researchgate.net/profile/Jose_Gomez-Soberon/publication/274079395_Experimental_study_of_XRD_FTIR_and_TGA_techniques_in_geopolymeric_materials/links/55154d890cf2f7d80a32bf4c.pdf)
- [81] R. Kumar, N. Dev, S. Ram, and M. Verma, "Investigation of dry-wet cycles effect on the durability of modified rubberised concrete," *Forces Mech.*, vol. 10, no. 2023, p. 100168, 2023, doi: 10.1016/j.finmec.2023.100168.
- [82] M. Nigam and M. Verma, "Effect of Nano-Silica on the Fresh and Mechanical Properties of Conventional Concrete," *Forces Mech.*, vol. 10, no. 22, p. 100165, 2023, doi: 10.1016/j.finmec.2022.100165.
- [83] H. M. Khater, "Studying the effect of thermal and acid exposure on alkali-activated slag

geopolymer,” *Adv. Cem. Res.*, vol. 29, no. 1, pp. 1–9, 2014, doi: 10.1680/adcr.11.00052.

- [84] W. K. W. Lee, J. S. J. Van Deventer, and J. S. J. Van Deventer, “The interface between natural siliceous aggregates and geopolymers,” *Cem. Concr. Res.*, vol. 34, no. July 2003, pp. 195–206, 2004, doi: 10.1016/S0008-8846(03)00250-3.

# Analyzing a higher order $q(t)$ model and its implications in the late evolution of the Universe using recent observational datasets

Madhur khurana,<sup>1,\*</sup> Himanshu Chaudhary,<sup>2,3,4,†</sup> Saadia Mumtaz,<sup>5,‡</sup> S. K. J. Pacif,<sup>3,§</sup> and G. Mustafa<sup>6,¶</sup>

<sup>1</sup>*Department of Applied Physics, Delhi Technological University, Delhi-110042, India*

<sup>2</sup>*Department of Applied Mathematics, Delhi Technological University, Delhi-110042, India*

<sup>3</sup>*Pacif Institute of Cosmology and Selfology (PICS), Sagara, Sambalpur 768224, Odisha, India*

<sup>4</sup>*Department of Mathematics, Shyamlal College, University of Delhi, Delhi-110032, India.*

<sup>5</sup>*Institute of Chemical Engineering and Technology,*

*University of the Punjab, Quaid-e-Azam Campus, Lahore-54590, Pakistan*

<sup>6</sup>*Department of Physics, Zhejiang Normal University, Jinhua 321004, People's Republic of China,*

In this research paper, we explore a well-motivated parametrization of the time-dependent deceleration parameter, characterized by a cubic form, within the context of late-time cosmic acceleration. The current analysis is based on the  $f(Q, T)$  gravity theory, by considering the background metric as the homogeneous and isotropic Friedmann–Lemaître–Robertson–Walker (FLRW) metric. Investigating the model reveals intriguing features of the late universe. To constrain the model, we use the recent observational datasets, including cosmic chronometer (CC), Supernovae (SNIa), Baryon Acoustic Oscillation (BAO), Cosmic Microwave Background Radiation (CMB), Gamma-ray Burst (GRB), and Quasar (Q) datasets. The joint analysis of these datasets results in tighter constraints for the model parameters, enabling us to discuss both the physical and geometrical aspects of the model. Moreover, we determine the present values of the deceleration parameter ( $q_0$ ), the Hubble parameter ( $H_0$ ), and the transition redshift ( $z_t$ ) from deceleration to acceleration ensuring consistency with some recent results of Planck 2018. Our statistical analysis yields highly improved results, surpassing those obtained in previous investigations. Overall, this study presents valuable insights into the higher order  $q(t)$  model and its implications for late-time cosmic acceleration, shedding light on the nature of the late universe.

## CONTENTS

I. Introduction	2	A. Comparison with the CC data points	11
II. Cosmological equation in $f(Q, T)$ gravity	3	B. Comparison with the type Ia supernova dataset	11
III. The Model	5	C. Relative difference between model and $\Lambda$ CDM	11
IV. Characterizing the Model through Dynamical Variables	5	VIII. Cosmography Parameters	12
V. Deriving Cosmological Parameters in terms of Redshift	6	A. The deceleration parameter	12
VI. Data Analysis	7	B. The jerk parameter	12
A. Methodology	7	IX. Statefinder diagnostic	13
B. Data Discription	7	X. Om Diagnostic	13
1. Cosmic Chronometers	7	XI. Physical Parameters	14
2. type Ia supernovae (SNIa)	8	A. Pressure Density $p$	14
3. Baryon Acoustic Oscillations	8	B. Energy Density $\rho$	14
4. Cosmic Microwave Background	9	C. Equation of State $\omega$	14
VII. Observational and theoretical comparisons of the Hubble Function and Distance Modulus Function	11	XII. Statistical Analysis	15
		XIII. Results	16
		XIV. Conclusion	18
		References	19

\* K.madhur2000@gmail.com

† himanshuch1729@gmail.com

‡ saadia.icet@pu.edu.pk

§ shibesh.math@gmail.com

¶ gmustafa3828@gmail.com

## I. INTRODUCTION

In contemporary times, one of the most significant breakthroughs is the discovery of cosmic accelerated expansion, which has been confirmed through a range of observational techniques [1, 2]. This expansion is accompanied by a mysterious energy component known as "dark energy" (DE), characterized by substantial negative pressure. Despite its enigmatic nature, DE constitutes about 70% of the cosmic content and plays a crucial role in maintaining the overall energy density of the universe consistent with predictions from inflationary theory. The investigation into the dominant constituents of the universe, namely dark energy and dark matter (DM), stands as a formidable challenge in modern physics, representing nearly 95% of the universe's imperceptible composition. Dark matter, an unseen type of matter exhibiting weak interactions with electromagnetic radiation, is detectable primarily through its gravitational effects on nearby ordinary matter. The existence of DM is supported by various observable phenomena, including rotation curves and mass discrepancies.

In the absence of any substantial evidence supporting the dark sources, alternative avenues have been explored extensively. The investigation of enigmatic approaches to these exotic terms has been delineated through two distinct methods: modifying matter sources or introducing additional degrees of freedom into the gravitational action. The initial approach involves the alteration of the matter sector within the Einstein-Hilbert Lagrangian density through the incorporation of various proposals. Numerous models have been proposed to explain DE, some of which correlate with observational evidence. Among these, the  $\Lambda$ CDM model is widely embraced due to its alignment with observations, although it also presents certain ambiguities such as fine-tuning and coincidence issues [3–7]. To address these limitations, alternative DE models like quintessence [8–10], phantom [11],  $k$ -essence [12–14], and Chaplygin gas [15] have been introduced, aiming to provide effective explanations for various cosmic inquiries and to explore diverse aspects of this exotic energy. Preceding the cosmic acceleration phase, the universe underwent a deceleration phase during its early epochs, where the impact of DE was relatively minor. It is believed that density perturbations during this period played a pivotal role in shaping cosmic structures. As a result, understanding the complete evolutionary timeline requires a cosmological model capable of describing both acceleration and deceleration phases.

The second approach involves an extension of the gravitational component in GR by introducing a DE source, while keeping the matter sector unchanged. As the first category holds intriguing implications, it has not gained as much support due to certain ambiguities.

In contrast, modified gravity frameworks have proven to be quite valuable due to their effective execution in the field of cosmology. Einstein formulated the concept of geometry-matter coupling whose adaptation has been integral to the development of General Relativity (GR). Several alternative theories of gravity, which incorporate additional curvature terms in the gravitational action, have been thoroughly examined in the scientific literature. Some well-known examples of these modified gravity theories encompass  $f(R)$  gravity [16], Gauss-Bonnet gravity [17],  $f(R, T)$  gravity [18], the scalar-tensor theory [19],  $f(T)$  gravity [20],  $f(T, T_G)$  gravity [21],  $f(Q)$  gravity [22], etc. These alterations are rooted in the metric tensor  $g_{ij}$  being considered a dynamic variable. Interestingly, an alternative approach to GR has been gaining momentum in the scientific literature, known as  $f(Q, T)$  gravity [23] appearing as an extension of symmetric teleparallel gravity, where  $Q$  and  $T$  correspond to the non-metricity and the trace of the energy-momentum tensor, respectively. This novel theory has sparked significant interest in investigating the late universe and has been the subject of different researchers in various contexts [24–27].

The Einstein field equations within the framework of General Relativity (GR) are notoriously intricate, comprising a set of nonlinear differential equations that pose significant challenges in terms of finding analytical solutions. To simplify these complexities, physicists often make certain physical assumptions, such as establishing relationships between the pressure and energy density of the universe's contents. However, when introducing components like dark energy, these equations become even more tangled. This complexity also extends to modified gravity (MG) theories, which incorporate higher-order derivative terms. To obtain tractable solutions for the field equations, whether within GR or MG theories, researchers employ a variety of techniques, including dynamical system analysis, autonomous systems, and notably, the model-independent approach. The model-independent approach, which typically involves the parametrization of any cosmological parameter, which is generally a functional form of the parameter. While the parametrization of cosmological parameters may initially appear adhoc, it proves to be a legitimate and powerful method when considering the broad spectrum of possible evolutions for the geometrical and physical parameters from a mathematical standpoint. This idea has garnered increasing interest in the realm of mathematical cosmology. This concept of cosmological parametrization has been extensively explored in the scientific literature, with comprehensive discussions by Pacif [28, 29], providing valuable insights into its application and significance in cosmological studies. There are also various studies on the model-independent approach by Eric V. Linder within some cosmological contexts. Specifically, he delved into the intricate studies of cosmological parametriza-



tion, a crucial facet of his research, which aimed to comprehensively study and understand a wide range of dark energy models that play a pivotal role in understanding the universe's accelerating expansion [30–34]

Modern cosmology advocates the investigation of kinematic quantities, often referred to as "Cosmography" or "Cosmo-kinetics." This approach relies on observational data and sidesteps prior assumptions about gravity theory or specific cosmological models. Cosmography's adoption of symmetry principles offers a means to navigate debates surrounding DE, dark matter, and related topics without invoking Einstein field equations (Friedmann equations). While cosmography does not directly involve the scale factor, it allows for some inference into its evolutionary history. The foundation of modern cosmology traces back to the work by Sandage [35], who introduced fundamental cosmographic parameters. These parameters include the Hubble parameter ( $H_0$ ), which dictates the rate of cosmic expansion, and the deceleration parameter ( $q_0$ ), responsible for accounting for the effects of gravity that slow down the expansion rate. Despite the emergence of DE altering the cosmological landscape, the dynamics of cosmic evolution remain intimately connected to the deceleration parameter ( $q$ ), defined as  $q = -\frac{a\ddot{a}}{\dot{a}^2}$ , where  $a(t)$  represents the scale factor. A positive value of  $q$  ( $\ddot{a} < 0$ ) indicates deceleration, while a negative value implies acceleration. The Hubble parameter characterizes the linear temporal evolution of  $a(t)$ , and a non-linear correction term ( $q_0$ ) introduces local instabilities and chaotic behavior [36]. The deceleration parameter ( $q$ ) also plays a vital role in examining the dynamics of observable galaxy numbers. Dark energy gains empirical support through various modifications to fit observational data, making its parameterization as a function of the scale factor ( $a(t)$ ) or redshift ( $z$ ) a valuable approach.

In the current paper, we have considered a parametrization of the deceleration parameter given in [37] within the classical gravity and extend the analysis in  $f(Q, T)$  gravity, which is also a special case of the parametrization considered in [28]. A plethora of parametric forms for the deceleration parameter have been explored, each with its own set of limitations. While some diverge at large cosmic times, others remain well-behaved at low redshifts ( $z \ll 1$ ) [38–74]. The adoption of parametric assumptions can occasionally lead to misinterpretations of the nature of DE, prompting an interest in non-parametric models that directly infer evolutionary mechanisms from observational data [75–79]. Nevertheless, the parameterization of  $q$  proves to be a more effective strategy for investigating the transition from cosmic deceleration to acceleration.

Highlighting the versatility and gravitational-theory independence of the parameterized  $q$  approach. In this study, we have made some analysis of the model pro-

duced by the cubic parametrization of the deceleration parameter and found tighter constraints to the model parameters involved. Additionally, we have done some cosmographic tests in order to differentiate a better-performing model when compared with cosmological data.

The paper is organized in the following manner. In section I, we provide an overview of the current state of modern cosmology, highlighting the various scenarios proposed to explain the late-time cosmic acceleration of the universe. We also outline the scope and objectives of the paper. In the section II, we introduce the  $f(Q, T)$  gravity framework and present the associated field equations. We derive the modified Friedmann equations within this context. In section III, we discuss a model-independent approach to parametrize the cosmological model. This section lays the foundation for our subsequent analyses. Section IV, focuses on characterizing the model through dynamical variables, providing insights into its behavior and evolution. In Section V, we derive all cosmological and physical parameters in terms of redshift, enhancing our understanding of the model's evolution. Section VI employs the Markov chain Monte Carlo (MCMC) method to constrain the model's parameters. We determine the best-fit values of model parameters by comparing them with the present Hubble rate function across different datasets. In Section VII, we validate the model's predictions by comparing them with observational data, Section VIII delves into the kinematic cosmographic parameters, including the deceleration and jerk parameters, providing a detailed analysis of their implications within the model. Sections IX and X are dedicated to the statefinder and  $Om$  diagnostic tests, offering further insights into the model's behavior and compatibility with observational data. In XI, we conduct a comprehensive statistical analysis to rigorously assess the model's performance and reliability. Sections XI and XII present the results of our study and draw conclusions based on our findings, summarizing the contributions and implications of this research.

## II. COSMOLOGICAL EQUATION IN $f(Q, T)$ GRAVITY

From the modification in Einstein-Hilbert action of general relativity by introducing a function of two scalar invariants,  $Q$  and  $T$ , which are constructed from the non-metricity and the trace of the energy-momentum tensor. The action in  $f(Q, T)$  is given by [80].

$$S = \int \left[ \frac{1}{16\pi} f(Q, T) + \mathcal{L}_M \right] \sqrt{-g} d^4x \quad (1)$$

where,  $g \equiv \det(g_{\mu\nu})$ , and  $\mathcal{L}_M$  is Lagrangian density. In Riemannian geometry, the metric tensor is always

symmetric. However, for our research, we take the  $f(Q, T)$  gravity, in which, we can take the non-symmetric part of the metric tensor called non-metricity. Which can be defined as

$$Q \equiv -g^{\mu\nu}(L_{\beta\mu}^\alpha L_{\nu\alpha}^\beta - L_{\beta\alpha}^\alpha L_{\mu\nu}^\beta) \quad (2)$$

where

$$L_{\beta\gamma}^\alpha \equiv -\frac{1}{2}g^{\alpha\lambda}(\nabla_\gamma g_{\beta\lambda} + \nabla_\beta g_{\lambda\gamma} - \nabla_\lambda g_{\beta\gamma}) \quad (3)$$

The trace of nonmetricity tensor is expressed as,

$$Q_\alpha \equiv Q_\alpha{}^\mu{}_\mu, \quad \tilde{Q}_\alpha \equiv Q^\mu{}_\alpha{}_\mu \quad (4)$$

The trace of the energy-momentum tensor and modification in the metric tensor are respectively

$$T_{\mu\nu} = -\frac{2}{\sqrt{-g}} \frac{\delta(\sqrt{-g}\mathcal{L}_M)}{\delta g^{\mu\nu}} \quad (5)$$

$$\Theta_{\mu\nu} = g^{\alpha\beta} \frac{\delta T_{\alpha\beta}}{\delta g^{\mu\nu}} \quad (6)$$

Finding the variation of action of the field equation Eq. (1) with respect to metric tensors.

$$8\pi T_{\mu\nu} = -\frac{2}{\sqrt{-g}} \nabla_\alpha (f_Q \sqrt{-g} P_{\mu\nu}^\alpha - \frac{1}{2} f g_{\mu\nu} + f_T (T_{\mu\nu} + \Theta_{\mu\nu}) - f_Q (P_{\mu\alpha\beta} Q_\nu^{\alpha\beta} - 2Q^{\alpha\beta\mu} P_{\alpha\beta\nu})) \quad (7)$$

Where super-momentum

$$P_{\mu\nu}^\alpha \equiv \frac{1}{4} \left[ -Q_{\mu\nu}^\alpha + 2Q_{(\mu}{}^\alpha{}_{\nu)} + Q^\alpha g_{\mu\nu} - \tilde{Q}^\alpha g_{\mu\nu} - \delta^\alpha{}_{(\mu} Q_{\nu)} \right] = -\frac{1}{2} L_{\mu\nu}^\alpha + \frac{1}{4} (Q^\alpha - \tilde{Q}^\alpha) g_{\mu\nu} - \frac{1}{4} \delta^\alpha{}_{(\mu} Q_{\nu)} \quad (8)$$

Taking the FLRW metric as follows,

$$ds^2 = -N(t)^2 dt^2 + a(t)^2 (dx^2 + dy^2 + dz^2), \quad (9)$$

where  $N(t)$  is the lapse function and  $a(t)$  is the scale factor. Hence,  $Q = 6H^2/N^2$ . We assume the value of  $N(t) = 1$ , for a standard case. Hence,  $Q = 6H^2$ .

To find the generalized Friedmann equations, assuming the matter content as the perfect fluid with the energy-momentum tensor  $T_\nu^\mu = \text{diag}(-\rho, p, p, p)$ . The tensor  $\Theta_\nu^\mu$  becomes,

$$\Theta_\nu^\mu = \delta_\nu^\mu p - 2T_\nu^\mu = \text{diag}(2\rho + p, -p, -p, -p) \quad (10)$$

For simplicity, taking  $F \equiv f_Q = dF/dt$  and  $8\pi\tilde{G} \equiv f_T = dF/dt$  the Friedmann equations we derived as follows,

$$8\pi\rho = \frac{f}{2} - 6FH^2 - \frac{2\tilde{G}}{1+\tilde{G}}(\dot{F}H + F\dot{H}) \quad (11)$$

$$8\pi p = -\frac{f}{2} + 6FH^2 + 2(\dot{F}H + F\dot{H}) \quad (12)$$

From Eqs (11) and (12), modified Einstein's field equations are derived

$$3H^2 = 8\pi\rho_{eff} = \frac{f}{4F} - \frac{4\pi}{F}[(1+\tilde{G})\rho + \tilde{G}p] \quad (13)$$

$$2\dot{H} + 3H^2 = -8\pi p_{eff} = \frac{f}{4F} - \frac{2\dot{F}H}{F} + \frac{4\pi}{F}[(1+\tilde{G})\rho + (2+\tilde{G})p] \quad (14)$$

From Eqs. (13) and (14) and the derivative of Eq. (13) we derive the continuity equation

$$\dot{\rho}_{eff} + 3H(\rho_{eff} + p_{eff}) = 0 \quad (15)$$

Although there are several forms of the function  $f(Q, T)$  is considered in the literature [81], we here only confined to the linear and additive form of  $f(Q, T)$  function [80, 82] in the form

$$f(Q, T) = \mu Q + \nu T \quad (16)$$

where,  $\mu$  and  $\nu$  are the non zero model constants. Hence the first derivatives  $f_Q = \mu$  and  $8\pi\tilde{G} = f_T = \nu$ . Solving the modified Friedmann equations and applying the barotropic equation of states  $p = \omega\rho$  we can find the equation of state parameter as follows

$$\omega = \frac{3H^2(8\pi + \nu) + \dot{H}(16\pi + 3\nu)}{\nu\dot{H} - 3H^2(8\pi + \nu)} \quad (17)$$

Hence the energy density equation turns out to be

$$\rho = \frac{-3H^2\mu(8\pi + \nu) + \mu\nu\dot{H}}{2(4\pi + \nu)(8\pi + \nu)} \quad (18)$$

To find the value of  $\dot{H}$  we use the relation  $a_0/a = 1 + z$  we can define a new relation between  $z$  and  $t$ .

$$\frac{d}{dt} = \frac{dz}{dt} \frac{d}{dz} = -(1+z)H(z) \frac{d}{dz} \quad (19)$$

normalizing the equation by taking the value of the scale factor as  $a_0 = a(t = t_0) = 1$ ,  $t_0$  refers the present time. Hence we can write the derivation Hubble parameter with respect to time in terms of red-shift as,

$$\dot{H} = -(1+z)H(z)\frac{dH}{dz} \quad (20)$$

### III. THE MODEL

The modified field equations, denoted as Eqs. (11) and (12) are two independent equations describing the dynamics of the model. These fundamental equations intertwined with three unknown variables:  $a$ ,  $\rho$ , and  $p$ . This implies that, in order to completely characterize and solve this system, we are in need of an additional equation. To address this, researchers in the field have adopt a widely accepted approach *the model independent way* study of model, often considers a scheme of parameterization of a cosmological parameter. In contemporary scientific literature, this approach is gaining prominence, as it is capable of solving the field equations in a model independent way that do not affect the background physics but provide a solution in simple way. It involves making an initial assumption regarding the scale factor, which can be done either directly or by expressing it in terms of cosmological parameters and their time derivatives. This strategic choice not only simplifies the equations but also allows for a more comprehensive exploration of the model's dynamics. To aid researchers in this endeavor, a wealth of parametrizations for various cosmological parameters has been meticulously compiled and organized in references [28] and [83]. These compilations serve as invaluable resources, facilitating the selection of appropriate parameterization schemes based on the specific characteristics and goals of a given study. There are a few intriguing models of dark energy and modified gravity based on various parametrization schemes of some geometrical parameters [84–88]. One noteworthy aspect of this concept is its ability to reconstruct the cosmic history and also the fate, while offering solutions to certain problems of standard cosmology.

In our own investigation, we opt to employ the parameterization of the deceleration parameter to effectively close the aforementioned system of equations. This choice aligns with established practices, streamlining our analytical framework while enabling us to delve deeply into the dynamics of our model. We employ here a scheme of parametrization of higher order time-dependent deceleration parameter ( $q(t)$ ) in the form,

$$q(t) = -1 + \frac{m}{n} - \frac{4}{n}t^3, \quad (21)$$

where  $m > 0$  and  $n > 0$  are two arbitrary constants. Similar form is also considered in the reference [37] in the

context of  $f(R, T)$  gravity. By using the definition of the deceleration parameter in the form of Hubble parameter,  $q = \frac{d}{dt} \left( \frac{1}{H} \right) - 1$ , we may obtain an explicit form of Hubble function as

$$H(t) = \frac{n}{t(m - t^3)} \quad (22a)$$

Note that this explicit form of the Hubble parameter can also be interpreted as a particular form of the parametrization given in Ref. [28] as a special case. Using the definition of the Hubble parameter in terms of scale factor will give the explicit form of scale factor as,

$$a = \beta \left( \frac{t^3}{m - t^3} \right)^{\frac{n}{3m}}, \quad (23)$$

The deceleration parameter plays a crucial role in characterizing the dynamics of the universe's expansion. It provides insights into whether the expansion is slowing down or accelerating. Specifically, when the deceleration parameter is greater than zero ( $q > 0$ ), it signifies a phase of decelerating expansion. Conversely, when the deceleration parameter is less than zero ( $q < 0$ ), it indicates an accelerating expansion phase. A particularly interesting point occurs when the deceleration parameter equals zero ( $q = 0$ ), marking a significant phase transition in the expansion of the universe. This phase transition at  $q = 0$  occurs at a specific time, denoted as  $t_{tr}$ , which can be calculated in this considered case as,  $t_{tr} = \sqrt[3]{\frac{m-n}{4}}$ . This implies,  $m$  must be greater than  $n$ . This moment represents a critical juncture in the universe's evolution, where the expansion dynamics shift fundamentally. We can observe in Eqs. (22a) and (23), for  $t = m^{1/3}$ ,  $H \rightarrow \infty$  and  $a \rightarrow \infty$ , implying the existence of a big rip singularity in this model, which is anticipated to happen in the near future, precisely at  $t = t_{end} = m^{1/3}$ . The universe's expansion reaches a point of extreme instability and divergence, characterized by this singularity.

### IV. CHARACTERIZING THE MODEL THROUGH DYNAMICAL VARIABLES

This section is dedicated to providing a comprehensive analysis of the physical behavior and properties inherent in our model. Within this context, we aim to elucidate the model's physical dynamics by examining key parameters, including energy density ( $\rho$ ), pressure ( $p$ ), and the equation of state parameter ( $\omega$ ). These parameters are essential in understanding the behavior of our model and the physical processes it encapsulates. The temporal or redshift evolution of these parameters serves as a window into the dynamic nature of our model, shedding light on various aspects of the Universe's evolution, particularly during its late stages and with regard to stability considerations.

Our discussion in this section serves as the bedrock upon which a deeper comprehension of our model is built. It lays the essential groundwork for subsequent analyses and discussions. To initiate our exploration, we leverage the field equations, denoted as Eqs. (11) and (12). These equations provide us with a means to express the energy density ( $\rho$ ) and pressure ( $p$ ) in specific mathematical forms. This mathematical representation is pivotal in uncovering the intricate details of our model's physical behavior and will serve as the basis for our further investigations and interpretations. The explicit expressions for  $\rho$  and  $p$  using Eqs. (22a) and (23) are found in this case as;

$$\rho(t) = \frac{-n\mu[\nu(m - 4t^3) + 3n(8\pi + \nu)]}{2(-mt + t^4)^2(4\pi + \nu)(8\pi + \nu)} \quad (24)$$

and

$$p(t) = \frac{n\mu[8\pi(-2m + 3n + 8t^3) + 3\nu(-m + n + 4t^3)]}{2(-mt + t^4)^2(4\pi + \nu)(8\pi + \nu)} \quad (25)$$

The concept of the equation of state parameter denoted as  $\omega$ , holds significant importance in the realm of cosmology. It serves as a fundamental tool for characterizing the relationship between pressure ( $p$ ) and energy density ( $\rho$ ) within a given system. This parameter is defined as the ratio of pressure to energy density, and it plays a crucial role in describing the thermodynamic behavior of fluids within the Universe. The value of  $\omega$  offers critical insights into how a substance or component reacts to variations in volume or temperature. These insights, in turn, have profound implications for the overall dynamics and fate of the Universe. In this context, we have already derived expressions for both energy density and pressure specific to our model.

Now, by employing Eqs. (24) and (25), we can deduce the expression for the equation of state parameter  $\omega$ , as follows:

$$\omega(t) = \frac{8\pi(2m - 3n - 8t^3) + 3\nu(m - n - 4t^3)}{\nu(m - 4t^3) + 3n(8\pi + \nu)} \quad (26)$$

## V. DERIVING COSMOLOGICAL PARAMETERS IN TERMS OF REDSHIFT

To effectively constrain our model parameters, we need to express all the obtained cosmological variables in terms of the redshift parameter, denoted as  $z$ . To achieve this, we utilize the relationship between the redshift ( $z$ ) and the scale factor ( $a$ ), given by

$$1 + z = \frac{a_0}{a}, \quad (27)$$

where  $a_0$  is the present value of the scale factor and generally normalized to be  $a_0 = 1$ . The  $t - z$  relationship in our case can be established as;

$$t = \sqrt[3]{k_1} \{1 + [\zeta(1 + z)]^{3\frac{m}{n}}\}^{-\frac{1}{3}} \quad (28)$$

Now, the Hubble parameter in terms of redshift  $z$  for the Model is,

$$H(z) = H_0 (1 + \zeta^{3\eta})^{-\frac{4}{3}} (1 + z)^{-3\eta} \{1 + [\zeta(1 + z)]^{3\eta}\}^{\frac{4}{3}}, \quad (29)$$

where  $\frac{m}{n} = \eta$ . Now, the deceleration parameter, energy density, pressure, and the equation of state parameter can be rewritten in terms of redshift  $z$  as follows;

$$q(z) = -1 + \eta - \frac{4\eta}{\{1 + [\zeta(1 + z)]^{3\eta}\}}, \quad (30)$$

$$p(z) = -\frac{\mu H_0^2 (1 + (\zeta(1 + z))^{3\eta})^{5/3} (16\pi + 3\nu)\eta (-3 + (\zeta(1 + z))^{3\eta}) - 3(8\pi + \nu)(1 + (\zeta(1 + z))^{3\eta})}{2(4\pi + \nu)(8\pi + \nu)(1 + \zeta^2(1 + z)^{3\eta})^{8/3}} \quad (31)$$

$$\rho(z) = -\frac{\left((1 + z)^{-6\eta}\mu H_0^2 \left(1 + ((1 + z)\zeta)^{3\eta}\right)^{5/3} \left(3(8\pi + \nu - \nu\eta) + (3(8\pi + \nu) + \nu\eta)((1 + z)\zeta)^{3\eta}\right)\right)}{2(4\pi + \nu)(8\pi + \nu)(1 + \zeta^3\eta^8)^{1/3}} \quad (32)$$

$$\omega(z) = \frac{(16\pi + 3\nu)\eta (-3 + ((1 + z)\zeta)^{3\eta}) - 3(8\pi + \nu) (1 + ((1 + z)\zeta)^{3\eta})}{3(8\pi + \nu - \nu\eta) + 3(8\pi + \nu) + \nu\eta((1 + z)\zeta)^{3\eta}} \quad (33)$$

To comprehensively explore the evolution of both geometric and physical parameters, as well as to rigorously assess the validity of our derived model, it becomes im-

perative to acquire precise values for the model parameters that play a pivotal role in our theoretical framework. Thus, we perform the data analysis in the subsequent sec-

tion to derive some precise values of the model parameter. Through this data analysis, we aim to gain a deeper understanding of the dynamics governing our cosmological model and evaluate its consistency with observation too.

## VI. DATA ANALYSIS

In this section, we conduct an extensive comparison between our proposed cosmological model and a wide range of available cosmological data. Our aim is to understand the fundamental characteristics of the model through various datasets, including the Cosmic Chronometers (CC), type Ia supernovae (SNIa), Gamma Ray Bursts (GRBs), Quasars (Q), Baryon Acoustic Oscillation (BAO) and Cosmic Microwave Background (CMB). This rigorous investigation aims to determine the optimal values of key model parameters, such as  $\eta$ ,  $\zeta$ . These datasets together describe how our model behaves in the context of the  $f(Q)$  gravity framework. Importantly, we also account for the present-day Hubble function, denoted as  $H_0$ , which plays a crucial role in shaping our results. To obtain best fit values our cosmological model we employ a robust Bayesian statistical methodology. This method relies on likelihood functions and a well-accepted technique called Markov Chain Monte Carlo (MCMC). Within this Bayesian framework, we construct a probabilistic assessment of how likely certain combinations of model parameters are based on actual observations. Our investigation reveals hidden aspects of our cosmological model, offering valuable insights into how it deeply connects with the observable universe.

### A. Methodology

Constraining the Hubble function through observational data involves a process known as parameter estimation or model fitting. In our case, our goal is to determine the optimal values of key model parameters, using various datasets, including Cosmic Chronometers (CC), Type Ia Supernovae (SNIa), Gamma-Ray Bursts (GRBs), Quasars (Q), Baryon Acoustic Oscillations (BAO), and the Cosmic Microwave Background (CMB). The initial step is to establish a likelihood function that measures the agreement between our model predictions and the observed data.

$$\mathcal{L}(\theta) = \exp \left( -\frac{1}{2} \sum_{i=1}^N \frac{(O_i - M_i(\theta))^2}{\sigma_i^2} \right) \quad (34)$$

Where  $O_i$  is the observed data point for the  $i$ th data point,  $M_i(\theta)$  is the model prediction for the  $i$ th data point based on the parameters  $\theta$ , and  $\sigma_i$  represents the uncertainty associated with the observed data point. Next, we perform Bayesian parameter estimation [89]. This requires defining prior distributions for each parameters we want to constrain, which should encapsulate any existing

knowledge or constraints. If strong prior information is lacking, relatively flat priors can be used. The posterior distribution, proportional to the likelihood function multiplied by the prior distribution, is then computed as:

$$P(\theta|D) \propto \mathcal{L}(\theta) \times \pi(\theta) \quad (35)$$

In the subsequent steps, Markov Chain Monte Carlo (MCMC) is employed to explore the posterior distribution and derive parameter constraints [90]. This widely employed sampling technique generates an extensive set of samples from the posterior distribution, allowing for the extraction of key statistical measures. These measures include mean, median, standard deviation, and credible intervals for each parameter, providing optimal parameter values and their associated uncertainties. Subsequently, we rigorously evaluate our model's performance against the Cosmic Chronometers (CC) and type Ia supernovae (SNIa) datasets. Visual comparisons and quantitative metrics, such as chi-squared values or the Akaike Information Criterion (AIC), are used to assess the model's fit. The ensuing discussions delve into the implications of the derived parameter constraints, emphasizing the model's alignment with observational data. We explore potential physical interpretations and ramifications, enriching our understanding of the underlying cosmological dynamics.

### B. Data Discription

#### 1. Cosmic Chronometers

In our analysis, we utilize thirty-one data points acquired through the cosmic chronometers (CC) technique for the determination of the Hubble parameter. This approach allows us to directly extract information about the Hubble function at various redshifts, extending up to  $z \lesssim 2$ . The selection of CC data is motivated by its reliability, as it primarily involves measurements of the age difference between two passively evolving galaxies that originated at the same time but have a slight separation in redshift. This technique enables us to compute  $\Delta z / \Delta t$ , making CC data preferable to methods based on absolute age determinations for galaxies [91]. Our chosen CC data points were sourced from independent [92–98]. Importantly, these references are not influenced by the Cepheid distance scale or any specific cosmological model. Nevertheless, it's worth noting that they do depend on the modeling of stellar ages, which is established using robust stellar population synthesis techniques (for more details, see [94, 96, 99–102] for analyses related to CC systematics). We evaluate the goodness of fit using the  $\chi^2_H$  estimator, which is expressed as follows:

$$\chi^2_{CC}(\Theta) = \sum_{i=1}^{31} \frac{(H(z_i, \Theta) - H_{\text{obs}}(z_i))^2}{\sigma_H^2(z_i)}, \quad (36)$$



Here,  $H(z_i, \Theta)$  represents the theoretical Hubble parameter values at redshift  $z_i$  with model parameters denoted as  $\Theta$ . The observational data for the Hubble parameter at  $z_i$  is given by  $H_{\text{obs}}(z_i)$ , with an associated observational error of  $\sigma_H(z_i)$ .

## 2. type Ia supernovae (SNIa)

Over the years, a multitude of supernova datasets has been established, including references such as [103–107]. Recently, a refreshed version of the Pantheon dataset, referred to as Pantheon+, has been introduced [108]. This updated compilation comprises 1701 data points of type Ia supernovae (SNIa), spanning the redshift interval  $0.001 < z < 2.3$ . SNIa observations have played a pivotal role in unveiling the phenomenon of the universe’s accelerating expansion. These observations serve as crucial tools for investigating the nature of the driving component behind this expansion, owing to SNIa’s status as luminous astrophysical objects. These objects, often treated as standard candles, enable the measurement of relative distances based on their intrinsic brightness. The Pantheon+ dataset stands as a valuable resource, offering insights into the accelerating universe’s characteristics. The chi-square statistic serves as a fundamental tool for comparing theoretical models with observational data. In the context of the Pantheon+ dataset, chi-square values are computed using the subsequent equation

$$\chi_{\text{Pantheon+}}^2 = \vec{D}^T \cdot \mathbf{C}_{\text{Pantheon+}}^{-1} \cdot \vec{D} \quad (37)$$

Here,  $\vec{D}$  represents the difference between the observed apparent magnitudes  $m_{Bi}$  of SNIa and the expected magnitudes given by the cosmological model.  $M$  represents the absolute magnitude of SNIa, and  $\mu_{\text{model}}$  is the corresponding distance modulus predicted by the assumed cosmological model. The term  $\mathbf{C}_{\text{Pantheon+}}$  denotes the covariance matrix provided with the Pantheon+ data, which includes both statistical and systematic uncertainties. The distance modulus is a measure of the distance to an object, defined as:

$$\mu_{\text{model}}(z_i) = 5 \log_{10} \left( \frac{D_L(z_i)}{(H_0/c) \text{ Mpc}} \right) + 25 \quad (38)$$

Here,  $D_L(z)$  represents the luminosity distance, which is calculated for a flat homogeneous and isotropic FLRW universe as:

$$D_L(z) = (1+z)H_0 \int_0^z \frac{dz'}{H(z')} \quad (39)$$

The Pantheon+ dataset differs from the previous Pantheon sample as it breaks the degeneracy between the absolute magnitude  $M$  and the Hubble constant  $H_0$ . This is achieved by rewriting the vector  $\vec{D}$  in terms of the distance moduli of SNIa in the Cepheid hosts. The distance moduli in the Cepheid hosts, denoted as  $\mu_i^{\text{Ceph}}$ ,

are measured independently using Cepheid calibrators. This allows for the independent constraint of the absolute magnitude  $M$ . The modified vector  $\vec{D}'$  is defined as:

$$\vec{D}'_i = \begin{cases} m_{Bi} - M - \mu_i^{\text{Ceph}} & \text{if } i \text{ is in Cepheid hosts} \\ m_{Bi} - M - \mu_{\text{model}}(z_i) & \text{otherwise} \end{cases} \quad (40)$$

With this modification, the chi-square equation for the Pantheon+ dataset can be rewritten as:

$$\chi_{\text{SN}}^2 = \vec{D}'^T \cdot \mathbf{C}_{\text{Pantheon+}}^{-1} \cdot \vec{D}' \quad (41)$$

This revised formulation allows for improved constraints on the absolute magnitude  $M$  and the cosmological parameters.

We’ve also expanded our investigation to include a subset of 162 Gamma Ray Bursts (GRBs) [109], spanning a redshift range of  $1.44 < z < 8.1$ . In this context, we define the  $\chi^2$  function as:

$$\chi_{\text{GRB}}^2(\phi_g^\nu) = \mu_g \mathbf{C}_{g,\text{cov}}^{-1} \mu_g^T, \quad (42)$$

Here,  $\mu_g$  denotes the vector encapsulating the differences between the observed and theoretical distance moduli for each individual GRB. Similarly, for our examination of 24 compact radio quasar observations [110], spanning redshifts in the range of  $0.46 \leq z \leq 2.76$ , we establish the  $\chi^2$  function as:

$$\chi_{\text{Q}}^2(\phi_q^\nu) = \mu_q \mathbf{C}_{q,\text{cov}}^{-1} \mu_q^T, \quad (43)$$

In this context,  $\mu_q$  represents the vector capturing the disparities between the observed and theoretical distance moduli for each quasar.

## 3. Baryon Acoustic Oscillations

To study Baryon Acoustic Oscillations (BAO), we utilize a dataset consisting of 333 measurements [111–122]. However, considering the potential error arising from data correlations, we select a smaller dataset of 17 BAO measurements for our analysis (please see table 1 of this work [123]). This selection helps to reduce errors and improve the accuracy of our results. One of the key measurements obtained from BAO studies in the transverse direction is the quantity  $D_H(z)/r_d$ , where  $D_H(z)$  represents the comoving angular diameter distance. It is related to the following expression [124, 125]:

$$D_M = \frac{c}{H_0} S_k \left( \int_0^z \frac{dz'}{E(z')} \right), \quad (44)$$

where  $S_k(x)$  is defined as:

$$S_k(x) = \begin{cases} \frac{1}{\sqrt{\Omega_k}} \sinh(\sqrt{\Omega_k}x) & \text{if } \Omega_k > 0 \\ x & \text{if } \Omega_k = 0 \\ \frac{1}{\sqrt{-\Omega_k}} \sin(\sqrt{-\Omega_k}x) & \text{if } \Omega_k < 0. \end{cases} \quad (45)$$

Additionally, we consider the angular diameter distance  $D_A = D_M/(1+z)$  and the quantity  $D_V(z)/r_d$ . The latter is a combination of the coordinates of the BAO peak and  $r_d$ , representing the sound horizon at the drag epoch. Furthermore, we can directly obtain "line-of-sight" or "radial" observations from the Hubble parameter using the expression:

$$D_V(z) \equiv [zD_H(z)D_M^2(z)]^{1/3}. \quad (46)$$

By studying these BAO measurements, we gain insights into the cosmological properties and evolution of the universe, while minimizing potential errors and considering relevant distance measures and observational parameters.

#### 4. Cosmic Microwave Background

The CMB distant prior measurements are taken [126]. The distance priors offer useful details about the CMB power spectrum in two ways: the acoustic scale  $l_A$  characterizes the CMB temperature power spectrum in the transverse direction, causing the peak spacing to vary, and the "shift parameter"  $R$  influences the CMB temperature spectrum along the line-of-sight direction, affecting the peak heights, which are defined as follows:

$$l_A = (1+z_d) \frac{\pi D_A(z)}{r_s}, \quad (47)$$

$$R(z) = \frac{\sqrt{\Omega_m} H_0}{c} (1+z_d) D_A(z) \quad (48)$$

The observables that [126] reports are:  $R_z = 1.7502 \pm 0.0046$ ,  $l_A = 301.471 \pm 0.09$ ,  $n_s = 0.9649 \pm 0.0043$  and  $r_s$  is an independent parameter, with an associated covariance matrix. (see table I in [126]). The points represent the inflationary observables as well as the CMB epoch expansion rate. In addition to the CMB points, we also take into account other data from the late Universe. The result is a successful test of the model in relation to the data.

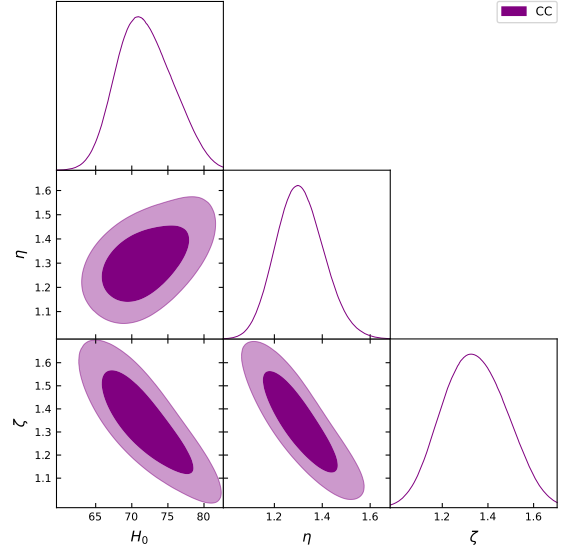


FIG. 1. The above figure shows the MCMC confidence contours at  $1\sigma$  and  $2\sigma$  obtained from the CC dataset.

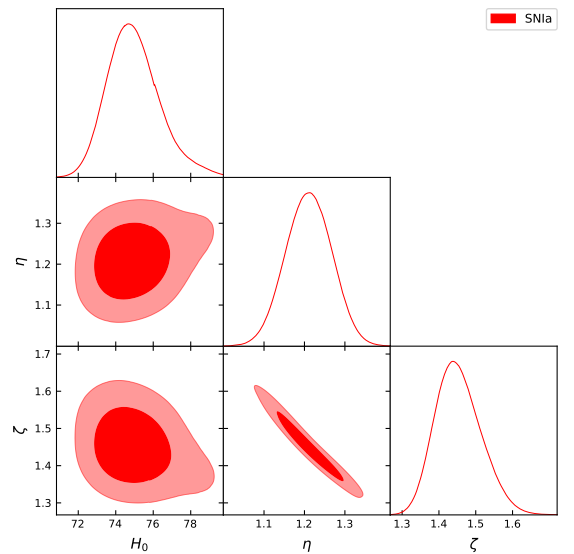


FIG. 2. The above figure shows the MCMC confidence contours at  $1\sigma$  and  $2\sigma$  obtained from the SNIa dataset.



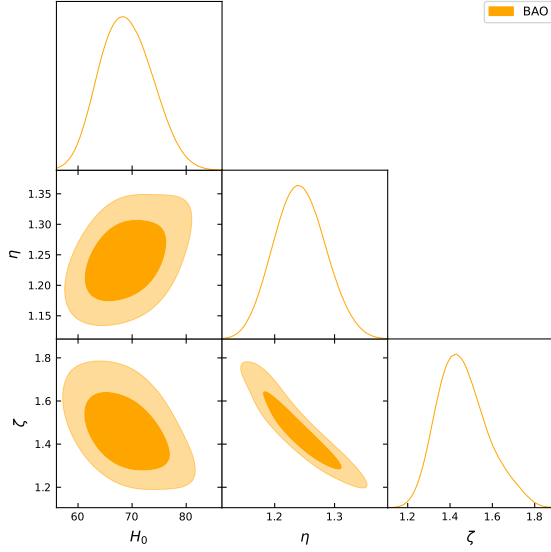


FIG. 3. The above figure shows the MCMC confidence contours at  $1\sigma$  and  $2\sigma$  obtained from BAO dataset.

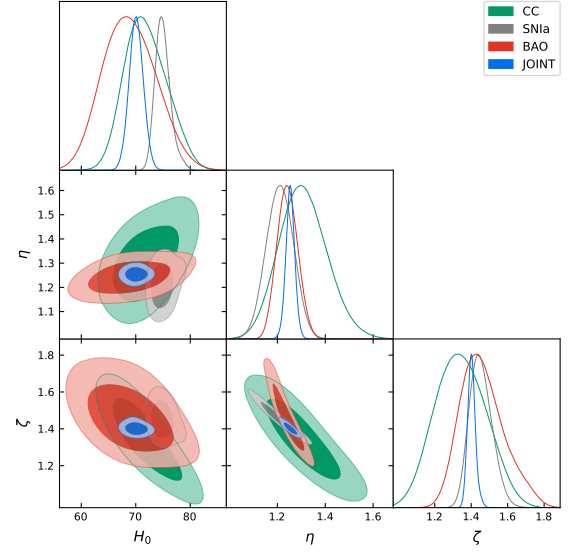


FIG. 5. The above figure shows the MCMC confidence contours at  $1\sigma$  and  $2\sigma$  with respect to All dataset

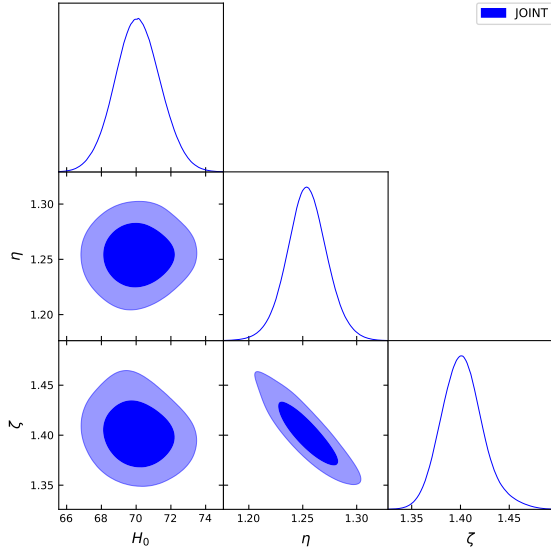


FIG. 4. The above figure shows the MCMC confidence contours at  $1\sigma$  and  $2\sigma$  obtained from the CC + SNIa + GRB + Q + BAO + CMB dataset.

MCMC Results			
Dataset	Parameter	Model	$\Lambda$ CDM Model
CC	$H_0$	$71.819162^{+3.670608}_{-5.331030}$	$68.080661^{+2.121480}_{-2.121480}$
	$\eta$	$1.306004^{+0.092016}_{-0.171492}$	
	$\zeta$	$1.341120^{+0.142478}_{-0.216258}$	
SNIa	$H_0$	$74.945259^{+1.405723}_{-2.176923}$	$74.174328^{+1.267662}_{-1.267662}$
	$\eta$	$1.210586^{+0.052141}_{-0.099643}$	
	$\zeta$	$1.451716^{+0.065137}_{-0.088348}$	
BAO	$H_0$	$68.882349^{+4.806346}_{-7.794739}$	$69.089209^{+4.396874}_{-4.396874}$
	$\eta$	$1.241916^{+0.041110}_{-0.077956}$	
	$\zeta$	$1.459481^{+0.116204}_{-0.202248}$	
Joint	$H_0$	$70.090386^{+1.244084}_{-2.612879}$	$69.854848^{+1.259100}_{-1.259100}$
	$\eta$	$1.253896^{+0.017497}_{-0.037139}$	
	$\zeta$	$1.401909^{+0.021304}_{-0.039846}$	

TABLE I. Summary of MCMC Results obtained in the article.

The contour plots derived from various datasets, including the CC dataset, BAO dataset, SNIa dataset, and the combined dataset (CC + SNIa + GRB + Q + BAO + CMB), are presented in Figure 1, Figure 2, Figure 3, and Figure 4, respectively. To provide a comprehensive overview, we have overlaid all four figures from these diverse datasets in Figure 5. In Table I, we have tabulated the best-fit values of the model parameters  $\eta$  and  $\zeta$ , along with the present-day Hubble function  $H_0$ , complete with their corresponding error bars.

## VII. OBSERVATIONAL AND THEORETICAL COMPARISONS OF THE HUBBLE FUNCTION AND DISTANCE MODULUS FUNCTION

After obtaining the best-fit values for our cosmological model parameters, it's crucial to compare our model with the widely accepted  $\Lambda$ CDM model. The  $\Lambda$ CDM model has consistently aligned with various observational datasets and stands as a robust framework for understanding the Universe's evolution. This comparative analysis deepens our comprehension of differences between the two models, shedding light on the implications of these disparities in cosmology. By examining deviations between our model and the  $\Lambda$ CDM model, we pinpoint the specific features that distinguish our parametrized model, such as the Universe's dynamics. This exploration provides valuable insights into our model's strengths and limitations, enriching our understanding of the cosmos. This comparison with the  $\Lambda$ CDM model acts as a benchmark, allowing us to assess the goodness-of-fit to observational data and gauge the alignment between our parametrized model and the established  $\Lambda$ CDM framework.

### A. Comparison with the CC data points

We evaluate our model's compatibility with observational data by conducting a comparative analysis with the Cosmic Chronometers dataset, consisting of 31 data points represented by orange dots, each accompanied by error bars represented by purple dots. This comparison is illustrated in Fig 6, where our model's predictions are depicted by the blue line. To provide a reference point, we also include the well established  $\Lambda$ CDM model, represented by the black line, with cosmological parameters  $\Omega_{m0} = 0.3$  and  $\Omega_{\Lambda} = 0.7$ . The results of this analysis reveal a remarkable correspondence between our model's predictions and the observed data points, as indicated by their close alignment. This alignment underscores the ability of our model to effectively capture the inherent features and trends within the Cosmic Chronometers dataset. Consequently, our model demonstrates its capability to reproduce the expansion history of the universe, as inferred from the Hubble data.

### B. Comparison with the type Ia supernova dataset

In this analysis, we carefully studied the  $\mu(z)$  distance modulus function for our Model alongside the data from type Ia supernovae, which includes a significant 1701 data points. We also compared our Model with standard  $\Lambda$ CDM Model. The results of this comparison are represented in Fig 7. The figures clearly illustrate that our model and the  $\Lambda$ CDM model exhibit

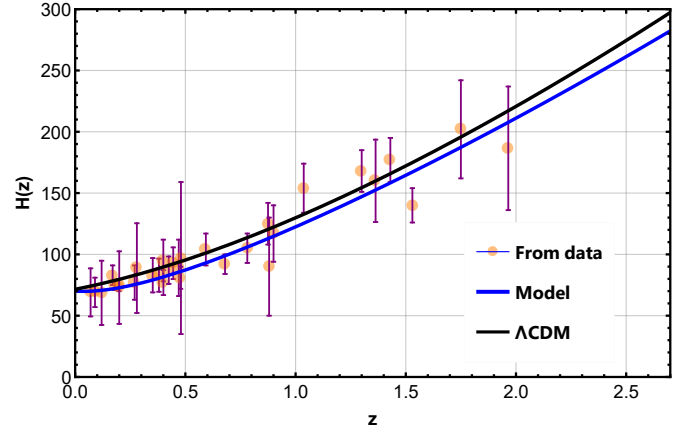


FIG. 6. Comparative analysis of our model ( Blue Line ) with 31 CC measurements ( Orange dots ) and  $\Lambda$ CDM model ( black line ).

a commendable level of agreement with the type Ia supernova dataset. This agreement indicates that these models effectively capture and replicate the observed distance measurements, signifying their consistency with empirical astronomical data.

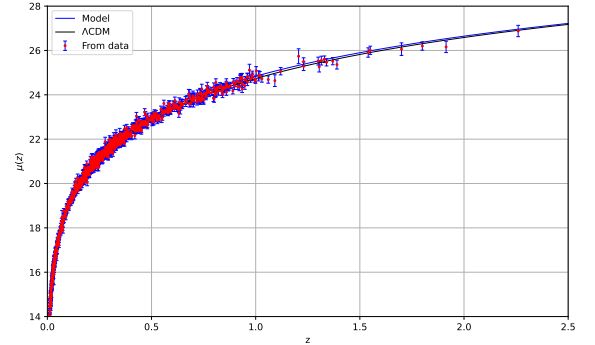


FIG. 7. Comparative analysis of our model ( Blue Line ) with 1701 type Ia supernova measurements ( Orange dots ) and  $\Lambda$ CDM model ( black line ).

### C. Relative difference between model and $\Lambda$ CDM

The Figure 8 illustrates the relative difference between our Model and the standard  $\Lambda$ CDM model. It is evident from the Figure that our Model exhibits distinct behavior from the typical  $\Lambda$ CDM model, particularly noticeable start from low redshift values ( $z > 0$ ). As the redshift increases, these disparities between our model and the  $\Lambda$ CDM Model become more pronounced.

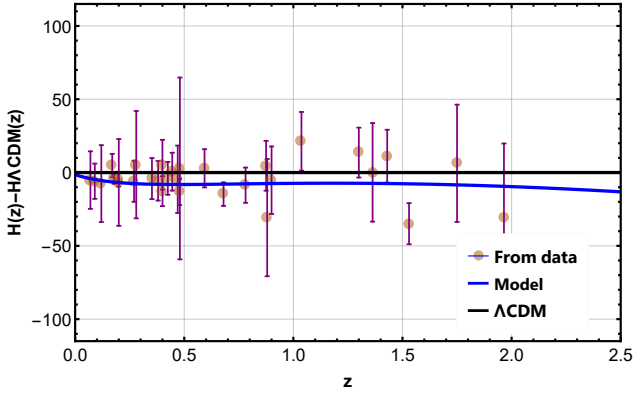


FIG. 8. Comparative analysis of our model ( Blue Line ) with 31 CC measurements ( Orange dots ) and  $\Lambda$ CDM model ( black line ).

## VIII. COSMOGRAPHY PARAMETERS

### A. The deceleration parameter

The deceleration parameter [127], denoted as " $q$ ," is a fundamental cosmological parameter used in the study of the expansion dynamics of the universe. It plays a crucial role in understanding the past and future evolution of the cosmos. Introduced by Edwin Hubble in the early 20th century, this parameter provides insights into whether the expansion of the universe is slowing down or accelerating. The deceleration parameter is defined in terms of the second derivative of the scale factor of the universe, which describes how the universe's size changes with time. Mathematically, it can be expressed as:

$$q = -\frac{a\ddot{a}}{\dot{a}^2}, \quad (49)$$

where  $q$  is the deceleration parameter.  $a(t)$  represents the scale factor of the universe as a function of time.  $\dot{a}$  represents the first derivative of the scale factor with respect to time.  $\ddot{a}$  represents the second derivative of the scale factor with respect to time. The value of the deceleration parameter is indicative of the nature of the cosmic expansion:

1.  $q > 0$  (Decelerating Universe): If the deceleration parameter is positive, it implies that the expansion of the universe is slowing down over time. In the past, this was the prevailing belief when the gravitational attraction of matter was thought to dominate the cosmic dynamics.
2.  $q = 0$  (Critical Universe): A deceleration parameter of zero suggests that the expansion is proceeding at a constant rate. In this scenario, the universe's expansion is neither accelerating nor decelerating, often referred to as a "critical universe."
3.  $q < 0$  (Accelerating Universe): A negative value for the deceleration parameter signifies that the uni-

verse's expansion is accelerating. This phenomenon gained significant attention in the late 20th century when the discovery of dark energy provided a compelling explanation for this observed acceleration.

In recent years, the study of the deceleration parameter has become increasingly important in cosmology, particularly in the context of understanding dark energy and the fate of the universe. It is a key parameter used in observational cosmology to probe the nature of the components of the universe, such as dark matter and dark energy, and to determine the overall geometry of the universe.

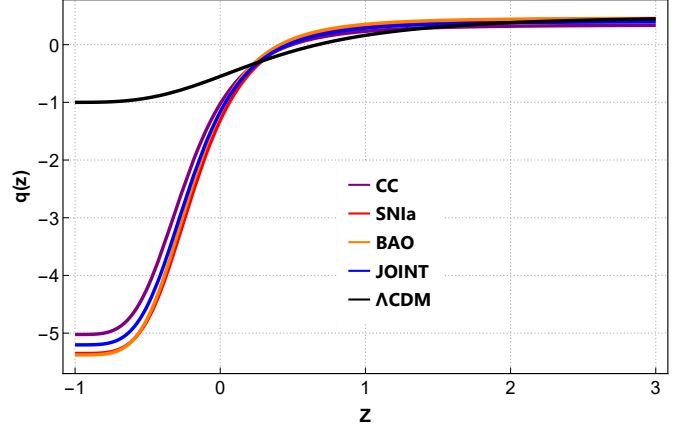


FIG. 9. Evolution of deceleration parameter with respect to the redshift.

### B. The jerk parameter

In the realm of cosmology, understanding the dynamics of the universe's expansion is of paramount importance. While the Hubble constant and the deceleration parameter have been essential tools in characterizing this expansion, a more nuanced parameter known as the "jerk parameter" has emerged as a valuable addition to our cosmological toolkit. The jerk parameter, denoted as " $j$ ," provides a deeper level of insight into the cosmic acceleration, complementing the information offered by the deceleration parameter [128]. The jerk parameter represents the third time derivative of the scale factor of the universe, building upon the concepts embodied by the Hubble parameter and the deceleration parameter. Mathematically, it can be expressed as:

$$j = \frac{1}{a} \frac{d^3 a}{d\tau^3} \left[ \frac{1}{a} \frac{da}{d\tau} \right]^{-3} = q(2q + 1) + (1 + z) \frac{dq}{dz}, \quad (50)$$

where:  $j$  is the jerk parameter,  $a(t)$  represents the scale factor of the universe as a function of time,  $\dot{a}$  represents the first derivative of the scale factor,  $\ddot{a}$  represents the second derivative of the scale factor,  $\dddot{a}$  represents the third derivative of the scale factor,  $z$  represents redshift.

In the context of the Taylor expansion of the scale factor around a reference time  $t_0$ , the expansion takes the form:

$$\frac{a(t)}{a_0} = 1 + H_0(t - t_0) - \frac{1}{2}q_0 H_0^2(t - t_0)^2 + \frac{1}{6}j_0 H_0^3(t - t_0)^3 + O[(t - t_0)^4]. \quad (51)$$

where the subscript 0 denotes current values. Here,  $H_0$  represents the Hubble constant at the reference time  $t_0$ ,  $q_0$  is the deceleration parameter, and  $j_0$  is the jerk parameter at the same reference time. In contemporary cosmology, the jerk parameter has played a crucial role in refining our knowledge of the universe's evolution.

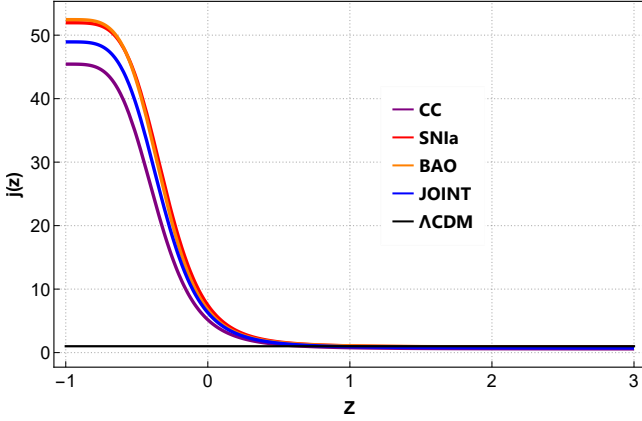


FIG. 10. Evolution of jerk parameter with respect to the redshift.

## IX. STATEFINDER DIAGNOSTIC

In the field of cosmology, the study of the universe's evolution, demands a comprehensive understanding of dark energy (DE) and its influence on cosmic expansion. To analyze this cosmic dynamic without bias toward any specific DE model, cosmologists employ a valuable tool known as the statefinder diagnostic parameter. This mathematical tool, pioneered by researchers [129–132], employs higher derivatives of the cosmic scale factor to characterize the universe's expansion. Its primary purpose is to distinguish and compare different DE models effectively. What sets the statefinder diagnostic apart is its model-independent nature, enabling the exploration of various cosmological scenarios, including those with diverse forms of dark energy. The statefinder diagnostic is encapsulated in a parameter pair, denoted as  $\{r, s\}$ . These parameters are defined as follows:

$$r = \frac{\ddot{a}}{aH^3}, \quad s = \frac{r - 1}{3(q - \frac{1}{2})}. \quad (52)$$

These parameters leverage higher-order derivatives of the scale factor, the Hubble parameter  $H$ , and the deceleration

parameter  $q$  to provide insights into cosmic expansion. Various possibilities in the  $\{r, s\}$  and  $\{q, r\}$  planes are exhibited to depict the temporal evolution of various DE models. With the assistance of the statefinder diagnostics pair. In these cases, some specific pairs typically correlate to classic DE models such as  $\{r, s\} = \{1, 0\}$  represents  $\Lambda$ CDM model and  $\{r, s\} = \{1, 1\}$  indicates standard cold dark matter Model (SCDM) in FLRW background. Also,  $(-\infty, \infty)$  yields static Einstein Universe. In the  $r - s$  plane,  $s > 0$  and  $s < 0$  define a quintessence-like model and phantom-like model of the DE, respectively. Moreover, the evolution from phantom to quintessence can be observed by deviation from  $r, s = 1, 0$ . On the other hand,  $\{q, r\} = \{-1, 1\}$  corresponds to the  $\Lambda$ CDM model while  $\{q, r\} = \{0.5, 1\}$  shows SCDM model. It is important to note that on a  $r - s$  plane if the DE model's trajectories deviate from these standard values, the resulting model differs from the normal cosmic model.

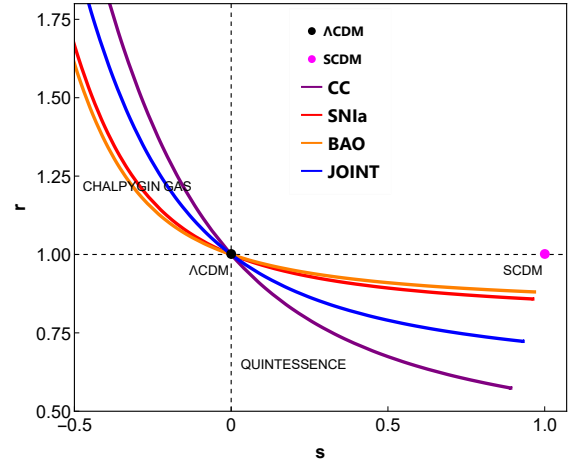


FIG. 11. This figure shows  $\{r, s\}$  Profile.

## X. OM DIAGNOSTIC

In the realm of cosmology, the  $Om(z)$  diagnostic [133–136] parameter plays a pivotal role in unraveling the cosmic mysteries. It's an essential tool for gauging the relative contribution of matter to the universe's total energy density, shedding light on the universe's overall geometry.  $\omega_m$ , the heart of this diagnostic, tells us about the ratio of current matter density to the critical density required for a flat universe. If  $\omega_m$  is less than 1, it hints at an open universe, while a value greater than 1 suggests a closed one. This parameter serves as a geometric yardstick and plays a crucial role in testing the  $\Lambda$ CDM model, which posits three primary components: dark matter, ordinary matter (baryonic), and enigmatic dark energy. One of the remarkable aspects of  $\omega_m$  is its ability to discern different dark energy models from the standard  $\Lambda$ CDM model. By altering the slope of  $Om(z)$ ,

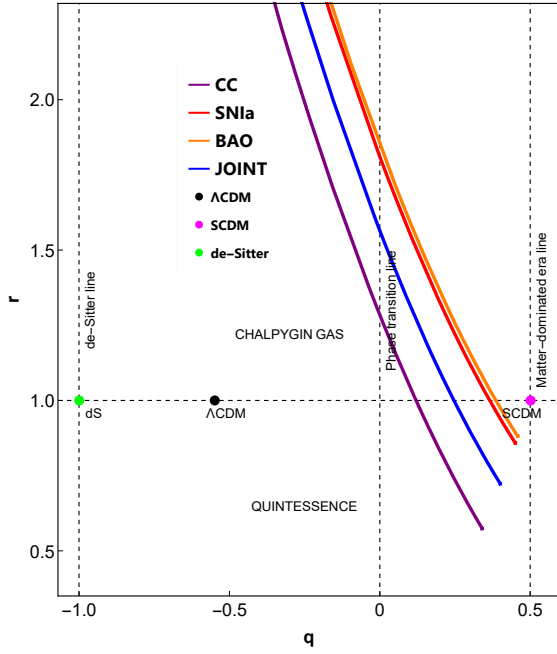


FIG. 12. This figure shows  $\{r, q\}$  profile.

this diagnostic efficiently distinguishes between various models. A positive slope signifies a quintessence model, while a negative slope points to a phantom model. A constant slope, on the other hand, aligns with the cosmological constant. This feature allows us to explore the intricate balance among these cosmic components. Much like the statefinder diagnostic,  $Om(z)$  serves as a powerful tool for understanding cosmic evolution, geometry, and the interplay of matter in shaping the universe's dynamics. In a flat universe,  $Om(z)$  is defined as:

$$Om(z) = \frac{\left(\frac{H(z)}{H_0}\right)^2 - 1}{(1+z)^3 - 1}. \quad (53)$$

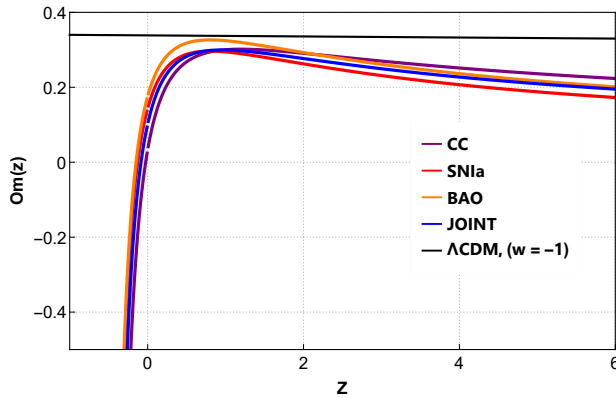


FIG. 13. This figure shows the  $Om(z)$  Profile.

## XI. PHYSICAL PARAMETERS

In cosmology, physical parameters such as pressure, equations of state, and the density parameter (denoted as  $\rho$ ) play crucial roles in describing the behavior and evolution of the universe. Here's a general overview of these concepts along with their associated formulas:

### A. Pressure Density $p$

Pressure density influences the overall behavior of the universe's expansion. The pressure associated with cosmic components, such as matter, radiation, and dark energy, affects how they contribute to the expansion rate. Positive pressure (like that of matter) tends to slow down the expansion, while negative pressure (like that of dark energy) can accelerate it.

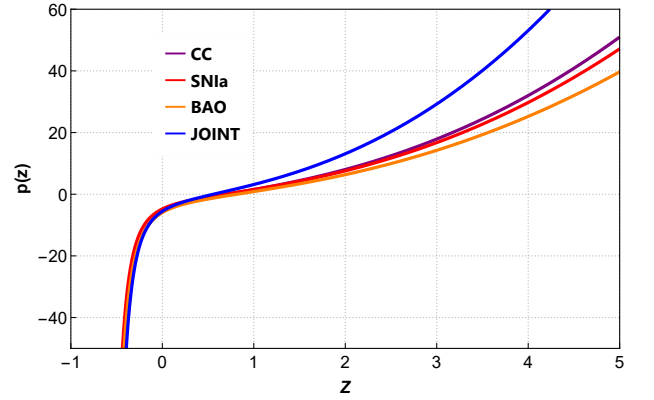


FIG. 14. Pressure density versus redshift for model.

### B. Energy Density $\rho$

Energy density refers to the amount of energy contained within a given volume of space. In cosmology, it quantifies the total energy content of a specific cosmic component. Energy density is a crucial parameter as it determines the gravitational effects of that component on the overall cosmic dynamics.

### C. Equation of State $\omega$

The equation of state in cosmology establishes a relationship between the pressure ( $P$ ) and energy density ( $\rho$ ) of a cosmic component, effectively describing how pressure and density interrelate. In its general form, this equation is expressed as  $P = \omega\rho$ , where " $\omega$ " serves as the parameter characterizing the equation of state. Different values of " $\omega$ " correspond to distinct cosmic components. For instance, when  $w = 0$ , it corresponds

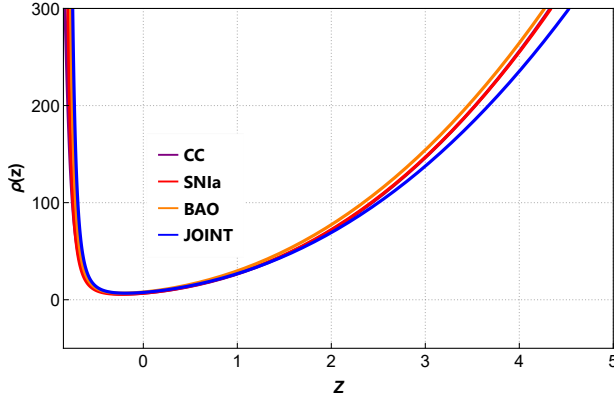


FIG. 15. Pressure density versus redshift for model

to non-relativistic matter (e.g., cold dark matter). A value of  $\omega = 1/3$  corresponds to radiation (e.g., photons), while  $\omega \approx -1$  aligns with dark energy, exhibiting properties akin to a cosmological constant ( $\Lambda$ ). These varying values of " $\omega$ " play a critical role in characterizing the behavior and properties of different cosmic components within the framework of cosmological models.

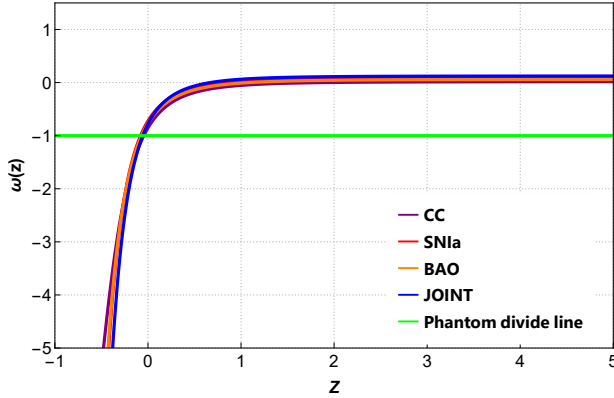


FIG. 16. Evolution of EoS Parameter.

## XII. STATISTICAL ANALYSIS

To determine the optimal model for our analysis, it is crucial to consider the number of free parameters associated with each model, in addition to the  $\chi^2_{\min}$  value obtained. While choosing among the various information criteria available in the literature is not a straightforward task [137], we opt for the most commonly used ones. One of these is the Akaike Information Criterion (AIC) [137–139], defined as:

$$AIC \equiv -2 \ln \mathcal{L}_{\max} + 2p_{\text{tot}} = \chi^2_{\min} + 2p_{\text{tot}} \quad (54)$$

Here,  $p_{\text{tot}}$  represents the total number of free parameters in the specific model, and  $\mathcal{L}_{\max}$  signifies the maximum likelihood of the model being considered. Addition-

ally, we incorporate the Bayesian Information Criterion (BIC), introduced by [137–139], which is defined as:

$$BIC \equiv -2 \ln \mathcal{L}_{\max} + p_{\text{tot}} \ln(N_{\text{tot}}) \quad (55)$$

By utilizing the definitions (54) and (55), we compute the discrepancies  $\Delta AIC$  and  $\Delta BIC$  relative to the  $\Lambda$ CDM model in question. In accordance with [140], if  $0 < |\Delta AIC| \leq 2$ , it implies that the compared models may be viewed as compatible with each other. Conversely, if  $|\Delta AIC| \geq 4$ , it indicates that the model with the higher AIC value is not supported by the data. Similarly, for  $0 < |\Delta BIC| \leq 2$ , the model exhibiting the higher BIC value is marginally less favored by the data. In cases where  $2 < |\Delta BIC| \leq 6$  ( $|\Delta BIC| > 6$ ), the model with the higher BIC values is significantly (highly) less favored. The specific distinctions among the investigated cosmological models are detailed in II.

In cosmology, the terms "P-value" and "L-statistic" are commonly used in statistical analyses to assess the significance of observations and test hypotheses.

- **P-value (Probability Value)** The P-value is a statistical measure that quantifies the evidence against a null hypothesis. It tells you the probability of observing data as extreme or more extreme than what you have, assuming that the null hypothesis is true. In cosmology, P-values are often used in the context of hypothesis testing [141–143]. For example, when analyzing cosmic microwave background (CMB) data, cosmologists may calculate P-values to assess whether the observed CMB power spectrum matches the predictions of a particular cosmological model. A low P-value suggests that the data is inconsistent with the model, while a high P-value suggests that the data is consistent with the model.
- **L-statistic (Likelihood Statistic)** The L-statistic, often referred to as the likelihood ratio [144–146], is a statistical measure used to compare the likelihood of observing data under two different hypotheses. In cosmology, the L-statistic is frequently used when comparing different cosmological models or parameter values. For instance, in Bayesian cosmological analyses, the likelihood function quantifies how well a given model explains the observed data. The L-statistic is used to compare the likelihoods of different models, and models with higher likelihoods are considered more compatible with the data.

In practical terms, when cosmologists perform statistical analyses in cosmology, they often calculate likelihoods and P-values to evaluate the goodness of fit between observational data and theoretical models. These statistical tools help cosmologists make inferences about the parameters of the universe, such as the density of dark matter, dark energy properties, and the geometry of the universe.



Model	$\chi_{\text{tot}}^2{}^{\text{min}}$	$\chi_{\text{red}}^2$	$\mathcal{K}_f$	$AIC_c$	$\Delta AIC_c$	$BIC$	$\Delta BIC$	P-value	L-statistic
$\Lambda$ CDM Model	1107.35	0.9645	3	1113.35	0	1128.83	0	0.712755	254.538726
Model	1108.32	0.9634	3	1114.32	0.97	1129.80	0.97	0.776201	251.087532

TABLE II. Summary of  $\chi_{\text{tot}}^2{}^{\text{min}}$ ,  $\chi_{\text{red}}^2$ ,  $AIC_c$ ,  $\Delta AIC_c$ ,  $BIC$ ,  $\Delta BIC$ , P-value and L-statistic for  $\Lambda$ CDM and  $f(Q, T)$  model.

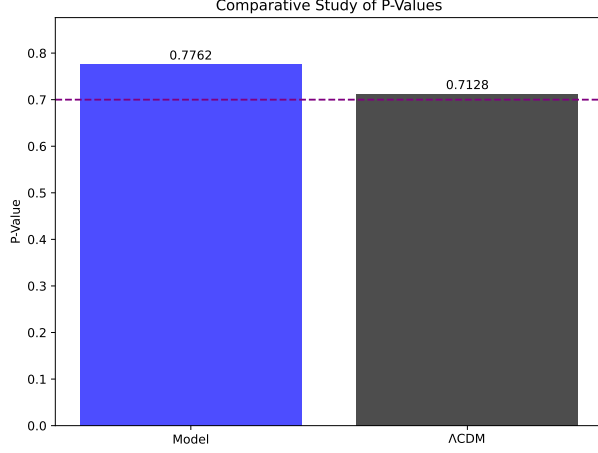


FIG. 17. The comparative Comparative Study of P-Values

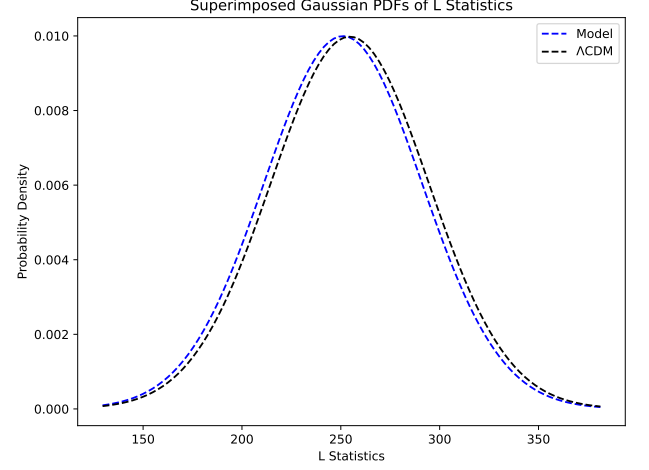


FIG. 19. The superimposed overlays of Gaussian Probability Density Functions (PDFs) on the L statistics for both models.

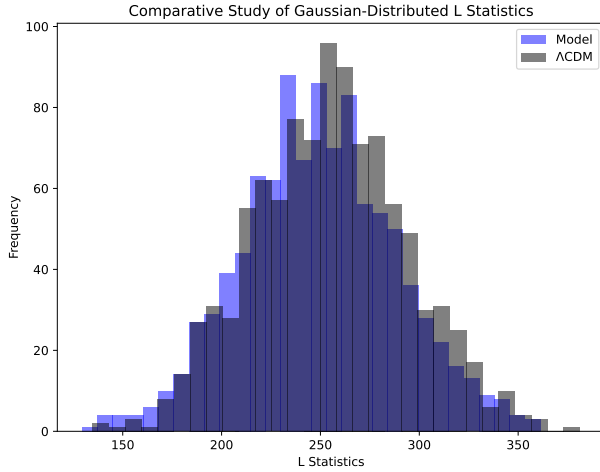


FIG. 18. The comparative histogram plot of the distribution of L statistics for both models

### XIII. RESULTS

*a. deceleration parameter* Fig 9 represents the redshift evolution of the deceleration parameter ( $q$ ) across various cosmological epochs, denoted by the redshift ( $z$ ), as constrained by different datasets, including Cosmic Chronometers (CC), Type Ia Supernovae

(SNIa), Baryon Acoustic Oscillations (BAO), and a Joint Dataset encompassing CC + SNIa + Gamma-Ray Bursts (GRB) + Quasars (Q) + BAO and  $\Lambda$ CDM paradigm. At high redshifts, corresponding to earlier epochs in the universe's history, the model exhibits a noteworthy convergence toward a consistent value of  $q$  across all analyzed datasets, specifically around 0.2763. As the universe evolves and redshift decreases, it undergoes a phase transition from a decelerating phase to an accelerating one. Nonetheless, the long-term behavior of the two models starkly diverges. In the  $\Lambda$ CDM model, the Universe ends in a de Sitter phase characterized by  $q = -1$ . Conversely, in the proposed model, a state of super-accelerated evolution emerges, with  $q(-1) \approx -4.7$  while in the  $\Lambda$ CDM model the Universe ends in a de Sitter phase, with  $q = -1$ , in the purposed model a super-accelerated evolution does occur, with  $q(-1) \approx -4.7$ . An intriguing and noteworthy observation is that the model predicts an acceleration at half-time higher than the  $\Lambda$ CDM model with an approximate value of  $q(0) \approx -1$ , whereas the  $\Lambda$ CDM model predicts a value of approximately  $q(0) \approx -0.5$ . This deceleration parameter maintains consistency across all the datasets employed in our comprehensive analysis.



*b. jerk parameter* Fig 10 illustrates the redshift-dependent behavior of the jerk parameter across the various datasets. At high redshifts, the value of  $j$  in the proposed model is across all datasets and the  $\Lambda$ CDM model consistently converges to approximately 1. However, important differences do exist at lower redshifts, with a very significant difference appearing at  $z = 0$ , where the proposed model predicts a value higher than the  $\Lambda$ CDM value of  $j = 1$  across all datasets. Determining the current value of  $j$  through observations can be a crucial test for the proposed model cosmological model. As we look far into the future, with decreasing redshifts, the jerk parameter consistently approaches a value of about  $j = 48$  across all datasets. This convergence strongly suggests that the proposed model predicts a coherent evolution towards a state of uniform acceleration in the distant cosmic era.

*c.  $\{r, s\}$  profile* The  $\{r, s\}$  profile of our obtained cosmological model, as depicted in Fig 11, reveals a fascinating cosmological evolution. At the initial stages, the model exhibits values in the range  $r < 1$  and  $s > 0$ , which correspond to characteristics akin to the quintessence region. In cosmology, quintessence is a dynamic form of dark energy with a varying energy density and negative pressure, potentially explaining the universe's accelerated expansion. Further by crossing intermediate  $\Lambda$ CDM fixed point  $\{0, 1\}$ , it transitions into a different region where  $r > 1$  and  $s < 0$ , represent Chaplygin gas. Chaplygin gas is a cosmological model featuring exotic matter with unique properties, used to explain cosmic expansion. Remarkably, this  $\{r, s\}$  profile consistently exhibits the same behavior with respect to all the datasets used in our analysis.

*d.  $\{r, q\}$  profile* The  $\{r, q\}$  profile of our derived cosmological model, as shown in Fig 12, reveals a fascinating evolution. Initially, the model exhibits values within the region  $q > 0$  and  $r < 1$ , which resembles characteristics commonly linked with the quintessence. As the Universe evolves, a distinctive transition occurs, leading the model into a different region characterized by  $q < 0$  and  $r > 1$ , represented by the Chaplygin gas scenario. Notably, this  $\{r, q\}$  profile consistently demonstrates the same behavior across all the datasets integrated into our analysis.

*e.  $Om$  Diagnostic* Fig. 13 illustrates how  $Om(z)$  changes as we look at different redshifts ( $z$ ).  $Om(z)$  is always lower than the current matter density parameter,  $\Omega_{m0}$ , indicating that the model falls into the quintessence domain in the early universe. For redshifts below zero  $z < 0$ ,  $Om(z)$  drops rapidly and becomes negative, indicating that the model enters a phase characterized as the "phantom region". The Phantom Dark Energy

Model is a cosmological theory where dark energy, with an equation of state parameter below  $-1$ , drives an accelerating universe and predicts a cataclysmic "Big Rip" end. This pattern remains consistent across all datasets.

*f. Physical Parameters* Figures 14, 15, and 16 provide the necessary illustration into the dynamic evolution of physical parameters, pressure ( $p$ ), energy density ( $\rho$ ), and the equation of state parameter ( $\omega$ ) at different redshifts. These figures meticulously portray how these parameters change as a function of redshift, with each curve representing a distinct set of model parameters derived from diverse datasets. The beauty of these plots lies in their ability to unveil the intricate phases of evolution exhibited by these physical parameters at different points in cosmic history. In particular, Figure 14 unveils the transition in the pressure parameter: it shifts from being positive in the early universe to becoming increasingly negative as time progresses, hinting at the onset of late-time cosmic acceleration. However, what truly sets this observation apart from conventional cosmological models like the Lambda Cold Dark Matter ( $\Lambda$ CDM) model is the intriguing revelation that the pressure doesn't just remain negative; it veers towards extreme negativity in the distant future, even diverging entirely. This divergence suggests the possibility of a cosmic fate like the "Big Rip." Meanwhile, Figure 15 charts the path of energy density ( $\rho$ ), illustrating how it initially diminishes from its higher values in the past to its present, smaller magnitude. However, rather unexpectedly, it then move towards substantial growth, reaching extraordinarily high values in the far-off future before, once again, diverging. This peculiar behavior aligns with the notion of a future singularity akin to the Big Rip, where cosmic densities surge towards infinity. Lastly, Figure 16 serves as a window into the universe's cosmic temperament through the equation of state parameter ( $\omega$ ). Here, we witness the universe's evolution from an early phase of decelerating expansion to the current epoch of accelerating expansion. Yet, it is the futuristic view of  $\omega$  that truly captivates; it plunges into profoundly negative territory, echoing the prospect of a cataclysmic Big Rip-type singularity in the cosmic evolution. In essence, these figures not only describes the evolution of the universe in the past but also show the far future behaviour.

*g. Statistical Analysis* Based on the table II. We conduct a comparative analysis between our proposed model and the  $\Lambda$ CDM Model. We will use several statistical criteria, including  $\chi^2_{\min}$ ,  $\chi^2_{\text{red}}$ ,  $AIC$ ,  $\Delta AIC$ ,  $BIC$ , and  $\Delta BIC$ . Additionally, we will assess the significance of  $\Delta AIC$  and  $\Delta BIC$  in the process of model selection. Our model has a  $\chi^2_{\min}$  value slightly higher than that of the  $\Lambda$ CDM Model, indicating a slightly worse goodness

of fit in terms of minimizing the total  $\chi^2$ . The  $\chi^2_{\text{red}}$  value for our model is also slightly higher, suggesting a slightly worse fit when accounting for the degrees of freedom. The  $AIC$  and  $\Delta AIC$  of our model have a higher  $AIC$  value compared to the  $\Lambda$ CDM Model, indicating that the  $\Lambda$ CDM Model is more favorable in terms of the balance between goodness of fit and model complexity according to  $AIC$ . The calculated  $\Delta AIC$  of 0.97 suggests that our model is less supported compared to the  $\Lambda$ CDM Model. This value falls within the range of models that are less strongly supported according to  $AIC$ . Our Model has a higher  $BIC$  value than the  $\Lambda$ CDM Model, further supporting the notion that the  $\Lambda$ CDM Model is preferred in terms of model complexity and goodness of fit according to  $BIC$ . The calculated  $\Delta BIC$  of 0.97 indicates that the  $\Lambda$ CDM Model is more favored compared to our model. This suggests that the  $\Lambda$ CDM Model is preferred over our model according to  $BIC$ . Fig. 17 shows that both models have relatively high p-values (above 0.7), indicating that the observed data is reasonably consistent with both the  $\Lambda$ CDM model and our Model. For the  $\Lambda$ CDM model, the L-statistic is 254.538726. For our Model, the L-statistic is 251.087532. In this case, the  $\Lambda$ CDM model has a slightly higher L-statistic compared to Our Model (254.538726 vs. 251.087532). This indicates that the  $\Lambda$ CDM model may provide a slightly better fit to the data. Fig. 18 provides the comparative histogram distribution of L statistics for  $\Lambda$ CDM model and our Model. The  $x$ -axis represents the L statistics values, while the  $y$ -axis represents the frequency or count of data points within each bin. You can observe how the L statistics are distributed for both models. Look for differences in the central tendencies (peaks) and spreads (widths) of the distributions. Since our model's histogram is similar to the  $\Lambda$ CDM histogram, it suggests that our model is a reasonable fit for the data, Fig. 19 provides the superimposed plot overlays Gaussian Probability Density Functions (PDFs) on the L statistics for both models. The  $x$ -axis represents the L statistics values, and the  $y$ -axis represents the probability density. Again the Gaussian PDF of our model is closely aligned with the histogram of L statistics, it suggests a better match between the  $\Lambda$ CDM model and the data.

#### XIV. CONCLUSION

In this study, we conducted a comprehensive and rigorous examination of the  $f(Q, T)$  cosmological model. We extend the analysis conducted in classical gravity by incorporating a parametrization of the deceleration parameter, as outlined in the reference [37]. Our investigation encompasses the realm of  $f(Q, T)$  gravity, allowing us to explore its implications and effects on the parametrization. This model was compared against a range of cosmological observations, encompassing 31 Cosmic Chronometers, 1071 type Ia supernovae measurements, 162 Gamma Ray Bursts (GRBs), 24

measurements from compact radio quasars, 17 Baryon Acoustic Oscillation (BAO) measurements, and CMB distant prior. To determine the best-fit value for the model parameters, we employed the Markov Chain Monte Carlo (MCMC) methodology, allowing us to derive the optimal fit for these parameters. Subsequently, utilizing these best-fit values, the data-fitting process yielded exceptionally accurate outcomes for both the CC and SNIa datasets. We observed the redshift evolution of the deceleration parameter ( $q$ ) across different cosmological epochs using various datasets. At high redshifts, the models converged to a consistent value of  $q \approx 0.2763$ . However, as the universe evolved, significant differences emerged. In the  $\Lambda$ CDM model, the universe ended in a de Sitter phase ( $q = -1$ ), while our proposed model exhibited super accelerated evolution ( $q \approx -4.7$ ), leading to intriguing differences in the predicted acceleration at present and in the distant future. We examined the redshift dependent behavior of the jerk parameter across datasets. Both the proposed model and the  $\Lambda$ CDM model consistently converged to  $j \approx 1$  at high redshifts. However, a significant difference was observed at the current redshift ( $z = 0$ ), where our model predicted a value higher than the  $\Lambda$ CDM value of  $j = 1$ . This presents a crucial test for the proposed model. The analysis of the  $\{r, s\}$  profile revealed a fascinating cosmological evolution. Our model exhibited characteristics resembling quintessence and Chaplygin gas phases, maintaining consistency across all datasets. The  $\{r, q\}$  profile depicted a dynamic evolution, transitioning from quintessence-like behavior to a Chaplygin gas scenario. This behavior remained consistent across all datasets. The  $Om(z)$  analysis indicated that our model fell into the quintessence domain in the early universe and entered a "phantom" region at redshifts below zero, consistent with the behavior of a Phantom Dark Energy Model. The analysis of pressure ( $p$ ), energy density ( $\rho$ ), and the equation of state parameter ( $\omega$ ) across varying redshifts reveals deviations from standard cosmological models like the  $\Lambda$ CDM model. These deviations hint at a unique cosmic fate marked by unprecedented negativity and divergence a phenomenon evocative of the enigmatic "Big Rip" singularity. The comparative analysis between our proposed model and the  $\Lambda$ CDM Model, as indicated by various statistical criteria including  $\chi^2_{\text{min}}$ ,  $\chi^2_{\text{red}}$ ,  $AIC$ ,  $\Delta AIC$ ,  $BIC$ , and  $\Delta BIC$ , reveals important insights. While our model exhibits slightly higher  $\chi^2_{\text{min}}$  and  $\chi^2_{\text{red}}$  values, suggesting a slightly worse goodness of fit and fit-adjusted for degrees of freedom, respectively, the  $\Lambda$ CDM Model stands as the more favorable choice in terms of both  $AIC$  and  $BIC$ .  $AIC$  and  $BIC$  differences indicate stronger support for the  $\Lambda$ CDM Model. However, both models display reasonably high p-values, implying compatibility with observed data. Although our model closely aligns with data distribution, the  $\Lambda$ CDM Model marginally outperforms it, as evident from L-statistics and PDF overlays. In conclusion, while our proposed model

provides intriguing insights and predicts unique cosmic behaviors, such as super acceleration, the standard  $\Lambda$ CDM Model remains favored by statistical criteria. The differences observed in various cosmological parameters and profiles underscore the complexity of

cosmological models and the need for ongoing research to refine our understanding of the universe's evolution. Further observations and refined datasets may shed more light on these intriguing cosmological phenomena.

- 
- [1] Saul Perlmutter, Goldhaber Aldering, Gerson Goldhaber, RA Knop, Peter Nugent, Patricia G Castro, Susana Deustua, Sebastien Fabbro, Ariel Goobar, Donald E Groom, et al. Measurements of  $\omega$  and  $\lambda$  from 42 high-redshift supernovae. *The Astrophysical Journal*, 517(2):565, 1999.
  - [2] Adam G Riess, Alexei V Filippenko, Peter Challis, Alejandro Clocchiatti, Alan Diercks, Peter M Garnavich, Ron L Gilliland, Craig J Hogan, Saurabh Jha, Robert P Kirshner, et al. Observational evidence from supernovae for an accelerating universe and a cosmological constant. *The astronomical journal*, 116(3):1009, 1998.
  - [3] Miguel Quartin, Mauricio O Calvao, Sergio E Joras, Ribamar RR Reis, and Ioav Waga. Dark interactions and cosmological fine-tuning. *Journal of Cosmology and Astroparticle Physics*, 2008(05):007, 2008.
  - [4] Sean M Carroll and Heywood Tam. Unitary evolution and cosmological fine-tuning. *arXiv preprint arXiv:1007.1417*, 2010.
  - [5] Hermano ES Velten, RF Vom Marttens, and Winifried Zimdahl. Aspects of the cosmological “coincidence problem”. *The European Physical Journal C*, 74:1–8, 2014.
  - [6] Navin Sivanandam. Is the cosmological coincidence a problem? *Physical Review D*, 87(8):083514, 2013.
  - [7] Varun Sahni. The cosmological constant problem and quintessence. *Classical and Quantum Gravity*, 19(13):3435, 2002.
  - [8] Ivaylo Zlatev, Limin Wang, and Paul J Steinhardt. Quintessence, cosmic coincidence, and the cosmological constant. *Physical Review Letters*, 82(5):896, 1999.
  - [9] Philippe Brax and Jerome Martin. Robustness of quintessence. *Physical Review D*, 61(10):103502, 2000.
  - [10] T Barreiro, Edmund J Copeland, and NJ a Nunes. Quintessence arising from exponential potentials. *Physical Review D*, 61(12):127301, 2000.
  - [11] Robert R Caldwell, Marc Kamionkowski, and Nevin N Weinberg. Phantom energy: dark energy with  $w < -1$  causes a cosmic doomsday. *Physical review letters*, 91(7):071301, 2003.
  - [12] Jaume Garriga and Viatcheslav F Mukhanov. Perturbations in k-inflation. *Physics Letters B*, 458(2-3):219–225, 1999.
  - [13] Takeshi Chiba, Takahiro Okabe, and Masahide Yamaguchi. Kinetically driven quintessence. *Physical Review D*, 62(2):023511, 2000.
  - [14] Christian Armendariz-Picon, V Mukhanov, and Paul J Steinhardt. Dynamical solution to the problem of a small cosmological constant and late-time cosmic acceleration. *Physical Review Letters*, 85(21):4438, 2000.
  - [15] Vittorio Gorini, Alexander Kamenshchik, and Ugo Moschella. Can the chaplygin gas be a plausible model for dark energy? *Physical Review D*, 67(6):063509, 2003.
  - [16] Thomas P Sotiriou and Valerio Faraoni.  $f(R)$  theories of gravity. *Reviews of Modern Physics*, 82(1):451, 2010.
  - [17] Guido Cognola, Emilio Elizalde, Shin’ichi Nojiri, Sergei D. Odintsov, and Sergio Zerbini. Dark energy in modified Gauss-Bonnet gravity: Late-time acceleration and the hierarchy problem. *Phys. Rev. D*, 73:084007, 2006.
  - [18] Tiberiu Harko, Francisco SN Lobo, Shin’ichi Nojiri, and Sergei D Odintsov.  $f(R, t)$  gravity. *Physical Review D*, 84(2):024020, 2011.
  - [19] Yasunori Fujii and Kei-ichi Maeda. *The scalar-tensor theory of gravitation*. Cambridge University Press, 2003.
  - [20] Asifa Ashraf, G Mustafa, Mushtaq Ahmad, and Ibrar Hussain. Lorentz distributed wormhole solutions in  $f(t)$  gravity with off-diagonal tetrad under conformal motions. *Modern Physics Letters A*, 35(29):2050240, 2020.
  - [21] G Mustafa, G Abbas, and T Xia. Wormhole solutions in  $f(t, tg)$  gravity under gaussian and lorentzian non-commutative distributions with conformal motions. *Chinese Journal of Physics*, 60:362–378, 2019.
  - [22] Zinnat Hassan, Ghulam Mustafa, and Pradyumn Kumar Sahoo. Wormhole Solutions in Symmetric Teleparallel Gravity with Noncommutative Geometry. *Symmetry*, 13(7):1260, 2021.
  - [23] Yixin Xu, Guangjie Li, Tiberiu Harko, and Shi-Dong Liang. Regular article-theoretical physics. *Eur. Phys. J. C*, 79:708, 2019.
  - [24] Simran Arora, JRL Santos, and PK Sahoo. Constraining  $f(q, t)$  gravity from energy conditions. *Physics of the Dark Universe*, 31:100790, 2021.
  - [25] Salim H Shekh, Aylin Caliskan, Dr G Mustafa, Ertan Gudekli, Anirudh Pradhan, and Sunil Kumar Maurya. Observational constraints on parameterized deceleration parameter with  $f(q, t)$  gravity. *Available at SSRN 4384140*.
  - [26] Antonio Nájera and Amanda Fajardo. Cosmological perturbation theory in  $f(q, t)$  gravity. *Journal of Cosmology and Astroparticle Physics*, 2022(03):020, 2022.
  - [27] SA Narawade, M Koussour, and B Mishra. Constrained  $f(q, t)$  gravity accelerating cosmological model and its dynamical system analysis. *Nuclear Physics B*, 992:116233, 2023.
  - [28] SKJ Pacif. Dark energy models from a parametrization of  $h$ : a comprehensive analysis and observational constraints. *The European Physical Journal Plus*, 135(10):1–34, 2020.
  - [29] Shibesh Kumar Jas Pacif, Ratbay Myrzakulov, and Shynaray Myrzakul. Reconstruction of cosmic history from a simple parametrization of  $h$ . *International Journal of Geometric Methods in Modern Physics*, 14(07):1750111, 2017.
  - [30] Eric V Linder. Exploring the expansion history of the universe. *Physical review letters*, 90(9):091301, 2003.

- [31] Eric V Linder. Cosmological parametrization of gamma ray burst intensity distribution. *Astronomy and Astrophysics*, v. 326, p. 29-33, 326:29–33, 1997.
- [32] Eric V Linder and Dragan Huterer. How many dark energy parameters? *Physical Review D*, 72(4):043509, 2005.
- [33] Eric V Linder. The dynamics of quintessence, the quintessence of dynamics. *General Relativity and Gravitation*, 40:329–356, 2008.
- [34] Eric V Linder. Mapping the cosmological expansion. *Reports on Progress in Physics*, 71(5):056901, 2008.
- [35] Allan Sandage. Beginnings of observational cosmology in hubble’s time: historical overview. *The Hubble Deep Field*, pages 1–26, 1998.
- [36] Yu L Bolotin, VA Cherkaskiy, OA Lemets, DA Yerokhin, and LG Zazunov. Cosmology in terms of the deceleration parameter. part i. *arXiv preprint arXiv:1502.00811*, 2015.
- [37] D Sofuoğlu, H Baysal, and RK Tiwari. Observational constraints on the cubic parametrization of the deceleration parameter in  $f(r, t)$  gravity. *The European Physical Journal Plus*, 138(6):1–14, 2023.
- [38] Salvatore Capozziello, Rocco D’Agostino, and Orlando Luongo. Model-independent reconstruction of  $f(t)$  teleparallel cosmology. *General Relativity and Gravitation*, 49:1–21, 2017.
- [39] Salvatore Capozziello, Rocco D’Agostino, and Orlando Luongo. Kinematic model-independent reconstruction of palatini  $f(r)$  cosmology. *General Relativity and Gravitation*, 51:1–19, 2019.
- [40] Salvatore Capozziello, Rocco D’Agostino, and Orlando Luongo. Thermodynamic parametrization of dark energy. *Physics of the Dark Universe*, 36:101045, 2022.
- [41] Salvatore Capozziello, Rocco D’Agostino, and Orlando Luongo. Cosmographic analysis with Chebyshev polynomials. *Mon. Not. Roy. Astron. Soc.*, 476(3):3924–3938, 2018.
- [42] S. Capozziello, R. D’Agostino, and O. Luongo. High-redshift cosmography: auxiliary variables versus Padé polynomials. *Mon. Not. Roy. Astron. Soc.*, 494(2):2576–2590, 2020.
- [43] Orlando Luongo. Cosmography with the Hubble parameter. *Mod. Phys. Lett. A*, 26:1459–1466, 2011.
- [44] Orlando Luongo. Dark energy from a positive jerk parameter. *Mod. Phys. Lett. A*, 28:1350080, 2013.
- [45] Orlando Luongo and Hernando Quevedo. A Unified Dark Energy Model from a Vanishing Speed of Sound with Emergent Cosmological Constant. *Int. J. Mod. Phys. D*, 23:1450012, 2014.
- [46] Orlando Luongo, Giovanni Battista Pisani, and Antonio Troisi. Cosmological degeneracy versus cosmography: a cosmographic dark energy model. *Int. J. Mod. Phys. D*, 26(03):1750015, 2016.
- [47] Orlando Luongo and Marco Muccino. Speeding up the universe using dust with pressure. *Phys. Rev. D*, 98(10):103520, 2018.
- [48] Alejandro Aviles, Christine Gruber, Orlando Luongo, and Hernando Quevedo. Cosmography and constraints on the equation of state of the Universe in various parametrizations. *Phys. Rev. D*, 86:123516, 2012.
- [49] Alejandro Aviles, Alessandro Bravetti, Salvatore Capozziello, and Orlando Luongo. Precision cosmology with Padé rational approximations: Theoretical predictions versus observational limits. *Phys. Rev. D*, 90(4):043531, 2014.
- [50] Christine Gruber and Orlando Luongo. Cosmographic analysis of the equation of state of the universe through Padé approximations. *Phys. Rev. D*, 89(10):103506, 2014.
- [51] Alejandro Aviles, Jaime Klapp, and Orlando Luongo. Toward unbiased estimations of the statefinder parameters. *Phys. Dark Univ.*, 17:25–37, 2017.
- [52] Alejandro Aviles, Jaime Klapp, and Orlando Luongo. Toward unbiased estimations of the statefinder parameters. *Phys. Dark Univ.*, 17:25–37, 2017.
- [53] Sergio del Campo, Ivan Duran, Ramon Herrera, and Diego Pavon. Three thermodynamically-based parameterizations of the deceleration parameter. *Phys. Rev. D*, 86:083509, 2012.
- [54] JV Cunha and Jose Ademir Sales de Lima. Transition redshift: new kinematic constraints from supernovae. *Monthly Notices of the Royal Astronomical Society*, 390(1):210–217, 2008.
- [55] J. V. Cunha. Kinematic Constraints to the Transition Redshift from SNe Ia Union Data. *Phys. Rev. D*, 79:047301, 2009.
- [56] Adam G. Riess et al. Type Ia supernova discoveries at  $z > 1$  from the Hubble Space Telescope: Evidence for past deceleration and constraints on dark energy evolution. *Astrophys. J.*, 607:665–687, 2004.
- [57] Lix-In Xu, Cheng-Wu Zhang, Bao-Rong Chang, and Hong-Ya Liu. Constraints to deceleration parameters by recent cosmic observations. *Mod. Phys. Lett. A*, 23:1939–1948, 2008.
- [58] Lixin Xu and Jianbo Lu. Cosmic constraints on deceleration parameter with SNe Ia and CMB. *Mod. Phys. Lett. A*, 24:369–376, 2009.
- [59] Remya Nair, Sanjay Jhingan, and Deepak Jain. Cosmokinetics: a joint analysis of standard candles, rulers and cosmic clocks. *Journal of Cosmology and Astroparticle Physics*, 2012(01):018, 2012.
- [60] Ozgur Akarsu, Tekin Dereli, Suresh Kumar, and Lixin Xu. Probing kinematics and fate of the Universe with linearly time-varying deceleration parameter. *Eur. Phys. J. Plus*, 129:22, 2014.
- [61] Beethoven Santos, Joel C Carvalho, and Jailson S Alcaniz. Current constraints on the epoch of cosmic acceleration. *Astroparticle Physics*, 35(1):17–20, 2011.
- [62] Yun-Gui Gong and Anzhong Wang. Reconstruction of the deceleration parameter and the equation of state of dark energy. *Phys. Rev. D*, 75:043520, 2007.
- [63] Michael S. Turner and Adam G. Riess. Do SNe Ia provide direct evidence for past deceleration of the universe? *Astrophys. J.*, 569:18, 2002.
- [64] Abdulla Al Mamon and Sudipta Das. A divergence free parametrization of deceleration parameter for scalar field dark energy. *Int. J. Mod. Phys. D*, 25(03):1650032, 2016.
- [65] Amine Bouali, Himanshu Chaudhary, Ujjal Debnath, Tanusree Roy, and G Mustafa. Constraints on the parameterized deceleration parameter in frw universe. *arXiv preprint arXiv:2301.12107*, 2023.
- [66] Himanshu Chaudhary, Amine Bouali, Ujjal Debnath, Tanusree Roy, and Ghulam Mustafa. Constraints on the parameterized deceleration parameter in frw universe. *Physica Scripta*, 2023.
- [67] Amine Bouali, Himanshu Chaudhary, Ujjal Debnath, Alok Sardar, and G Mustafa. Data analysis of three

- parameter models of deceleration parameter in frw universe. *arXiv preprint arXiv:2304.13137*, 2023.
- [68] Amine Bouali, BK Shukla, Himanshu Chaudhary, Rishi Kumar Tiwari, Mahvish Samar, and G Mustafa. Cosmological tests of parametrization  $q = \alpha - \beta h$  in  $f(q)$  frw cosmology. *International Journal of Geometric Methods in Modern Physics*, page 2350152, 2023.
- [69] Amine Bouali, Himanshu Chaudhary, Saadia Mumtaz, G Mustafa, and SK Maurya. Observational constraining study of new deceleration parameters in frw universe. *Fortschritte der Physik*, page 2300033, 2023.
- [70] Amine Bouali, BK Shukla, Himanshu Chaudhary, Rishi Kumar Tiwari, and Marco San Martin. Cosmographic studies of  $q = \alpha - \beta h$  parametrization in  $f(t)$  framework. *International Journal of Geometric Methods in Modern Physics*, 2023.
- [71] Himanshu Chaudhary, Aditya Kaushik, and Ankita Kohli. Cosmological test of  $\sigma\theta$  as function of scale factor in  $f(r, t)$  framework. *New Astronomy*, 103:102044, 2023.
- [72] Amine Bouali, Himanshu Chaudhary, Amritansh Mehrotra, and SKJ Pacif. Model-independent study for a quintessence model of dark energy: Analysis and observational constraints. *arXiv preprint arXiv:2304.02652*, 2023.
- [73] Abdulla Al Mamon and Sudipta Das. A parametric reconstruction of the deceleration parameter. *The European Physical Journal C*, 77(7):495, 2017.
- [74] Abdulla Al Mamon. Constraints on a generalized deceleration parameter from cosmic chronometers. *Modern Physics Letters A*, 33(10n11):1850056, 2018.
- [75] Bijan Berenji, Jennifer Gaskins, and Manuel Meyer. Constraints on axions and axionlike particles from fermi large area telescope observations of neutron stars. *Physical Review D*, 93(4):045019, 2016.
- [76] Tonatiuh Matos and L Arturo Urena-Lopez. Further analysis of a cosmological model with quintessence and scalar dark matter. *Physical Review D*, 63(6):063506, 2001.
- [77] Tonatiuh Matos and L Arturo Urena-Lopez. Quintessence and scalar dark matter in the universe. *Classical and Quantum Gravity*, 17(13):L75, 2000.
- [78] Tonatiuh Matos, F Siddhartha Guzmán, and Dario Nunez. Spherical scalar field halo in galaxies. *Physical Review D*, 62(6):061301, 2000.
- [79] L Arturo Urena-Lopez and Tonatiuh Matos. New cosmological tracker solution for quintessence. *Physical Review D*, 62(8):081302, 2000.
- [80] Yixin Xu, Guangjie Li, Tiberiu Harko, and Shi-Dong Liang.  $f(q, t)$  gravity. *The European Physical Journal C*, 79:1–19, 2019.
- [81] Yixin Xu, Tiberiu Harko, Shahab Shahidi, and Shi-Dong Liang. Weyl type  $f(q, t)$  gravity, and its cosmological implications. *The European Physical Journal C*, 80(5):1–22, 2020.
- [82] Simran Arora and PK Sahoo. Energy conditions in  $f(q, t)$  gravity. *Physica Scripta*, 95(9):095003, 2020.
- [83] Matt Visser. General relativistic energy conditions: The hubble expansion in the epoch of galaxy formation. *Physical Review D*, 56(12):7578, 1997.
- [84] M Koussour, SKJ Pacif, M Bennai, and PK Sahoo. Dynamical dark energy models from a new hubble parameter in  $f(q)$  gravity. *arXiv preprint arXiv:2208.04723*, 2022.
- [85] SKJ Pacif, Md Salahuddin Khan, LK Paikroy, and Shalini Singh. An accelerating cosmological model from a parametrization of hubble parameter. *Modern Physics Letters A*, 35(05):2050011, 2020.
- [86] SKJ Pacif, Simran Arora, and PK Sahoo. Late-time acceleration with a scalar field source: Observational constraints and statefinder diagnostics. *Physics of the Dark Universe*, 32:100804, 2021.
- [87] Ritika Nagpal and Shibesh Kumar Jas Pacif. Cosmological aspects of  $f(r, t)$  gravity in a simple model with a parametrization of  $q$ . *The European Physical Journal Plus*, 136(8):875, 2021.
- [88] Ritika Nagpal, SKJ Pacif, JK Singh, Kazuharu Bamba, and A Beesham. Analysis with observational constraints in  $\lambda$ -cosmology in  $f(r, t)$  gravity. *The European Physical Journal C*, 78:1–17, 2018.
- [89] Roberto Trotta. Bayesian methods in cosmology. *arXiv preprint arXiv:1701.01467*, 2017.
- [90] Joël Akeret, Sebastian Seehars, Adam Amara, Alexandre Refregier, and André Csillaghy. Cosmohammer: Cosmological parameter estimation with the mcmc hammer. *Astronomy and Computing*, 2:27–39, 2013.
- [91] Raul Jimenez and Abraham Loeb. Constraining cosmological parameters based on relative galaxy ages. *The Astrophysical Journal*, 573(1):37, 2002.
- [92] Cong Zhang, Han Zhang, Shuo Yuan, Siqu Liu, Tong-Jie Zhang, and Yan-Chun Sun. Four new observational  $h(z)$  data from luminous red galaxies in the sloan digital sky survey data release seven. *Research in Astronomy and Astrophysics*, 14(10):1221, 2014.
- [93] Raul Jimenez, Licia Verde, Tommaso Treu, and Daniel Stern. Constraints on the equation of state of dark energy and the hubble constant from stellar ages and the cosmic microwave background. *The Astrophysical Journal*, 593(2):622, 2003.
- [94] Michele Moresco, Lucia Pozzetti, Andrea Cimatti, Raul Jimenez, Claudia Maraston, Licia Verde, Daniel Thomas, Annalisa Citro, Rita Tojeiro, and David Wilkinson. A 6% measurement of the hubble parameter at  $z = 0.45$ : direct evidence of the epoch of cosmic re-acceleration. *Journal of Cosmology and Astroparticle Physics*, 2016(05):014–014, 2016.
- [95] Joan Simon, Licia Verde, and Raul Jimenez. Constraints on the redshift dependence of the dark energy potential. *Physical Review D*, 71(12):123001, 2005.
- [96] Michele Moresco, Andrea Cimatti, Raul Jimenez, Lucia Pozzetti, Gianni Zamorani, Micoll Bolzonella, James Dunlop, Fabrice Lamareille, Marco Mignoli, Henry Pearce, et al. Improved constraints on the expansion rate of the universe up to  $z = 1.1$  from the spectroscopic evolution of cosmic chronometers. *Journal of Cosmology and Astroparticle Physics*, 2012(08):006–006, 2012.
- [97] Daniel Stern, Raul Jimenez, Licia Verde, Marc Kamionkowski, and S Adam Stanford. Cosmic chronometers: constraining the equation of state of dark energy. i:  $H(z)$  measurements. *Journal of Cosmology and Astroparticle Physics*, 2010(02):008, 2010.
- [98] Michele Moresco. Raising the bar: new constraints on the hubble parameter with cosmic chronometers at  $z = 2$ . *Monthly Notices of the Royal Astronomical Society: Letters*, 450(1):L16–L20, 2015.
- [99] Adrià Gómez-Valent and Luca Amendola.  $H_0$  from cosmic chronometers and type ia supernovae, with gaus-

- sian processes and the novel weighted polynomial regression method. *Journal of Cosmology and Astroparticle Physics*, 2018(04):051, 2018.
- [100] M López-Corredoira and A Vazdekis. Impact of young stellar components on quiescent galaxies: deconstructing cosmic chronometers. *Astronomy & Astrophysics*, 614:A127, 2018.
- [101] M López-Corredoira, A Vazdekis, CM Gutiérrez, and N Castro-Rodríguez. Stellar content of extremely red quiescent galaxies at  $z_i$  2. *Astronomy & Astrophysics*, 600:A91, 2017.
- [102] Licia Verde, Pavlos Protopapas, and Raul Jimenez. The expansion rate of the intermediate universe in light of planck. *Physics of the Dark Universe*, 5:307–314, 2014.
- [103] Marek Kowalski, David Rubin, Greg Aldering, RJ Agostinho, A Amadon, R Amanullah, C Balland, K Barbary, G Blanc, PJ Challis, et al. Improved cosmological constraints from new, old, and combined supernova data sets. *The Astrophysical Journal*, 686(2):749, 2008.
- [104] Rahman Amanullah, Chris Lidman, D Rubin, G Aldering, P Astier, K Barbary, MS Burns, A Conley, KS Dawson, SE Deustua, et al. Spectra and hubble space telescope light curves of six type ia supernovae at 0.511  $z_i$  1.12 and the union2 compilation. *The Astrophysical Journal*, 716(1):712, 2010.
- [105] N Suzuki, D Rubin, C Lidman, G Aldering, R Amanullah, K Barbary, LF Barrientos, J Botyanszki, M Brodwin, N Connolly, et al. The hubble space telescope cluster supernova survey. v. improving the dark-energy constraints above  $z_i$  1 and building an early-type-hosted supernova sample. *The Astrophysical Journal*, 746(1):85, 2012.
- [106] MEA Betoule, R Kessler, J Guy, J Mosher, D Hardin, R Biswas, P Astier, P El-Hage, M König, S Kuhlmann, et al. Improved cosmological constraints from a joint analysis of the sdss-ii and snls supernova samples. *Astronomy & Astrophysics*, 568:A22, 2014.
- [107] Daniel Moshe Scolnic, DO Jones, A Rest, YC Pan, R Chornock, RJ Foley, ME Huber, R Kessler, Gautham Narayan, AG Riess, et al. The complete light-curve sample of spectroscopically confirmed sne ia from pan-starrs1 and cosmological constraints from the combined pantheon sample. *The Astrophysical Journal*, 859(2):101, 2018.
- [108] Dan Scolnic, Dillon Brout, Anthony Carr, Adam G Riess, Tamara M Davis, Arianna Dwomoh, David O Jones, Noor Ali, Pranav Charvu, Rebecca Chen, et al. The pantheon+ type ia supernova sample: the full dataset and light-curve release. *arXiv preprint arXiv:2112.03863*, 2021.
- [109] Marek Demianski, Ester Piedipalumbo, Disha Sawant, and Lorenzo Amati. Cosmology with gamma-ray bursts-i. the hubble diagram through the calibrated ep, i-eiso correlation. *Astronomy & Astrophysics*, 598:A112, 2017.
- [110] Carl Roberts, Keith Horne, Alistair O Hodson, and Alasdair Dorkenoo Leggat. Tests of  $\Lambda$  cdM and conformal gravity using grb and quasars as standard candles out to  $z \sim 8$ . *arXiv preprint arXiv:1711.10369*, 2017.
- [111] Will J Percival, Beth A Reid, Daniel J Eisenstein, Neta A Bahcall, Tamas Budavari, Joshua A Frieman, Masataka Fukugita, James E Gunn, Zeljko Ivezić, Gillian R Knapp, et al. Baryon acoustic oscillations in the sloan digital sky survey data release 7 galaxy sample. *Monthly Notices of the Royal Astronomical Society*, 401(4):2148–2168, 2010.
- [112] Florian Beutler, Chris Blake, Matthew Colless, D Heath Jones, Lister Staveley-Smith, Lachlan Campbell, Quentin Parker, Will Saunders, and Fred Watson. The 6df galaxy survey: baryon acoustic oscillations and the local hubble constant. *Monthly Notices of the Royal Astronomical Society*, 416(4):3017–3032, 2011.
- [113] Timothee Delubac, James Rich, Stephen Bailey, Andreu Font-Ribera, David Kirkby, J-M Le Goff, Matthew M Pieri, Anze Slosar, Éric Aubourg, Julian E Bautista, et al. Baryon acoustic oscillations in the  $ly\alpha$  forest of boss quasars. *Astronomy & Astrophysics*, 552:A96, 2013.
- [114] Lauren Anderson, Eric Aubourg, Stephen Bailey, Dmitry Bizyaev, Michael Blanton, Adam S Bolton, Jon Brinkmann, Joel R Brownstein, Angela Burden, Antonio J Cuesta, et al. The clustering of galaxies in the sdss-iii baryon oscillation spectroscopic survey: baryon acoustic oscillations in the data release 9 spectroscopic galaxy sample. *Monthly Notices of the Royal Astronomical Society*, 427(4):3435–3467, 2012.
- [115] Hee-Jong Seo, Shirley Ho, Martin White, Antonio J Cuesta, Ashley J Ross, Shun Saito, Beth Reid, Nikhil Padmanabhan, Will J Percival, Roland De Putter, et al. Acoustic scale from the angular power spectra of sdss-iii dr8 photometric luminous galaxies. *The Astrophysical Journal*, 761(1):13, 2012.
- [116] Ashley J Ross, Lado Samushia, Cullan Howlett, Will J Percival, Angela Burden, and Marc Manera. The clustering of the sdss dr7 main galaxy sample–i. a 4 per cent distance measure at  $z = 0.15$ . *Monthly Notices of the Royal Astronomical Society*, 449(1):835–847, 2015.
- [117] Rita Tojeiro, Ashley J Ross, Angela Burden, Lado Samushia, Marc Manera, Will J Percival, Florian Beutler, J Brinkmann, Joel R Brownstein, Antonio J Cuesta, et al. The clustering of galaxies in the sdss-iii baryon oscillation spectroscopic survey: galaxy clustering measurements in the low-redshift sample of data release 11. *Monthly Notices of the Royal Astronomical Society*, 440(3):2222–2237, 2014.
- [118] Julian E Bautista, Mariana Vargas-Magaña, Kyle S Dawson, Will J Percival, Jonathan Brinkmann, Joel Brownstein, Benjamin Camacho, Johan Comparat, Hector Gil-Marín, Eva-Maria Mueller, et al. The sdss-iv extended baryon oscillation spectroscopic survey: baryon acoustic oscillations at redshift of 0.72 with the dr14 luminous red galaxy sample. *The Astrophysical Journal*, 863(1):110, 2018.
- [119] E De Carvalho, A Bernui, GC Carvalho, CP Novaes, and HS Xavier. Angular baryon acoustic oscillation measure at  $z = 2.225$  from the sdss quasar survey. *Journal of Cosmology and Astroparticle Physics*, 2018(04):064, 2018.
- [120] Metin Ata, Falk Baumgarten, Julian Bautista, Florian Beutler, Dmitry Bizyaev, Michael R Blanton, Jonathan A Blazek, Adam S Bolton, Jonathan Brinkmann, Joel R Brownstein, et al. The clustering of the sdss-iv extended baryon oscillation spectroscopic survey dr14 quasar sample: first measurement of baryon acoustic oscillations between redshift 0.8 and 2.2. *Monthly Notices of the Royal Astronomical Society*, 473(4):4773–4794, 2018.



- [121] TMC Abbott, FB Abdalla, A Alarcon, S Allam, F Andrade-Oliveira, J Annis, S Avila, Mandakranta Banerji, N Banik, K Bechtol, et al. Dark energy survey year 1 results: Measurement of the baryon acoustic oscillation scale in the distribution of galaxies to redshift 1. *Monthly Notices of the Royal Astronomical Society*, 483(4):4866–4883, 2019.
- [122] Z Molavi and A Khodam-Mohammadi. Observational tests of gauss-bonnet like dark energy model. *The European Physical Journal Plus*, 134(6):254, 2019.
- [123] David Benisty and Denitsa Staicova. Testing late-time cosmic acceleration with uncorrelated baryon acoustic oscillation dataset. *Astronomy & Astrophysics*, 647:A38, 2021.
- [124] Natalie B Hogg, Matteo Martinelli, and Savvas Nesseris. Constraints on the distance duality relation with standard sirens. *Journal of Cosmology and Astroparticle Physics*, 2020(12):019, 2020.
- [125] Matteo Martinelli, Carlos Jose Amaro Parente Martins, S Nesseris, Domenico Sapone, I Tutusaus, Anastasios Avgoustidis, Stefano Camera, Carmelita Carbone, S Casas, Stéphane Ilić, et al. Euclid: Forecast constraints on the cosmic distance duality relation with complementary external probes. *Astronomy & Astrophysics*, 644:A80, 2020.
- [126] Lu Chen, Qing-Guo Huang, and Ke Wang. Distance priors from planck final release. *Journal of Cosmology and Astroparticle Physics*, 2019(02):028, 2019.
- [127] Matt Visser. Cosmography: Cosmology without the einstein equations. *General Relativity and Gravitation*, 37:1541–1548, 2005.
- [128] Matt Visser. Jerk, snap and the cosmological equation of state. *Classical and Quantum Gravity*, 21(11):2603, 2004.
- [129] Ujjaini Alam, Varun Sahni, Tarun Deep Saini, and AA Starobinsky. Exploring the expanding universe and dark energy using the statefinder diagnostic. *Monthly Notices of the Royal Astronomical Society*, 344(4):1057–1074, 2003.
- [130] Kazuharu Bamba, Salvatore Capozziello, Shin’ichi Nojiri, and Sergei D Odintsov. Dark energy cosmology: the equivalent description via different theoretical models and cosmography tests. *Astrophysics and Space Science*, 342:155–228, 2012.
- [131] Varun Sahni, Tarun Deep Saini, Alexei A Starobinsky, and Ujjaini Alam. Statefinder—a new geometrical diagnostic of dark energy. *Journal of Experimental and Theoretical Physics Letters*, 77:201–206, 2003.
- [132] Edmund J Copeland, Mohammad Sami, and Shinji Tsujikawa. Dynamics of dark energy. *International Journal of Modern Physics D*, 15(11):1753–1935, 2006.
- [133] Matteo Martinelli, Carlos Jose Amaro Parente Martins, S Nesseris, Domenico Sapone, I Tutusaus, Anastasios Avgoustidis, Stefano Camera, Carmelita Carbone, S Casas, Stéphane Ilić, et al. Euclid: Forecast constraints on the cosmic distance duality relation with complementary external probes. *Astronomy & Astrophysics*, 644:A80, 2020.
- [134] Ujjaini Alam, Varun Sahni, Tarun Deep Saini, and Alexei A Starobinsky. Is there supernova evidence for dark energy metamorphosis? *Monthly Notices of the Royal Astronomical Society*, 354(1):275–291, 2004.
- [135] Varun Sahni, Tarun Deep Saini, Alexei A Starobinsky, and Ujjaini Alam. Statefinder—a new geometrical diagnostic of dark energy. *Journal of Experimental and Theoretical Physics Letters*, 77:201–206, 2003.
- [136] Ujjaini Alam, Varun Sahni, Tarun Deep Saini, and AA Starobinsky. Exploring the expanding universe and dark energy using the statefinder diagnostic. *Monthly Notices of the Royal Astronomical Society*, 344(4):1057–1074, 2003.
- [137] Andrew R Liddle. How many cosmological parameters. *Monthly Notices of the Royal Astronomical Society*, 351(3):L49–L53, 2004.
- [138] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.
- [139] Savvas Nesseris and Juan Garcia-Bellido. Is the jeffreys’ scale a reliable tool for bayesian model comparison in cosmology? *Journal of Cosmology and Astroparticle Physics*, 2013(08):036, 2013.
- [140] Harold Jeffreys. *The theory of probability*. OUP Oxford, 1998.
- [141] Martin Kerscher and Jochen Weller. On model selection in cosmology. *SciPost Physics Lecture Notes*, page 009, 2019.
- [142] Nathan J Secrest, Sebastian von Hausegger, Mohamed Rameez, Roya Mohayaee, and Subir Sarkar. A challenge to the standard cosmological model. *The Astrophysical journal letters*, 937(2):L31, 2022.
- [143] Yashar Akrami, Tomi S Koivisto, and Marit Sandstad. Cosmological constraints on ghost-free bigravity: background dynamics and late-time acceleration. In *THE THIRTEENTH MARCEL GROSSMANN MEETING: On Recent Developments in Theoretical and Experimental General Relativity, Astrophysics and Relativistic Field Theories*, pages 1252–1254. World Scientific, 2015.
- [144] Christopher M Hirata and Uroš Seljak. Analyzing weak lensing of the cosmic microwave background using the likelihood function. *Physical Review D*, 67(4):043001, 2003.
- [145] E O’MONGAIN. Application of statistics to results in gamma ray astronomy. *Nature*, 241(5389):376–379, 1973.
- [146] Alma X Gonzalez-Morales, Robert Poltis, Blake D Sherwin, and Licia Verde. Are priors responsible for cosmology favoring additional neutrino species? *arXiv preprint arXiv:1106.5052*, 2011.



# Android Malware Analysis using Coefficient of Multiple Correlation

Shresth Jain  
Dept. of Applied Mathematics  
Delhi Technological University  
New Delhi, India  
shresthjain\_2k19mc121@dtu.ac.in

Sarthak Kapoor  
Dept. of Applied Mathematics  
Delhi Technological University  
New Delhi, India  
sarthakkapoor\_2k19mc115@dtu.ac.in

Anshul Arora  
Department of Applied Mathematics  
Delhi Technological University  
Delhi, India  
anshularora@dtu.ac.in

Sachin Kumar Sharma  
Dept. of Applied Mathematics  
Delhi Technological University  
New Delhi, India  
sachinkumarsharma\_2k19mc111@dtu.ac.in

**Abstract**—Android being the most famous Operating System (OS) for smart hand-held devices also serves as a prime attraction for cyber-criminals and black hat hackers. Hacking into these devices through malware applications gives them access to the data which can be used for personal gains. These applications have been a prime source of cyber crimes, information leaks, financial frauds, and much more. In order to get control over malware applications, it is of utmost importance to detect these applications before they are installed on any system. This can prevent huge losses for mankind. Malware detection has become a challenge due to constant technological advancements. Most of the data about an app including permissions, activities, services, etc can be extracted using its manifest file. The concept of Multiple Correlation Coefficient has been used to rank these permissions and then build a machine learning and deep learning model using various classifiers in a k-fold setup. The results suggest that the presented model gives a detection accuracy of 97.54% with Random Forest Classifier. This accuracy is observed with the top 210 permissions as ranked by Multiple Correlation Coefficient.

**Keywords:** Mobile Security, Mobile Malware, Malware Detection, Permissions, Multiple Correlation Coefficient.

## I. INTRODUCTION

Currently, as hand-held devices become more and more prevalent, smartphones are gaining access to an increasing amount of private information. Advanced malware, specifically Android spyware, acquires or uses such data without the consent of the user. Therefore, it becomes highly important to provide constructive techniques for evaluating and identifying these threats.

The hand-held device operating system (OS) popularly known as Android was developed specifically for touchscreen devices like smart mobile phones and tablets.

The important milestones of the Android platform are displayed in Table I. Every program on Android runs in a security isolation container known as the "Sandbox". The permission system controls how resources are accessed by Android applications and verify that they do so legally and are not doing it maliciously. Different methods are used to create restrictions. To safeguard data, storage isolation is sometimes used; alternatively, "Permission List" mechanism constraints

are set up to limit access to important APIs. Some of these secured APIs are listed below: Place of the camera (GPS) SMS/MMS, Bluetooth phone, network, and data (GSM and Wi-Fi).

## II. RELATED WORK

This subsection focuses on the works suggested in the literature for detecting Android malware. The detection mechanisms for Android malware are classified into three main categories, which are Static, Dynamic, and Hybrid, based on the detection features utilized. These techniques will be further discussed in the upcoming sections.

### A. Static Detection

Ahmed et al. [1] proposes an Android malware detection approach based on static feature analysis using four different machine learning algorithms: K-Nearest Neighbor (KNN), Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM). The RF algorithm achieved the highest detection accuracy of 99.5%, while the KNN algorithm achieved the highest F1-score of 0.999. Alazab et al.[2] uses a combination of permission requests and API calls on a dataset of 10,000 android applications employing machine learning techniques.

By leveraging the static properties of permissions, intents, and APIs, Taheri et al. [3] were able to determine the similarities between malware and non-malware apps by applying Hamming Distance.

According to Selvaganapathy et. al.[4] the proficiency of malware developers has increased in order to evade detection by anti-malware software.

Arora et al. [5] examined permissions in pairs in order to find malware. In order to identify malware in Android applications, Wang et al. [6] combined API calls and associated function call graphs with text properties like permissions and intents. From the Android samples, the FAMD (Fast Android Malware Detection) model [7] extracted permissions and Dalvik code segments. They also used the CatBoost classifier for malware identification and the N-gram approach to

TABLE I  
MAJOR MILESTONES IN ANDROID OPERATING SYSTEM

Date	Event
1-July-2005	Google acquired Android Inc.
12-November-2007	Android was released.
28-August-2008	Announcement of Android market.
23-September-2008	Release of Android 1.0 platform.
21-November-2008	Android was open-sourced.
13-February-2009	USA Android Market started accepting Paid applications.
2009-2022	Android platform grew over multiple versions and still gets continuous updates.
The latest version is Android 12.0 Snow Clone	

minimize the dimensionality of the feature space. To identify fraudulent Android apps, Firdaus et al. [8] applied Factorization Machine architecture to the collection of manifest components.

For the purpose of identifying malware, the approach proposed by Vu et al. [9] examined API calls and their call graphs. Each application's adjacency matrix was created using the APK source code. APK source code was transformed into call-graphs using API calls, and CNN was then trained to recognise malicious programmes. A couple of more studies, such as [10] and [11] extract static features for malware detection on the Android platform. A feature transformation-based Android malware detector is discussed by Han et al. [12], which presents three new forms of feature transformations that reversibly shift well-known features into a new feature domain. Y. Zhang et al. [13] used the code semantic structure features to reflect deep semantic information and proposed a pre-processing method of APK files to generate graphics that reflect the code semantic features, whereas Azar et al. [14] proposed a model with the capabilities of both binary file feature representation and feature selection for malware detection.

To provide a real-time and responsive detection environment on mobile devices against malware, Feng et al. [15] used customised deep neural networks, and Bibi et al. [16] proposed a strong, scalable, and effective Cuda-empowered multi-class malware detection technique using Gated Recurrent Unit (GRU) to identify sophisticated Android malware. Common malicious system call codes were discovered by Surendran et al. [17] in the system call sequence of many malware types.

### B. Dynamic Detection

The use of static detection mechanisms may not have the ability to identify the malicious component as they do not run the applications. This means that during updates, these harmful components can be downloaded without being detected. Therefore, researchers proposed the use of dynamic solutions to detect malware in Android. Some research has focused on detecting malicious apps that use features based on the Android operating system. By using Natural Language Processing techniques on the HyperText Transfer Protocol (HTTP) headers, Wang et al. performed virus detection in [18].

### C. Hybrid Detection

In the literature, there are hybrid solutions that aim to merge the benefits of both static and dynamic approaches. These solutions combine static and dynamic features to develop a hybrid detection model. Machine learning approaches were used by Mahindru et al. [19] after important features including app ratings, dynamic API calls, permissions, and the number of people that downloaded the app were extracted. Mehtab et al. [20] introduced the "AdDroid model," which analyses Android behaviours like uploading a file to a server, connecting to the internet, installing packages on the device, etc. to detect malicious activity on a device.

Zhu et al. [21] discovered Android malware by keeping an eye on run-time-related events, critical APIs, and permissions. According to experimental findings cited by Khariwal et al. [22], the best feature set for Android malware analysis included 25 features, including 5 permissions, 2 intentions, 9 activities, 3 content providers, 4 hardware components, 1 service, and 1 broadcast receiver. They implemented the information gain to rank the manifest file components. In order to identify zero-day malware families, Qiu et al. [23] employed multi-label classification models to the set of extracted features, such as permissions, API calls, network addresses, etc. Recently developed methods by Kim et al. [24] and Zhu et al. [25], have also looked at manifest properties for malware identification. Ribeiro et al. [26] used features that show how memory, CPU, battery life, and network traffic are consumed when running mobile applications for virus detection. As a result of the computational complexity and significant overheads associated with extracting dynamic data such system calls as opposed to static techniques, a static malware detection approach is described in this research. Khariwal et al. [27], information acquired was used to rate permissions and intentions, and this rating was then used to identify dangerous apps. Arora et al. [28] joined the network traffic with permissions to introduce two unique hybrid detection approaches. Similar to this, static and dynamic feature combinations were examined for malware detection by Arora et al. [29] and arshad et al. [30].

## III. PROPOSED METHODOLOGY

In this section, we introduce the proposed method for identifying harmful Android applications. We acquired Android APK files and extracted the permissions specified in the AndroidManifest.xml files. These permissions served as



### C. Machine Learning Classifiers

In our model, we applied three machine learning classifiers namely, Naïve Bayes, Random Forest, and few more. The following discussion gives details on the classifiers which we have used.

**i. Naïve-Bayes:** Naïve-Bayes is a probabilistic, supervised machine learning technique that applies the famous Bayes theorem to classification tasks.

**ii. Random Forest:** Random Forest is a popular and efficient ensemble model that is based on Decision Trees with further enhancements resulting in a low bias and low variance final ensemble.

**iii. Decison trees:** A decision tree is a widely used machine learning approach for classification and regression issues.

**iv. Logistic Regression:** Logistic regression is a widely used statistical method for binary classification problems.

**iv. Extreme Gradient Boosting(XG Boost):** XGBoost (eX-treme Gradient Boosting) is a well-known and very efficient machine learning technique for supervised learning tasks.

### D. Deep Learning Classifiers

**i. Recurrent Neural Networks (RNNs)** Recurrent Neural Networks (RNNs) are a class of neural networks that are particularly suited for processing sequential data.

## IV. RESULTS AND DISCUSSION

This section presents an analysis of the outcomes generated by the proposed model.

**Table II: Top Ranked Permissions**

Permission	CMC Score
READ PHONE STATE	0.7186
SYSTEM ALERT WINDOW	0.6117
CHANGE WIFI STATE	0.5404
MOUNT UNMOUNT FILESYSTEMS	0.5372
GET TASKS	0.5140

In the first step, we identify the most significant permissions ranked by the CMC Score. The Table II, which displays the top 5 permissions ranked by the CMC Score, highlights that the permission "READ PHONE STATE" has the highest CMC Score. The top-ranked features have a positive impact on the proposed model's detection accuracy.

### A. Detection Results With Machine Learning and Deep Learning Algorithms

In this section, we report the outcomes of our proposed approach for detecting potential issues. First we present the detection results obtained by utilizing the Logistic Regression algorithm, where we evaluated different sets of permissions in each iteration. Table III presents a summary of the detection outcomes, where the accuracy is measured for each set of permissions used. The results show that an accuracy of 95.72% is achieved when utilizing the top 110 ranked permissions as sorted by the Coefficient of Multiple Correlation Value. Similarly, other results in the table can be interpreted accordingly.

**Table III : Detection Results with Logistic Regression**

No. of Ranked Permissions	Accuracy
Without Feature Selection	95.68
Top 10	91.78
Top 20	94.34
Top 30	94.62
Top 40	95.15
Top 50	95.10
Top 60	95.14
Top 70	95.55
Top 80	95.55
Top 90	95.59
Top 100	95.70
<b>Top 110</b>	<b>95.72</b>
Top 120	95.70
Top 130	95.69
Top 140	95.68
Top 150	95.68

Table IV presents the outcomes of utilizing the Naive Bayes classifier for identifying potentially harmful Android applications, with different sets of top-ranked permissions. We found that the highest accuracy achieved was 88.36% when using the top 30 ranked permissions. We observed that increasing the number of permissions beyond 30 did not result in a further improvement in the detection accuracy.

**Table IV: Detection Results with Naive Bayes Classifier**

No. of Ranked Permissions	Accuracy
Without Feature Selection	52.74
Top 10	88.10
Top 20	87.43
<b>Top 30</b>	<b>88.36</b>
Top 40	84.94
Top 50	84.21
Top 60	84.18
Top 70	83.47
Top 80	83.65
Top 90	82.32
Top 100	81.50

The outcomes of using the Decision Tree Classifier for detecting Android malware are summarized in Table V. The top 200 ranked permissions yielded the highest accuracy of 96.86% with no improvement in detection accuracy upon further increasing the number of permissions.

Table VI presents the outcomes of using the Random Forest Classifier algorithm for detecting Android malware. The top 210 ranked permissions resulted in the highest accuracy of 97.54%. As with the other classifiers, increasing the number of permissions beyond 210 did not lead to any improvement in the detection accuracy.

Table VII outlines the results obtained from utilizing the XGBoost algorithm for detecting Android malware. The highest accuracy of 87.32% is achieved when utilizing the top 140 ranked permissions. Similar to the other classifiers, we observed that including more permissions did not lead to a significant improvement in detection accuracy beyond this point.

Table V: Detection Results with Decision Tree Classifier

No. of Ranked Permissions	Accuracy
Without Feature Selection	96.84
Top 10	92.63
Top 20	95.63
Top 30	96.24
Top 40	96.47
Top 50	96.54
Top 60	96.72
Top 70	96.78
Top 80	96.80
Top 90	96.83
Top 100	96.82
Top 110	96.83
Top 120	96.84
Top 130	96.84
Top 140	96.84
Top 150	96.83
Top 160	96.82
Top 170	96.82
Top 180	96.83
Top 190	96.81
<b>Top 200</b>	<b>96.86</b>
Top 210	96.83
Top 220	96.80
Top 230	96.82
Top 240	96.84

Table VII : Detection Results with XGBoost Algorithm

No. of Ranked Permissions	Accuracy
Without Feature Selection	87.27
Top10	77.12
Top20	83.31
Top30	84.84
Top40	85.66
Top50	85.90
Top60	86.524
Top70	86.72
Top80	87.07
Top90	87.14
Top100	87.17
Top110	87.17
Top120	87.22
Top130	87.29
<b>Top140</b>	<b>87.32</b>
Top150	87.31
Top160	87.29
Top170	87.30
Top180	87.30
Top190	87.29
Top200	87.30
Top210	87.29
Top220	87.29
Top230	87.29
Top240	87.29

Table VI: Detection Results with Random Forest Classifier

No. of Ranked Permissions	Accuracy
Without Feature Selection	97.51
Top 10	92.61
Top 20	95.84
Top 30	96.73
Top 40	97.01
Top 50	97.18
Top 60	97.36
Top 70	97.39
Top 80	97.42
Top 90	97.43
Top 100	97.45
Top 110	97.45
Top 120	97.47
Top 130	97.49
Top 140	97.50
Top 150	97.50
Top 160	97.51
Top 170	97.50
Top 180	97.50
Top 190	97.49
Top 200	97.51
<b>Top 210</b>	<b>97.54</b>
Top 220	97.51
Top 230	97.53
Top 240	97.52

Table VIII: Detection Results with Recurrent Neural Network

No. of Ranked Permissions	Accuracy
Without Feature Selection	81.58
<b>Top10</b>	<b>87.64</b>
Top20	82.11
Top30	82.92
Top40	83.10
Top50	83.58
Top60	83.10
Top70	84.09
Top80	83.59
Top90	81.84
Top100	82.44
Top110	81.81
Top120	81.94
Top130	82.39
Top140	79.94
Top150	80.15

Table VIII summarizes the detection results obtained from using the RNN (Recurrent Neural Network) algorithm for identifying Android malware. The results demonstrate that, the highest accuracy of 87.64% is achieved when utilizing the top 10 ranked permissions. Increasing the number of permissions did not result in a significant improvement in detection accuracy.

Some of the malicious instances had very few permissions,

typically 1 or 2, listed in their manifest file. Consequently, detecting these types of malicious instances with a limited number of permissions becomes challenging. Additionally, certain malware instances employ stealthy techniques by downloading malicious components during updates. Traditional static techniques that rely on permissions alone cannot identify such stealthy instances. For example, BaseBridge is a sample that downloads malicious components during updates, making it undetectable using the permissions-based approach. To address these challenges, our future work aims to incorporate dynamic features like system calls and network traffic analysis. We also intend to implement our approach on more recent and stealthier Android malware samples.

## V. CONCLUSION AND FUTURE WORK

This article suggests a method for detecting Android malware by using a static model that analyzes permissions. Given a large number of permissions available in Android, the approach aims to prioritize them based on a scalar value termed as CMC Score. This helps eliminate unnecessary permissions and improve the accuracy of malware detection. The model then uses machine learning and deep learning algorithms on the top-ranked permissions. Experimental results show that the model achieved a 97.5% accuracy rate with the top 210 ranked permissions. The authors plan to expand their analysis in the future by including additional features, such as system calls, network traffic, and CPU and memory usage, for both static and dynamic analysis.

## REFERENCES

- [1] A. S. Shatnawi, Q. Yassen, and A. Yateem, "An android malware detection approach based on static feature analysis using machine learning algorithms," *Procedia Comput. Sci.*, vol. 201, no. C, p. 653–658, jan 2022.
- [2] M. Alazab, M. Alazab, A. Shalaginov, A. Mesleh, and A. Awajan, "Intelligent mobile malware detection using permission requests and api calls," *Future Generation Computer Systems*, vol. 107, pp. 509–521, 2020.
- [3] R. Taheri, M. Ghahramani, R. Javidan, M. Shojafar, Z. Pooranian, and M. Conti, "Similarity-based android malware detection using hamming distance of static binary features," *Future Generation Computer Systems*, vol. 105, pp. 230–247, 2020.
- [4] S. G. Selvaganapathy, S. Sadasivam, and V. Ravi, "A review on android malware: Attacks, countermeasures and challenges ahead," *Journal of Cyber Security and Mobility*, vol. 10, p. 10, 2021.
- [5] A. Arora, S. K. Peddoju, and M. Conti, "Permpair: Android malware detection using permission pairs," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1968–1982, 2020.
- [6] W. Wang, Z. Gao, M. Zhao, Y. Li, J. Liu, and X. Zhang, "Droidensemble: Detecting android malicious applications with ensemble of string and structural static features," *IEEE Access*, vol. 6, pp. 31 798–31 807, 2018.
- [7] J. Feng, L. Shen, Z. Chen, Y. Wang, and H. Li, "A two-layer deep learning method for android malware detection using network traffic," *IEEE Access*, vol. 8, pp. 125 786–125 796, 2020.
- [8] A. Firdaus, N. B. Anuar, A. Karim, and M. Razak, "Discovering optimal features using static analysis and a genetic search based method for android malware detection," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, pp. 712–736, 2018.
- [9] L. N. Vu and S. Jung, "Admat: A cnn-on-matrix approach to android malware detection and classification," *IEEE Access*, vol. 9, pp. 39 680–39 694, 2021.
- [10] H. Zhang, S. Luo, Y. Zhang, and L. Pan, "An efficient android malware detection system based on method-level behavioral semantic analysis," *IEEE Access*, vol. 7, pp. 69 246–69 256, 2019.
- [11] V. Dharmalingam and V. Palanisamy, "A novel permission ranking system for android malware detection—the permission grader," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, pp. 2835–2847, 2020.
- [12] Q. Han, V. Subrahmanian, and Y. Xiong, "Android malware detection via (somewhat) robust irreversible feature transformations," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3511–3525, 2020.
- [13] Y. Zhang and B. Li, "Malicious code detection based on code semantic features," *IEEE Access*, vol. 8, pp. 176 728–176 737, 2020.
- [14] L. Yazar-Azar, L. Hamey, V. Varadharajan, and S. Chen, "Byte2vec: Malware representation and feature selection for android," *The Computer Journal*, vol. 63, no. 1, pp. 1125–1138, 2020.
- [15] R. Feng, S. Chen, X. Xie, G. Meng, S.-W. Lin, and Y. Liu, "A performance-sensitive malware detection system using deep learning on mobile devices," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1563–1578, 2021.
- [16] I. Bibi, A. Akhunzada, J. Malik, J. Iqbal, A. Musaddiq, and S. Kim, "A dynamic dl-driven architecture to combat sophisticated android malware," *IEEE Access*, vol. 8, pp. 129 600–129 612, 2020.
- [17] R. Surendran, T. Thomas, and S. Emmanuel, "On existence of common malicious system call codes in android malware families," *IEEE Transactions on Reliability*, vol. 70, no. 1, pp. 248–260, 2021.
- [18] S. Wang and et al., "Detecting android malware leveraging text semantics of network flows," *IEEE Transactions On Information Forensics And Security*, vol. 13, pp. 1096–1109, 2018.
- [19] A. Mahindru and A. Sangal, "Mldroid—framework for android malware detection using machine learning techniques," *Neural Computing & Applications*, 2020.
- [20] A. Mehtab, F. Rana, M. Nazir, A. Zeb, and T. Saba, "Addroid: Rule-based machine learning framework for android malware analysis," *Mobile Networks and Applications*, vol. 25, pp. 180–192, 2020.
- [21] H. Zhu, H. Lu, P. Liu, and Y. Zhang, "Hemd: a highly efficient random forest-based malware detection framework for android," *Neural Computing & Applications*, vol. 30, pp. 3353–3361, 2018.
- [22] K. Khariwal, R. Gupta, J. Singh, and A. Arora, "R-mfdroid: Android malware detection using ranked manifest file components," *International Journal of Innovative Technology and Exploring Engineering*, vol. 10, pp. 55–64, 2021.
- [23] J. Qiu, J. Liu, H. Zheng, Z. Zhou, and J. Lai, "A3cm: Automatic capability annotation for android malware," *IEEE Access*, vol. 7, pp. 147 156–147 168, 2019.
- [24] T.-H. Kim, J.-H. Kim, Y. Jang, H. Kim, and S. Kang, "A multimodal deep learning method for android malware detection using various features," *IEEE Transactions on Information Forensics and Security*, vol. 14, 2019.
- [25] H. Zhu, H. Lu, and P. Liu, "Droiddet: Effective and robust detection of android malware using static analysis along with rotation forest model," *Neurocomputing*, vol. 272, pp. 638–646, 2018.
- [26] J. Ribeiro, F. B. Saghezchi, G. Mantas, J. Rodriguez, and R. A. Abd-Alhameed, "Hidroid: Prototyping a behavioral host-based intrusion detection and prevention system for android," *IEEE Access*, vol. 8, pp. 23 154–23 168, 2020.
- [27] K. Khariwal, J. Singh, and A. Arora, "Ipdroid: Android malware detection using intents and permissions," in *4th World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*. London, United Kingdom: IEEE, 2020, pp. 197–202.
- [28] A. Arora and S. Peddoju, "Ntpdroid: A hybrid android malware detector using network traffic and system permissions," in *17th IEEE TrustCom*, 2018.
- [29] A. Arora, S. Peddoju, V. Chauhan, and A. Chaudhary, "Hybrid android malware detection by combining supervised and unsupervised learning," in *24th ACM MobiCom*, 2018.
- [30] S. M. Arshad, M. A. Shah, A. Wahid, A. Mehmood, H. Song, and H. Yu, "Samadroid: A novel 3-level hybrid malware detection model for android operating system," *IEEE Access*, vol. 6, pp. 4321–4339, 2018.





# Blood pressure estimation and classification using a reference signal-less photoplethysmography signal: a deep learning framework

Pankaj<sup>1</sup> · Ashish Kumar<sup>1,2</sup> · Rama Komaragiri<sup>1</sup> · Manjeet Kumar<sup>3</sup> 

Received: 26 January 2023 / Accepted: 21 August 2023  
© Australasian College of Physical Scientists and Engineers in Medicine 2023

## Abstract

The markers that help to predict the function of a cardiovascular system are hemodynamic parameters like blood pressure (BP), stroke volume, heart rate, and cardiac output. Continuous analysis of hemodynamic parameters such as BP can detect abnormalities earlier, preventing cardiovascular diseases (CVDs). However, sometimes due to motion artifacts, it becomes difficult to monitor the BP accurately and classify it. This work presents an optimized deep learning model having the capability to estimate the systolic blood pressure (SBP) and diastolic blood pressure (DBP) and classify the BP stages simultaneously from the same network using only a single channel photoplethysmography (PPG) signal. The proposed model is designed by exploiting the deep learning framework of a convolutional neural network (CNN), exhibiting the inherent ability to extract features automatically. Moreover, the proposed framework utilizes the superlet transform method to transform a 1-D PPG signal into a 2-D super-resolution time–frequency (TF) spectrogram. A superlet transform separates the peaks related to true PPG signal components and motion artifacts components. Thus, the superlet provides a robust realtime approach to accurately estimating and classifying BP using a single PPG sensor signal and does not require additional ECG and PPG sensor signals for reference. Using a super-resolution spectrogram and CNN model makes the method profitable in motion artifact removal, feature selection, and extraction. Hence the proposed framework becomes less complex for deployment on wearable devices having limited battery resources. The performance of the proposed framework is demonstrated on the publicly available larger dataset MIMIC-III. This work obtained a mean absolute error (MAE) of 2.71 mmHg and 2.42 mmHg for SBP and DBP, respectively. The classification accuracy for the SBP prediction is about 96.79%, whereas it is 98.94% for DBP. From a motion artifact-affected PPG signal, SBP and DBP are estimated. Then the estimated BP is classified into three categories: normotension, prehypertension, and hypertension, and is compared with the state of art methods to show the effectiveness of the proposed optimized framework.

**Keywords** Blood pressure · Classification · Convolutional Neural Network · Deep Learning · Hypertension · Photoplethysmography · Regression · Wearable device

✉ Manjeet Kumar  
manjeetchhillar@gmail.com

Pankaj  
er.pankaj08@gmail.com

Ashish Kumar  
akumar.1june@gmail.com

Rama Komaragiri  
rama.komaragiri@gmail.com

<sup>1</sup> Department of Electronics and Communication Engineering, Bennett University, Greater Noida, India

<sup>2</sup> School of Computer science engineering and technology, Bennett University, Greater Noida, India

<sup>3</sup> Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India

## Introduction

According to the World Health Organization, cardiovascular diseases (CVDs) are responsible for one-third of the total death rate worldwide. Around 1.28 billion adults worldwide, aged 30–79, face hypertension. Out of these, approximately 46% of adults are unaware they are suffering from hypertension. More than 720 million people were not receiving the needed treatment, as they did not feel any symptoms. People with BP usually SBP 180 mmHg and DBP 120 mmHg face symptoms like chest pain, blurred vision, and headaches. Timely identification and treatment of hypertension save people from major health conditions like heart stroke, heart disease, and kidney disease. The increase in death due to

CVDs manifests the need for realtime health monitoring devices to facilitate efficient healthcare services. Hence, the interest of the healthcare industry in exploiting photoplethysmography (PPG) signals for continuous measurement of BP has been increasing [1]. Realtime continuous monitoring of blood pressure (BP) is a preferred method to characterize the health status of a subject. It is also considered a valuable parameter for medical practitioners in diagnosing CVDs. The commonly used BP measurement devices are cuff-based mercury sphygmomanometers. Electronic sphygmomanometers are slowly becoming a commonly found healthcare device in homes. Still, sphygmomanometers are unsuited for continuous monitoring due to cuff discomfort and the prolonged interval required to obtain the BP reading [2].

The techniques used to measure BP continuously comprise arterial BP (ABP) monitoring and cuffless sensor monitoring. For accurate and reliable BP monitoring, ABP measurement is a gold-standard technique. ABP is an invasive process, hence limited to hospital settings. Cuffless monitoring is a non-invasive approach, allowing the user to monitor the BP continuously, but it requires multiple sensors; electrocardiogram (ECG) electrodes and PPG sensors [3]. Thus, recent research on single-sensor-based continuous BP monitoring is emerging.

In clinical practice, invasive and non-invasive are the two measurement techniques used to monitor BP. In an invasive measurement, BP is measured by inserting a catheter into an artery. However, the measurement using this gold-standard technique is accurate and reliable but with a high risk of infection and requires an expert medical practitioner.

In non-invasive methods, cuff type and cuff-less are the two commonly used BP monitoring methods. The cuff-based techniques are accurate, most popular, and widely used. However, cuff-based approaches cannot measure BP continuously since a two-minute pause is required to minimize the error between two immediate measurements. Subjects who undergo BP measurement feel uncomfortable due to the cuff inflation. Recently, researchers made the oscillometry cuff-based method to measure BP so simple that BP can be measured at home without the help of trained medical personnel. However, the inaccuracy with movement and cuff positioning, the inability to interface with communication systems for realtime monitoring, and the discontinuous nature have made these devices insufficient to develop realtime monitoring wearable devices. Thus, a novel cuff-less framework is highly desirable to measure BP non-invasively and continuously in realtime. Therefore, researchers are trying to implement an accurate and user-friendly approach for continuous realtime monitoring of BP [4].

Over the past decade, the role of artificial intelligence (AI) has been increasing in designing portable and wearable healthcare devices such as fitness tracking, vital sign estimation, and abnormality diagnosis in realtime. Further, with

advances in sensor technology, combining medical sensors with AI has acquired the attention of the healthcare industry to design realtime monitoring devices. PPG and ECG are the two promising techniques used with AI to monitor and diagnose CVD. Out of these sensors, PPG is using more due to its inexpensive, easy-to-use, portable, and wearable. Researchers have recently been interested in designing PPG sensor-enabled wearable devices to measure BP in realtime [5].

## Background and objectives

PPG is an optical technique used to extract the pulse wave signal due to blood volume variations during the systolic and diastolic phases of a cardiac cycle. Initially, the PPG signal-based cuffless BP estimation, pulse transit time (PTT), pulse arrival time (PAT), and pulse wave velocity (PWV) are commonly analyzed.

The pulse transit time (PTT) is initially estimated using PPG and ECG signals for continuous BP measurement. PTT is the time delay for the blood pulse to travel between two arterial sites. To estimate BP, in [6] author recorded both PPG and ECG signals synchronously for calculating PTT. Recently PTT based BP monitoring approach at home has been proposed [7].

For a cardiac cycle, PAT is the time difference between the peak of the PPG signal and the R-peak of the ECG signal. With a known distance, PWV is the velocity of the pulse wave travel through the same arterial branch with a known distance between two PPG sensors. Recently, BP estimation that uses two PPG sensors has been proposed [8]. The time difference between the two subsequent PPG waveforms is used to obtain the PWV feature. Once the PTT, PAT, or PWV is measured, the measured value is used to estimate the SBP and DBP values. Although the results obtained using these parameters are superior, these methods need both ECG and PPG devices to acquire the signal simultaneously. The measurement requires synchronization between the two devices. The devices require calibration for an individual subject and are not simple. Thus, the PPT, PWV, and PAT measurement schemes increase the complexity of the measurement and system.

The trend of using machine learning (ML) algorithms to estimate BP has recently increased to overcome these issues. Using pulse wave analysis (PWA) with ML has become a key parameter for cuff-less calibration-free BP measurement. In the PWA approach, various features are extracted from the morphological structure of the PPG signal. The features that show a good correlation with the BP are used to train the ML model.

In [9], the extracted morphological feature of the PPG signal and the PTT value is fed to a recursive neural network to obtain the BP value.

In [10], a Kalman filter and PAT feature-based framework is proposed to estimate the BP. A BP estimation model based on morphological features extracted from the PPG and ECG signal is proposed in [11]. The relevant features with high correlation with the target value are extracted and used to train the various artificial intelligence-based model to reduce the dimensionality.

In [12], to improve the BP estimation performance, a BP estimation model is proposed. The model is trained using the features containing the time, energy, and amplitude information of the ECG, PPG, and ballistocardiogram signal. A semi-classical signal analysis approach is used to extract the relevant features from the PPG signal [13]. The extracted features train the feed-forward neural network for BP estimation. However, the semi-classical approach shows superior results, requiring two sensors to collect the signal, which increases the complexity of the method.

Recently, to improve the accuracy, a fusion of PPG morphological features and demographic features has been proposed in [14] to obtain the BP. In [15], a higher-order derivative of the PPG signal-based BP estimation framework using optimized machine learning approaches is proposed. In [16], a derivative of the PPG signal and PPG signal are used to estimate the BP using the least absolute shrinkage and selection operator.

The performance of the above method depends on the PPG signal, the first and the second derivative of the PPG signal, increasing the computation complexity of the device. Therefore, the need for BP monitoring using a single PPG sensor has emerged. In [17], a technique that uses pulse wave velocity and learning-based regression is proposed to estimate the BP using a single PPG signal. The work proposed in [18] feeds the neural network with twenty-one features extracted from a single PPG signal.

The methods above show that coupling handcrafted features with machine learning models can effectively estimate BP in realtime. This method uses all the features extracted from the signal. Using too many features increases computational complexity and the probability of overfitting. Hence selecting BP-oriented features is essential to improve the performance of an algorithm.

In [19], a filter-based feature ranking algorithm is proposed to reduce the irrelevant feature extracted from the PPG signal. The features exhibit a high correlation with the target value finally selected to be fed as input to the LASSO-LSTM model for BP estimation.

The performance of the above state of art approaches depends on handcrafted features extracted from the PPG signal. The morphological structure of a PPG signal depends on the change in cardiovascular characteristics, which varies from person to person. Thus, extracting handcrafted features from a PPG signal may provide erroneous features. In addition, the optical properties of skin, tissue, and motion

artifacts could affect the morphological structure of the PPG signal.

Henceforth, the dependence of BP on these handcrafted features extracted from the PPG signals provides inaccurate estimation when tested in realtime.

Thus, to avoid the need for handcrafted features for model training, recently, researchers have taken full advantage of deep learning models for automatic feature extraction, regression, and classification. The deep learning method for BP estimation mainly consists of feature-based and end-to-end learning approaches. In a feature-based approach, the spectral and temporal features of an input PPG signal are manually extracted and fed to the model to estimate systolic blood pressure (SBP) and diastolic blood pressure (DBP). A 2-D spectrogram is input in an end-to-end learning approach and processed to implicitly obtain the features to estimate SBP and DBP.

In [20], a method is proposed to construct a deep belief network-restricted Boltzmann machine to predict the BP. This work shows the potential of the PPG signal for non-invasive and continuous BP measurement.

In [21], a long-term recurrent convolutional network-based model is introduced to estimate the BP. A cascaded approach of CNN and long short-term memory (LSTM) network to estimate the BP is proposed in [22]. In [23], a receptive field parallel attention shrinkage network-based BP estimation framework is proposed. This work uses the multi-scale large receptive field convolution module to capture the long-term time dynamics of the PPG signal.

A single-channel PPG and ECG signal-based BP estimation approach is proposed in [24]. This work proposes an LSTM and multi-scale convolution network-based approach to estimate BP. In [25], a transfer learning-based BPCNN framework is proposed to estimate BP to avoid the need for handcrafted feature-based estimation. A recurrent gated unit maintains the temporal dependencies among the feature extracted through the CNN layer. A visibility graph of PPG signal-based training of deep neural networks is proposed in [26] to reduce the computational training time of the model.

In [27], PPG and ECG signals are acquired from a sensor installed in a wearable device. Then the features are extracted from the two signals to train the model.

Recently to replace the conventional handcrafted feature extraction process in [28], a CNN and LSTM network-based approach has been proposed. In [29], a U-net deep learning framework is proposed to estimate the BP using a PPG signal. The U-Net framework uses a skip-connection approach to reduce feature information loss by preserving detailed information. However, the skip connections approach used in the U-Net architecture does not consist of temporal information of the PPG signal. Thus, to improve the performance of the U-Net architecture for BP estimation, [30] introduces

a self-attention layer with U-Net to capture the temporal information.

In [31], a pattern-fusion method that utilizes a cardiovascular coupling model is proposed to enhance the adaptability and reliability of the BP estimation method. A continuous wavelet transforms (CWT) TF analysis-based approach is proposed in [32]. CWT localizes the PPG signal well in time and faces tradeoffs in frequency resolution with increased frequency. Thus, with increasing frequency, the CWT provides an excellent temporal resolution throughout the spectrum but degrades in frequency resolution. Thus, to overcome the TF uncertainty limitations of single-order CWT, a multi-order wavelet named superlet transform is proposed in this work.

Most of the works found in the literature require signals from two sensors to estimate BP. Additionally, most of the work reported in the literature used raw corrupted PPG signals to train the deep neural network. The feature extraction and model training using raw PPG signal provides inaccurate realtime BP estimation as the PPG signal acquired using wearable device corrupted with motion artifacts.

Moreover, most of the works reported in the literature analyze a PPG signal either for BP estimation or classification. The estimation and classification of BP using a wearable device in realtime enables the users to know their BP condition and can also become a potential early warning system.

Thus, this paper proposes a novel BP estimation and classification framework for PPG signals using a convolutional neural network with a superlet transform-based super-resolution spectrogram.

The highlight of the proposed work is as follows:

1. The proposed work not only estimates the accurate BP value but also simultaneously classifies the obtained BP value into three different BP categories based on the Joint National Committee (JNC7), thus helping the user to know the classification status of their BP.
2. The proposed framework does not depend on the additional sensor reference signal to suppress motion artifacts and estimate BP in realtime. Hence, the computational complexity is reduced.
3. With the help of a super-resolution spectrogram, the proposed framework separates the motion artifacts and noise components from the acquired PPG signal that can provide the actual condition of cardiac health.
4. The proposed framework optimized the CNN model combined with a super-resolution TF spectrogram by varying the value of different model parameters. The model parameters at which the model performs best are finalized to compare the results.

## Methodology

The flowchart of the proposed framework to estimate real-time SBP and DBP value from a PPG signal is illustrated in Fig. 1. The acquired PPG signal and the simultaneously acquired ABP signals are the inputs to the framework. The ABP signal is used to provide the reference SBP and DBP values during the training and testing phase of the proposed framework. Input PPG and the ABP signal are segmented into an overlapping window containing three hundred samples in each window. Each segmented PPG signal is transformed into a super-resolution spectrogram. The obtained spectrogram after that was used as input to the CNN model during training and testing. The estimated SBP and DBP values are then compared to the reference SBP and DBP values to analyze the performance of the proposed framework. The same model uses the estimated SBP and DBP values to classify the BP into three categories. The steps involved in this process are shown in the block diagram in Fig. 1 and described below.

## Dataset

The first stage of the proposed block diagram is related to the data collection stage. The multiparameter intelligent monitoring in intensive care (MIMIC) waveform database MIMIC-III [33], a publicly available dataset, is used in this work. The MIMIC-III dataset contains multiple parameter recordings of more than thirty thousand subjects from an intensive care unit. This database is extensively used in BP estimation and classification research. The database comprises more than 67,000 recordings, including many biological signals such as fingertip PPG signals and ABP and ECG signals recorded simultaneously at a sampling frequency of 125 Hz.

## Pre-processing stage

The recordings containing PPG and the ABP signal are considered for analysis from the MIMIC dataset. After extracting the records consists both PPG and the ABP signal, all recordings are checked manually. The recording having missing values (NaN), flat peaks, and flat lines are discarded from the dataset.

5,73,603 samples are acquired using this process from the MIMIC dataset. For continuous monitoring and prediction of the BP state, the complete dataset is segregated into a number of windows. The acquired PPG and the ABP signals from the MIMIC database are used as the predicted parameter and the target or true parameter.



The SBP and DBP to provide target BP (reference value) are calculated for a 2.4-s duration from ABP signals, resulting in the 300-sample window.

The first window ( $k=1$ ) consists of samples numbered from 1 to 300, as shown in Fig. 2, which are considered to estimate the BP. After the first window, each consecutive window overlaps the immediate prior window by  $(T-w)$  samples. Here  $w$  is the step size of length 1.6 s, and  $T$  is the total number of samples in a window. Thus, the last 100 samples from the first window are concatenated with 200 new samples to obtain 300 samples in the next window [34]. Therefore, BP estimation includes new data for 1.6 s in each consecutive window. The pictorial representation of window designing is represented in Fig. 2.

With the overlapping procedure, 3204 (2.4-s) windows are designed, which are used to train and test the proposed deep neural network framework.

Simultaneously recorded PPG datasets are segregated similarly with consecutive windows of 2.4-s duration with a 1.6-s shift. The last 0.8-s data of the  $(k-1)$ th window is replicated as the first 0.8-s data of the  $k$ th window. Figure 2 shows the PPG and ABP signals segregated into windows of three hundred samples with an overlapping of one hundred samples.

Different noises affect all the recorded PPG signals, such as baseline wandering, motion artifacts, and other noises. For reliable and accurate estimation of BP, separating the noises from the PPG signal is necessary. Thus, data pre-processing is the most vital. In the pre-pre-processing stage, the segmented 2.4-s window PPG signal data is subjected through a 4th-order Butterworth bandpass filter. The cutoff frequencies of the filter are chosen to be 0.5 Hz and 8 Hz to filter the baseline wandering (values below 0.5 Hz) and other undesired high-frequency noises above 8 Hz.

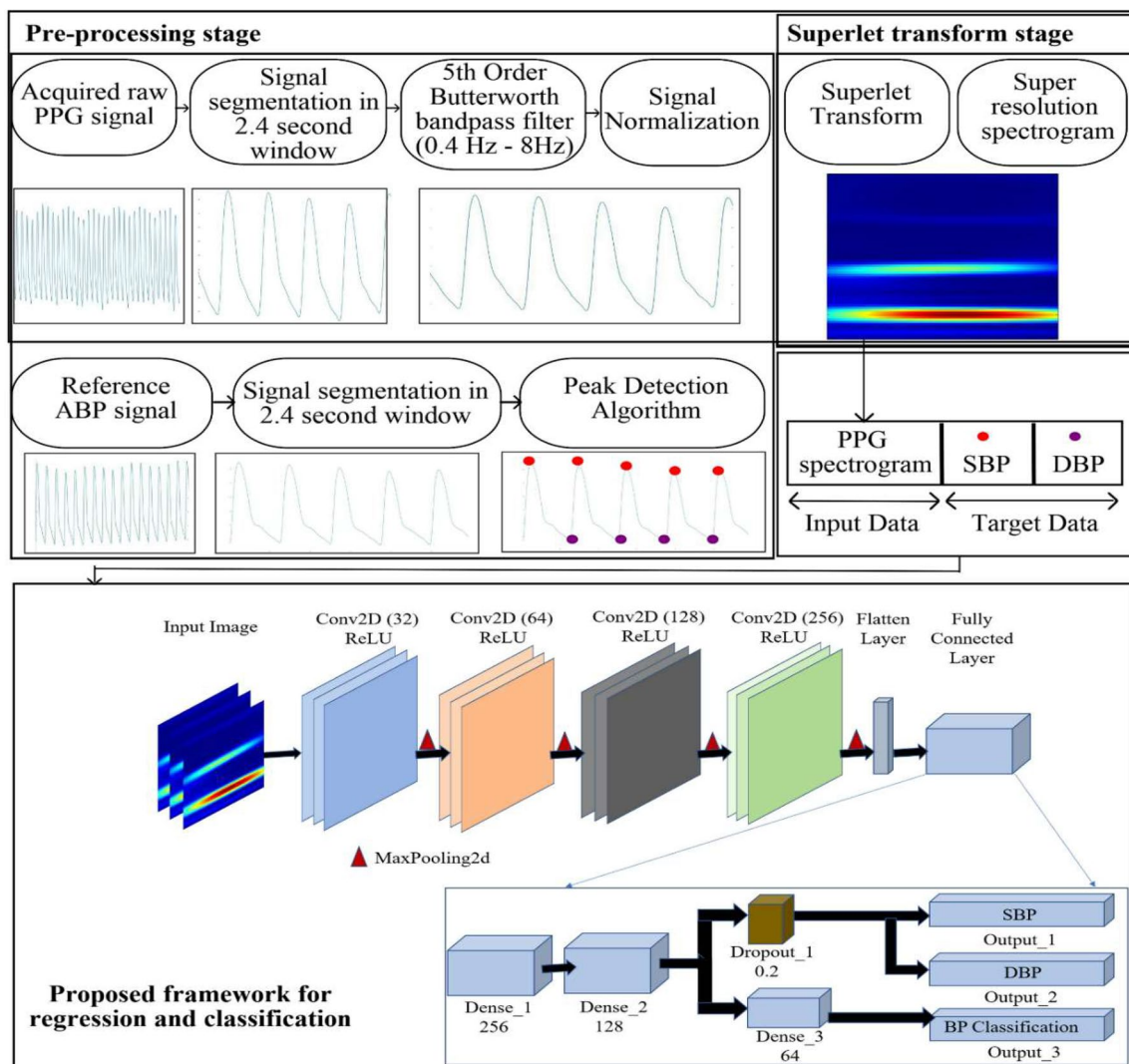


Fig. 1 Proposed framework for regression and classification using PPG signal

The 0.8-s window overlapping between two consecutive windows is introduced for continuous estimation in subsequent windows and to avoid the chance of missing information at the boundaries of each window. The size of the window and the overlapping values are selected empirically. Finally, the filtered data is normalized using Eq. (1), where  $x_k$  represents the  $k^{\text{th}}$  PPG signal window.

$$X_{\text{normalize}(k)} = \frac{x_k - \min(x_k)}{[\max(x_k) - \min(x_k)]} \quad (1)$$

The normalized  $k^{\text{th}}$  PPG signal window and the  $k^{\text{th}}$  ABP signal, SBP, and DBP value are input and target data to train the proposed CNN model.

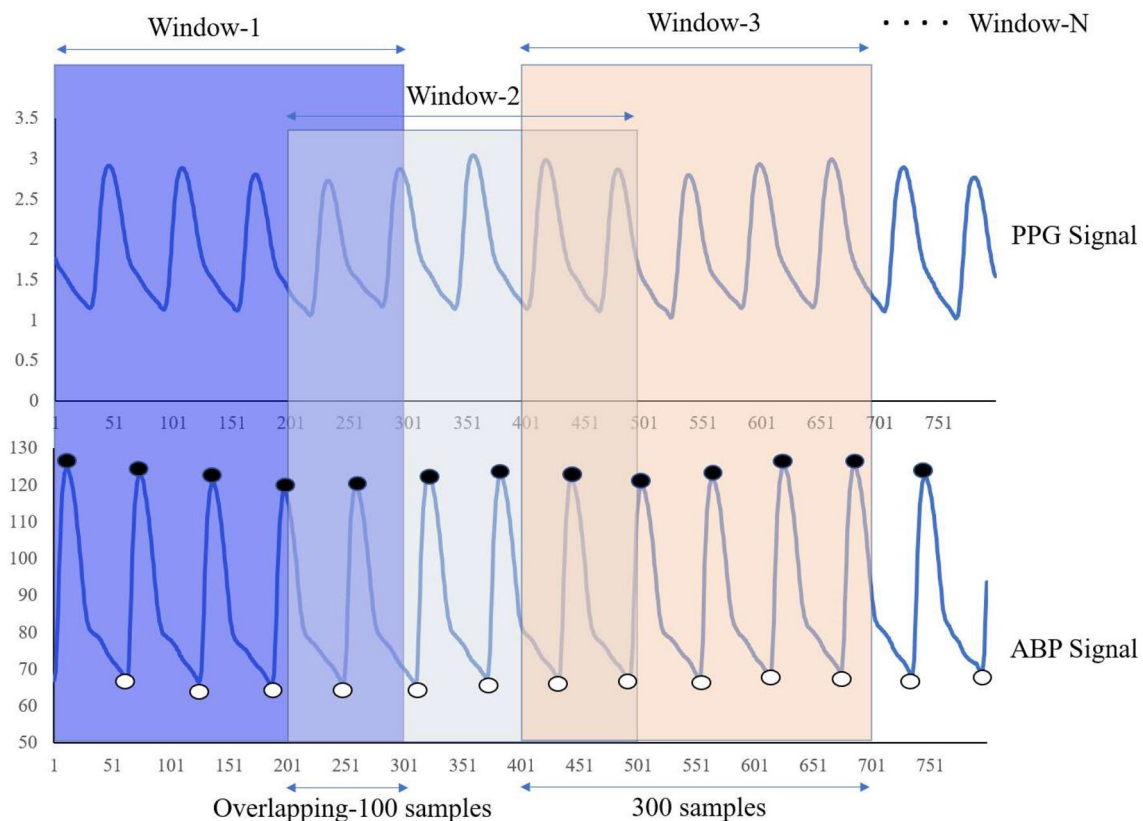
Figures 3 and 4 show the histogram plot of all reference SBP and DBP values used as true values in the proposed framework obtained using the ABP signal. The histograms in Figs. 3 and 4 show that the pre-processed dataset contains all three BP classes per the JNC7 standard. All values are distributed widely from the lower to the highest range of SBP and DBP.

## Reference SBP and DBP estimation

The segmented 2.4-s ABP window is passed through the peak detection block to obtain the reference SBP and DBP values. For all the windows, the algorithm provides the SBP and DBP values. The reference SBP value of a window is calculated by computing the mean of all synchronous wave peaks that occur in that window, shown by a black dot in Fig. 2. DBP is computed by inverting the ABP signal to detect the wave trough of each pulse in a window. The mean synchronous wave trough of a window shown by a white dot is used to calculate the reference DBP value.

## PPG signal transformation

The accurate and reliable estimation and classification of BP in realtime is the key feature of the proposed framework compared to the existing methods. In a recorded PPG signal, noise and motion artifacts signal components and clean PPG signal have overlapping frequency spectra between 0.4 Hz to 8 Hz. Hence, separating noise and motion artifacts from a recorded PPG signal is challenging, and the estimation of BP can be inaccurate. The superlet transform can represent the



**Fig. 2** Formation of consecutive windows with the number of samples equal to 300

motion artifact contaminated PPG signal in TF by splitting the PPG components related to BP and motion artifacts, thus providing a dependable approach for accurate estimation of BP in realtime [35]. Superlet transform involves multiple wavelets with a fixed center frequency to provide super-resolution by utilizing different cycles. The mathematical representation of a superlet is given by Eq. (2).

$$SL_{f,g} = \left\{ \psi_{f,d} \middle| d_i = d_1, d_2, \dots, d_g \right\}, \quad (2)$$

where  $g$  represents the number of wavelets used in the superlet (order). The SLT improves frequency resolution at higher frequencies by increasing the order of the superlet.  $d_1$  is the first wavelet base cycle used in the analysis. In superlet, the value of different cycles that define the order of the wavelet can follow a multiplicative or additive rule.

With multiple wavelets of different cycles, the response of Superlet (SL) to the pre-processed PPG signal  $S[n]$  is given by the geometric mean (GM) of the individual wavelet response represented in Eq. (3).

$$R[SL_{f,g}] = \sqrt[g]{\prod_{i=1}^g R[\psi_{f,d_i}]} \quad (3)$$

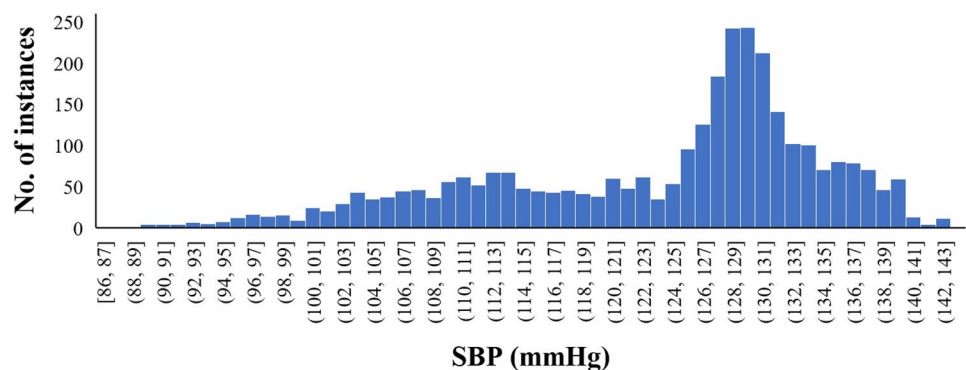
$R$  represents the response of  $i$ th wavelet of the given signal PPG signal.

With wavelet order greater than one, superlet transform starts with the lower order. It increases the order as a function of frequency and the combined response using Eq. (3). Multiple order wavelet helps to extract and magnify TF information from the PPG signal.

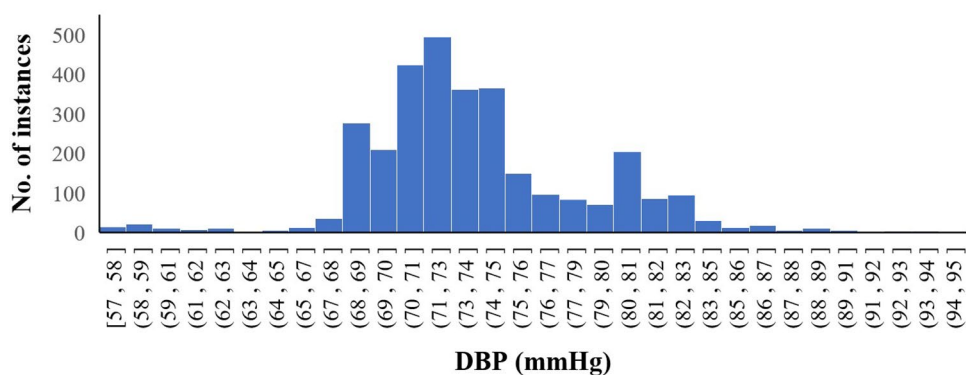
With multiple wavelets, the PPG signal components with the same feature will show the same TF information with multiple-order wavelet-based superlet transform. A combination of information retrieve with multiple wavelets represents a bright red peak. This bright red color peak indicates a large amplitude, and this information is used to extract BP. The component which does not have the same feature with multiple wavelets will be represented by dark blue and yellow colors, indicating a low amplitude. Thus, the superlet transform can potentially separate the motion artifacts component and desired BP component. A 2-D representation of the SLT spectrogram shown in Fig. 5 indicates the separation of the BP signal component from a motion artifact corrupted PPG signal.

The TF spectrum corresponds to the superlet transform of the different spectral elements in the spectrogram of a signal; it determines the rate of change of the signal spectral components over time. Changes in blood volume modulate the PPG signal; hence BP information is concentrated as lobes in the spectro-temporal spectrum. The BP components are encoded in the superlet transform spectral lobes, and artifacts appear in other regions, thus improving the robustness

**Fig. 3** Histogram plot of SBP signals indicating all three BP ranges, namely normotension, prehypertension, and Hypertension

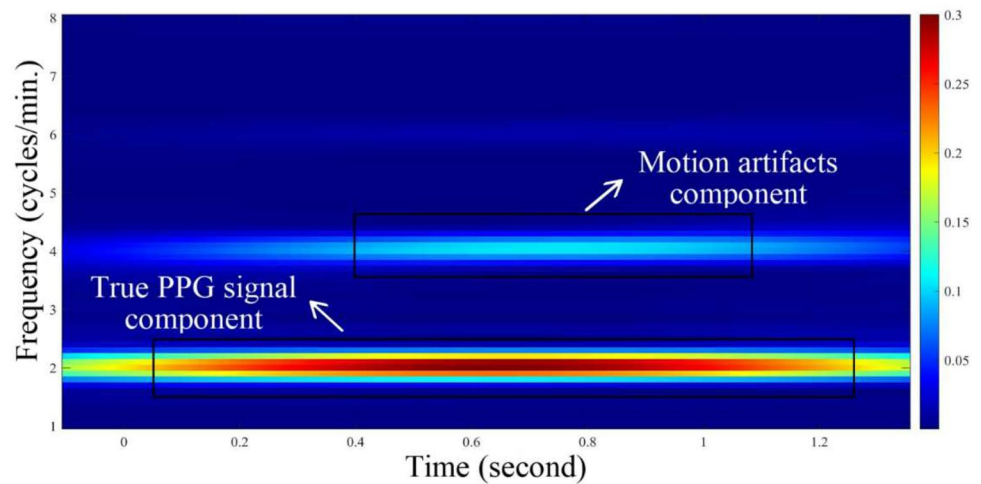


**Fig. 4** Histogram plot of DBP signals indicating all three BP ranges, namely normotension, prehypertension, and Hypertension





**Fig. 5** The 2-D TF super-resolution spectrogram of the PPG signal



of BP estimation to artifacts. Due to the nature of the PPG signals component, the spectral content of the PPG signal will change at different rates during different physical motions. Thus, multi-resolution superlet transform-based spectral domain analysis improves separability between desired signal and noise. Multiple spectrograms generated with multiple wavelet orders can be combined by their geometric mean, which is optimal for an entropic criterion and can increase the resolution. Hence, the final output data from the spectrogram contains the intensity data as a function of time and frequency. One graphical representation of the data is shown in Fig. 5.

The color bar on the right side of Fig. 5 indicates the different color intensity magnitude values. The darker or hotter color will correspond to the maximum intensity magnitude value, and the lighter color will correspond to the lower intensity magnitude value. The time point of the darker lobe corresponds to the center of the wavelet, where it is aligned with the data. Suppose the convolution of the wavelet with the signal gets a negative value. In that case, the output of the wavelet spectrum gets a deep blue color. If more positive, it will get a deep red color lobe in the superlet transform spectrogram.

Thus, the rate-of-change of the PPG TF spectral components is represented by a high-intensity lobe centered on a frequency corresponding to the BP.

Artifacts appear in other regions, thus improving the robustness of BP estimation by separating the artifacts.

In the proposed work, the TF spectral component of the PPG signal with true BP value is fed as input to the model during training. The CNN model updates its weight and bias by correlating the TF spectral component of the PPG signal with its true BP value. The updated weight and bias

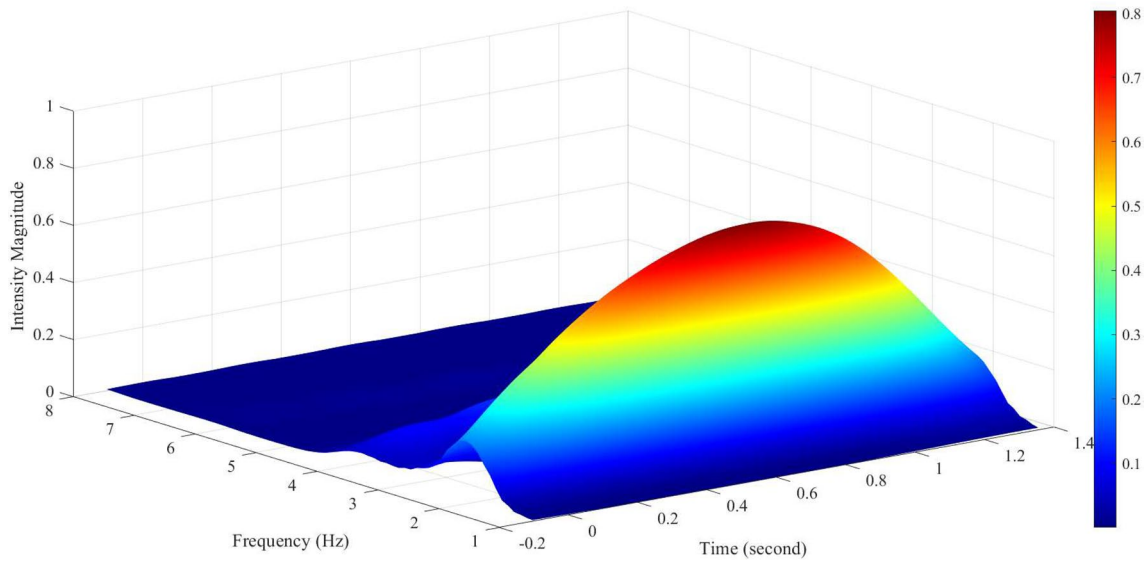
of the CNN model are used directly in the testing phase for prediction when an unseen dataset is fed into it. The SBP and DBP prediction values from the trained model are verified using mean absolute error (MAE) performance matrices, the mean value of the difference between true BP and predicted BP.

The core idea of the proposed superlet transform-based spectrogram is to distinguish the true PPG signal and motion artifacts component in the TF spectrogram, helping the model to correlate the true PPG components with reference BP effectively.

The 2D-CNN layer automatically learns the features from the 2-D superlet spectrogram. If the input spectrogram data is sufficient, then the deep CNN model can predict the BP reliably and accurately.

Superlet transform provides a super-resolution 2-D spectrogram of a PPG signal. In the 2-D spectrogram, the spectral components of the PPG signal are represented by a high-intensity lobe located on a frequency corresponding to the BP. The superlet high-resolution spectrogram is composed of pixels that describe the amplitude associated with a range of frequencies at a specific time step. Figure 6 represents the 3-D view of the super-resolution spectrogram. The  $x$ -axis represents the time information, the  $y$ -axis represents the frequency, and the  $z$ -axis represents the normalized amplitude of the PPG signal.

The brighter the pixel is, the higher the energy of the associated frequency. Thus, the time step, frequency, and amplitude of a high-intensity peak collectively identify features in the PPG signal through the superlet spectrogram. The BP estimation process using the proposed framework is shown in Algorithm 1.



**Fig. 6** The 3-D representation of the super-resolution spectrogram

---

**Algorithm 1:** Estimation of SBP and DBP by optimized deep neural network

---

Input: PPG and ABP signal

Output: SBP and DBP

BP detection Process:

1. Begin
2. Divide the PPG and ABP signal into a 2.4-second window containing  $T = 300$  samples.
 
$$S_{PPG}^k = (S_1((k-1) \times w) + 1 \dots S_1((k-1) \times w) + T),$$

$$S_{ABP}^k = (S_2((b-1) \times w) + 1 \dots S_2((b-1) \times w) + T),$$

Where  $w$  is the step size equal to 200 samples

3. for  $k = 1$  to  $N$ , do
4. Obtain the pre-processed PPG signal  $S_{PPG}^k$
5. Assign Morlet wavelet parameter values: the central frequency ( $f$ ), wavelet order ( $g$ ), the number of cycles ( $d$ ), and the time spread parameter ( $B_d$ ) in second
6. Obtain the geometric mean (GM) of the multiple wavelets of different cycles using the relation

$$R[SL_{f,g}] = \sqrt[g]{\prod_{i=1}^g R[\psi_{f,d_i}]}$$

8. Obtain the Superlet transform time-frequency 2-D spectrogram using step 6
  9. Pass the  $S_{ABP}^k$  signal from peak detection block to calculate reference SBP and DBP value
  10. end for
  11. Train the deep neural network using the  $N$  window spectrogram with reference SBP and DBP value
  12. Return performance metrics
-

## Deep neural network

The deep CNN model automatically learns the features from the super-resolution TF spectrogram and correlates these features with true SBP and DBP values. Most state-of-the-art techniques require separate training to estimate SBP and DBP from the PPG signal. Hence separate models to estimate SBP and DBP in realtime are required. As a result, the computational complexity of the system increases.

The proposed framework uses a single training model to estimate SBP and DBP and classify the BP state simultaneously. Moreover, the estimation of BP does not depend on any manual feature selection and extraction steps, thus reducing the computational complexity of the system. As a result, the model becomes a more generalized method for realtime BP estimation. Further, the proposed framework obtained a MAE of SBP and DBP is 2.71 mmHg and 2.42 mmHg for 3204 windows used for training and testing the proposed model. Thus, the proposed framework can potentially enable realtime implementation in healthcare applications.

## Training

This section discusses the training process of the proposed framework consisting of the BP estimation and classification stage. To evaluate the efficacy of the proposed work, the performance metrics used are MAE and the root mean square error (RMSE). This section provides the performance of the proposed model to facilitate comparison with the cuffless BP estimation algorithm.

This study proposes a 2-D CNN-based cuff-less BP estimation technique from the PPG signal. This work introduces a model parameter setting to obtain high-performance accuracy and minimal error. The first step in parameter optimization is the ratio of training, validation, and test datasets out of the complete dataset. This step prevents the model from overfitting. The final accuracy

of the validation and test dataset is analyzed to determine that the parameter of the model is accurately optimized.

For each super-resolution TF spectrogram, the spectral and temporal information is extracted by a deep neural block consisting of four convolutional layers. Every convolutional layer is followed by the ReLU activation layer, which reduces the training time of the network. For dimensionality reduction output of each layer is passed through the max pooling layer.

The 2-D array output from the last convolutional layer must be flattened to match the fully connected layer. The fully connected layer is used to analyze the extracted feature for prediction. The fully connected layers are used as a final stage of the deep neural network, changing the dimensionality of the output from the preceding layer per the model objective.

The final output of the network is applied to the ReLU activation layer. The output of the ReLU activation layer is used to predict SBP and DBP for the given 2.4-s super-resolution window.

An adaptive learning rate process is used as a model optimization parameter that changes during training to increase the model estimation and classification performance and speed up the training process. The model is trained with an initial learning rate of  $10^{-4}$  and drops with a factor of  $10^{-1}$ . The batch size is also used as an optimized parameter while finalizing the best model. A batch size of 32 is fixed in this work. An early stopping criterion is used to generalize the model. The early stopping helps to avoid overfitting the training data when no improvement in performance metric is observed during any consequent five epochs.

The deep neural network parameters, such as the number of convolutional layers, convolutional layer filter size, selection of pooling layer, the number of epochs, initial learning rate, and batch size, are valuable parameters while optimizing the proposed framework. Algorithm 2 shows the steps to train and test the proposed regression and classification model.

---

Algorithm 2: Training the proposed Regression + Classification deep neural network framework.

---

Require: Training  $S_{PPG-Train}^k$  superlet spectrogram, Test  $S_{PPG-test}^k$  superlet spectrogram.

Ensure: Optimized neural network model for BP estimation and classification

---

1. Initialize the model parameter.
  2. # Training Phase
  3. Load training  $S_{PPG-train}^k$  superlet spectrogram.
  4. Update the parameter of the model with the loss function of the equation
  5. # Testing Phase
  6. Load test  $S_{PPG-test}^k$  superlet spectrogram.
  7. Load trained model and test with an unseen dataset.
  8. Compute the loss function and accuracy of the trained model with the test dataset.
-

## Evaluation and Statistical Analysis

The model estimation performance is evaluated using RMSE and MAE. MAE is the mean of the absolute difference between the true and predicted values and is calculated using the relation mentioned in Eq. (4).

$$MAE = \frac{1}{N} \sum_{k=1}^N |BP_{true}(k) - BP_{predicted}(k)| \quad (4)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{k=1}^N (BP_{true}(k) - BP_{predicted}(k))^2} \quad (5)$$

For regression tasks, RMSE is the most commonly used performance metric. RMSE, the squared difference mentioned in Eq. (5), helps identify how large the difference is between the squares of the true value and the predicted value.

The effectiveness of the proposed framework to estimate BP is validated by comparing the results with the BP estimation standard criteria introduced by the British Hypertension Society (BHS) and the Association for the Advancement of Medical Instrumentation (AAMI). The standard percentage criterion for a specific BHS grade is mentioned in Table 1. A BP estimation device must achieve a B-grade for SBP and DBP per BHS standards.

According to the AAMI standard, the proposed BP estimation framework is validated only if the MAE and standard deviation (SD) lie within the standard range, as shown in Table 2. There should be at least 85 subjects before computing the AAMI standard.

Besides SBP and DBP estimation, the same model can classify the State of BP in realtime. The BP can be classified from the estimated values of SBP and DBP. Based on the Hypertension guidelines from the American College of Cardiology/American Heart Association (ACC/AHA) 2017, BP is classified into three categories: Normotension, Prehypertension,

**Table 2** Performance of the proposed framework with AAMI standard

	MAE (mmHg)	SD (mmHg)	Number of subjects
AAMI standard			
BP	$\leq 5$ mmHg	$\leq 8$ mmHg	$\geq 85$
Proposed work			
SBP	2.71	3.95	1557
DBP	2.42	3.29	1557

and Hypertension. Figure 7a shows the range of SBP and DBP values for the three categories of BP. In Fig. 7a, the white box shows the range of BP not present in the given dataset. Figure 7b shows the number of windows assigned to each class of BP.

To analyze the performance of the proposed framework for BP Classification, the evaluation metrics used are accuracy, specificity, precision, sensitivity, and F1-score. These metrics are predicted by using true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), represented by Eqs. (6) and (7).

$$Accuracy(\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (6)$$

$$Specificity(\%) = \frac{TN}{TN + FP} \times 100 \quad (7)$$

$$Precision(\%) = \frac{TP}{TP + FP} \times 100 \quad (8)$$

$$Sensitivity(\%) = \frac{TP}{TP + FN} \times 100 \quad (9)$$

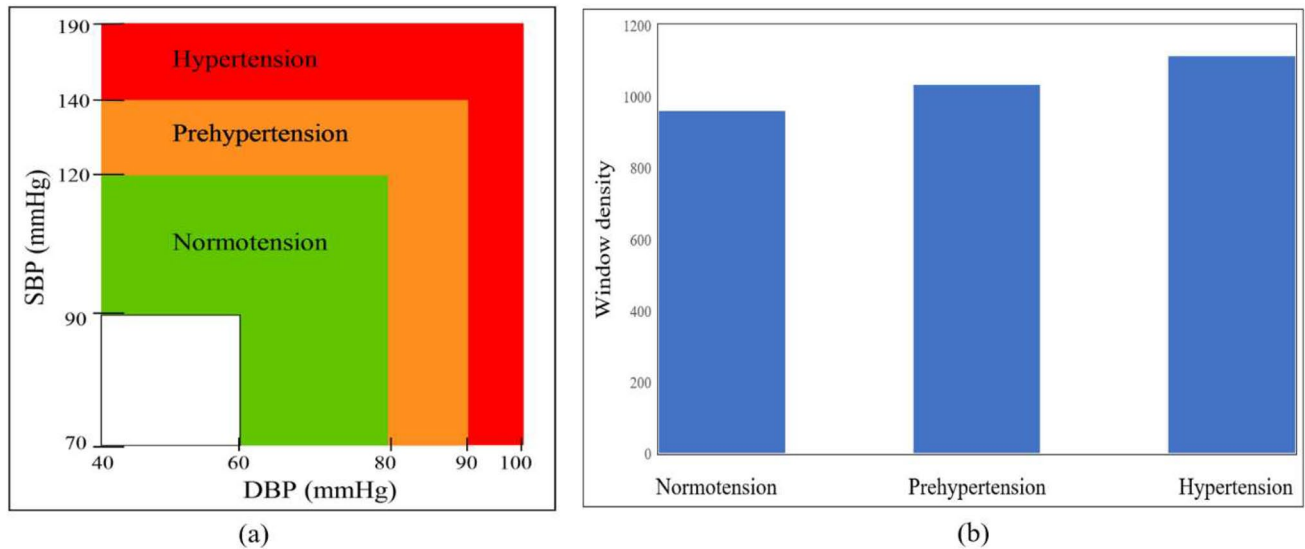
$$F1-score(\%) = \frac{2 \times precision \times sensitivity}{precision + sensitivity} \times 100 \quad (10)$$

**Table 1** Performance comparison of the proposed framework with BHS standard

	Cumulative error percentage		
	$\leq 5$ mmHg (%)	$\leq 10$ mmHg (%)	$\leq 15$ mmHg (%)
BHS Standard			
Grade A	60	85	95
Grade B	50	75	90
Grade C	40	65	85
Proposed work			
SBP	76.89	98.02	99.49
DBP	90.88	97.63	99.75

## Results

The results obtained with the proposed framework are validated using 1557 subjects. Table 1 compares the BHS standard and Table 2 for the AAMI standard. Out of the total 3204 windows, the SBP-MAE of 76.89% and 90.88% of windows of DBP-MAE lie below the  $\leq 5$  mmHg. For all three parameters of cumulative error percentage, the obtained SBP and DBP of the proposed work meet the requirements for Grade A. The proposed BP estimation framework obtained MAE  $\pm$  SD of  $2.71 \pm 3.95$  mmHg and  $2.42 \pm 3.29$  mmHg for SBP and DBP estimation, respectively. MAE and SD values



**Fig. 7** (a) Hypertension criterion as per JNC7, (b) Histogram plot representation of window's density of each class

for SBP and DBP lie within the range of AAMI standards. Thus, the proposed work achieved the AAMI and Grade-A in the BHS standards for both SBP and DBP.

All the obtained MAE and SD values lie within the BHS and AAMI standard range. Furthermore, the dropout rate is an important hyperparameter used to reduce the proposed model complexity and prevent the model from overfitting. Tuning of the dropout rate dramatically affects the BP estimation performance of the proposed model. In this work, the dropout rate is selected from 40 to 10% with a stride by a factor of 5%. Table 3 lists different dropout rates, and the results are obtained for the proposed CNN model. As shown in Table 3, the proposed framework can obtain the best BP estimation performance when the dropout rate equals 0.2. The proposed framework achieves the best MAE of 2.71 mmHg and 2.42 mmHg and RMSE of 3.97 mmHg and 3.58 mmHg for SBP and DBP estimation, respectively. Therefore, the dropout rate is fixed at 0.2 in the proposed framework.

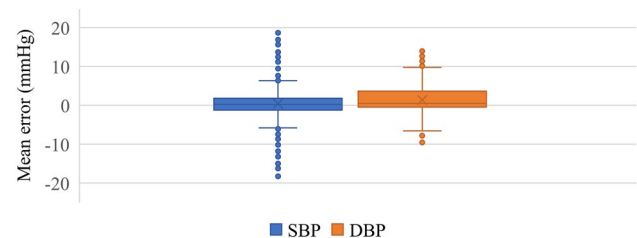
Figure 8 shows a box plot of the mean error value of SBP and DBP obtained using the proposed method.

The boxes and median value of SBP and DBP show that the MAE value for maximum windows lies under the acceptable range of BP estimation mentioned in Table 2, thus, demonstrating the potential of the proposed framework.

From an application aspect, it is more valuable to classify the State of BP of a user in realtime instead of providing the values of SBP and DBP. Table 4 lists the classification performance for each hypertension class corresponding to SBP and DBP values in terms of specificity, precision, sensitivity, and F1-score. The overall accuracy for the SBP prediction is about 96.79%, and 98.94% for DBP.

**Table 3** BP Measurement Performance of the Proposed Framework with different dropout factors

Dropout rate	Index	SBP	DBP
0.4	MAE	11.05	7.52
	RMSE	13.83	9.85
0.35	MAE	9.45	5.54
	RMSE	11.87	7.98
0.3	MAE	7.23	5.47
	RMSE	10.89	7.21
0.25	MAE	4.32	4.02
	RMSE	7.56	6.84
0.2	MAE	2.71	2.42
	RMSE	3.97	3.58
0.15	MAE	3.89	3.43
	RMSE	5.43	5.26
0.1	MAE	4.41	4.06
	RMSE	6.21	5.89



**Fig. 8** Box plot of mean error for SBP and DBP



## Discussion

Over the past decade, various algorithms have been proposed to estimate cuffless BP. Most techniques in the literature require data from PPG and ECG sensors for BP estimation using PTT and PAT. Despite providing accurate results, these algorithms cannot be accepted in medical practice as the accuracy of these algorithms depends on individual physiological conditions. Furthermore, these methods require two sensors. Table 5 presents a performance comparison with various works found in the literature. Compared with the state-of-the-art, factors like calibration, evaluation matrices, and dataset size make it hard to compare these techniques.

Further, the publicly available MIMIC-II and MIMIC-III datasets have been used tremendously in the literature. Thus, in this study, for a fair comparison with the literature, Table 5 lists the calibration-free techniques tested on the MIMIC dataset, and MAE, the standard deviation of absolute errors (SDAE), is used as the performance metric. The  $\text{MAE} \pm \text{SDAE}$  values of the proposed framework for SBP and DBP are  $(2.71 \pm 3.29)$  and  $(2.42 \pm 3.94)$ , respectively. The obtained MAE and the SDAE value for both SBP and DBP outperform other methods.

Table 5 indicates that MAE and SDAE value obtained with the proposed work is lower than the prior works except for the recent work proposed in [36, 37, 41, 46–48]. However, work present in [36, 37, 41, 46, 47] requires an additional ECG signal with a PPG signal to extract the feature. The need for additional sensor signal increases the overall system complexity and energy consumption. The work reported in [48] requires a manual feature extraction step for model training. The accuracy and reliability of the model depend on the value of the feature computed in realtime, as the morphological structure of the PPG signal depends on the nature of motion artifacts and disease specification. Further, some of the recent work not listed in Table 5 based on BP estimation using PPG signal are proposed in the literature. In [50], a PAT analysis-based approach is proposed for cuffless BP estimation. This work meets the AAMI standard for DBP but fails to attain the same for SBP. Moreover, using two sensor signals for analysis makes this approach complex. The work proposed in [51] used a single PPG sensor to

overcome the need for two sensors and uses the time domain features extracted from a PPG signal.

This method achieved an MAE of 4.51 mmHg for SBP and 2.6 mmHg for DBP. However, the accuracy of this work depends on the accurate extraction of the first and second derivatives of the PPG signal.

In realtime analysis, the correct extraction of features is challenging as the signal shape varies with age, disease, and body movement. In [38], a transfer learning-based approach is proposed to estimate BP using the PPG signal, first derivative PPG, and second derivative PPG signal. The signal is processed through five ResNet blocks and the GRU layer. The method in [38] has high computational complexity and low performance compared to the state of the art work reported in the literature. The work proposed in [52] incorporates many subjects for model training. Still, the use of PPG signal, first derivative, and second derivative PPG signal increase the computational complexity of the system, thus not feasible for wearable device development. In [53], two cascaded U-Net architecture-based BP estimation approach is proposed. The computational complexity demanded by this approach makes them impractical to implement in a wearable device for realtime analysis. The work proposed in [54] Fourier transform of PPG signal as input to the CNN model. The model used both PPG and ECG signals to train the CNN model, thus increasing the complexity of the model. Thus, from the perspective of realtime analysis using a wearable device, this work proposes a single PPG sensor and single PPG signal-based BP classification and estimation approach, thereby reducing the computational complexity of the system.

Analysis of realtime acquired PPG signals using wearable devices can provide inaccurate prediction, as the signal is affected by motion artifacts. Thus, for accurate analysis and prediction from a PPG signal, the separation of motion artifacts components and the true PPG signal components is necessary. Therefore, this work used the superlet transform-based TF spectrogram of PPG signals to train the deep neural network for estimation and classification.

The work reported in the literature used raw corrupted PPG signals to train the deep neural network. Notably, a deep neural network has the potential to extract relevant features automatically and map these features with the target class. It is assumed that training a model with a corrupted

**Table 4** Classification performance of the proposed framework

Class	SBP				DBP			
	Pre%	Sen%	Spec%	F1-Score	Pre%	Sen%	Spec%	F1-Score
Normotension	97.3	97.3	97.4	97.3	98.7	97.4	98.7	98.1
Prehypertension	97.4	98.7	98.1	97.4	97.4	98.7	98.1	97.4
Hypertension	96.1	94.9	98	95.4	95.5	98.6	97.4	95.5



**Table 5** Comparison of result analysis of proposed work with the state of art work

Author	Sensor (number of Subjects)	Input data pre-processing for featurig	Methodology used for regression	MAE $\pm$ SDAE	
Tanveer [36] 2019	PPG/ECG (39)	Bandpass by TQWT, feature extracted from PPG and ECG signal	ANN-LSTM	SBP DBP	$1.10 \pm 1.56$ $0.58 \pm 0.85$
Yan [37] 2019	PPG/ECG (604)	Segmented data into a 10-s window	CNN	SBP DBP	$3.09 \pm 2.76$ $2.11 \pm 2.0$
Slapnicar [38] 2019	PPG (510)	4th order bandpass filter, PPG, first derivative PPG, second derivative PPG	ResNet and LOSO	SBP DBP	$9.43 \pm \text{NA}$ $6.88 \pm \text{NA}$
Baek [39] 2019	PPG/ECG (604 Subject)	Random cropping, fast Fourier transform, and signal derivative	CNN	SBP DBP	$5.32 \pm 5.54$ $3.38 \pm 3.82$
Eom [40] 2020	ECG/PPG/ BCG (15 subjects)	IInd order bandpass filter, data segmented into an 8-s window	Bi-GRU	SBP DBP	$4.06 \pm 4.04$ $3.33 \pm 3.42$
Hsu [41] 2020	PPG/ECG	Physiological parameters extracted from ECG and PPG	ANN	SBP DBP	$3.21 \pm \text{NA}$ $2.23 \pm \text{NA}$
Lie [42] 2020	PPG/ECG	Seven features extracted from PPG and ECG	LSTM	SBP DBP	$4.63 \pm \text{NA}$ $3.15 \pm \text{NA}$
Aguiree [43] 2021	PPG (1131)	Bandpass filtering, 15-s window segmentation	RNN/GRU/ Attention	SBP DBP	$12.08 \pm 15.67$ $5.56 \pm 7.32$
Lee [44] 2021	ECG/PPG (18)	Seven features extracted from PPG and ECG	BiLstm	SBP DBP	$5.82 \pm 6.82$ $5.24 \pm 6.06$
Harfiya [45] 2021	PPG	Translation of PPG to ABP signal	LSTM	SBP DBP	$4.05 \pm \text{NA}$ $2.41 \pm \text{NA}$
Sakib [46] 2022	PPG/ECG (982)	Normalization, the derivative of PPG signal	CNN + ANN	SBP DBP	$2.33 \pm 0.713$ $2.73 \pm 1.16$
Rastegar [47] 2023	PPG/ECG	Savitzky–Golay filtering	CNN + SVR	SBP DBP	$1.23 \pm 2.45$ $3.08 \pm 5.67$
Qiu [31] 2023	PPG/ECG	Pattern diffusion method	Feed-forward model	SBP DBP	$3.65 \pm \text{NA}$ $4.56 \pm \text{NA}$
Nour [48] 2023	PPG	Time domain and chaotic features	Regression models	SBP DBP	$3.069 \pm \text{NA}$ $1.721 \pm \text{NA}$
Qin [49] 2023	PPG	Multi-scale feature extraction	ResNet34	SBP DBP	$5.98 \pm \text{NA}$ $3.24 \pm \text{NA}$
Proposed work	PPG (1557 subjects)	Superlet transform-based super-resolution spectrogram	CNN	<b>SBP</b> <b>DBP</b>	<b><math>2.71 \pm 3.29</math></b> <b><math>2.42 \pm 3.94</math></b>

The results represents in bold indicates the results obtained in the proposed work

PPG signal consisting of random and irregular peaks due to strong physical exercises degrades the generalized capability of the model when tested in realtime.

Thus, the proposed work trains the deep neural network model with the super-resolution TF spectrogram to increase the accuracy of the model.

The PPG signal was segregated with consecutive windows of 2.4-s duration to generate the super-resolution TF spectrogram. However, in general, window length significantly impacts the model's performance. A longer window length can result in more accurate and precise data analysis. It provides a larger sample size for analysis

and can capture more complex patterns or trends. However, longer window length may also result in increased computational complexity, longer processing times, and potentially reduced system responsiveness.

On the other hand, using a shorter window length may result in faster processing times and a more responsive system. However, it may also result in less accurate or precise data analysis, as it provides a smaller sample size and may miss important patterns or trends.

Thus, in the future, the proposed framework's performance will be analyzed with different window lengths to achieve optimal performance.

Moreover, no realtime acquired PPG signal is used to test the proposed model. Thus, in the future, we will deploy the model on a processor, an interface with a PPG sensor for realtime monitoring and classification of BP. Further, no standard optimization technique is used in the proposed work to optimize the model's computational complexity. Thus, in the future, model parameters and hardware need to be optimized to reduce the computational complexity of the method.

## Conclusions

For diagnosis and treatment of hypertension, continuous, cuff-less, and non-invasive user-friendly BP measurement technique is a strongly desired requirement in the health care industry. Hence, the demand for realtime PPG-based continuous monitoring of BP is increasing. Considering this objective, this work proposes a superlet transform and optimized deep neural network-based SBP and DBP estimation framework. Moreover, the proposed framework also classifies BP to detect possible hypertension early.

The proposed framework introduces a superlet transform-based super-resolution spectrogram giving a 2-D TF spectrogram that separates the noise component and the true PPG BP signal component. Separation of true PPG signal components from motion artifacts affected signal helps the deep network model to learn the feature more accurately. Thus, the model relates a high correlation between an extracted feature, reference SBP, and DBP signal. The obtained SBP and DBP values also satisfy the standard provided by AAMI and BHS. The proposed framework requires a single model for the estimation and classification tasks, which helps the user better monitor the realtime BP. The proposed framework can be a possible solution for designing a wearable device to monitor BP in realtime.

**Funding** No funding available.

## Declarations

**Conflict of interest** All Authors of this work declare no conflict of interest.

**Ethical approval** This article contains no studies with human participants or animals performed by authors.

## References

- Esgalhado F, Fernandes B, Vassilenko V, Batista A, Russo S (2021) The application of deep learning algorithms for PPG signal processing and classification. *Computers* 10(12):1–15. <https://doi.org/10.3390/computers10120158>
- Sharma M et al (2017) Cuff-less and continuous blood pressure monitoring: a methodological review. *Technologies* (Basel) 5(2):21. <https://doi.org/10.3390/technologies5020021>
- Pankaj, Kumar A, Komaragiri R, Kumar M (2022) A review on computation methods used in photoplethysmography signal analysis for heart rate estimation. *Arch Comput Methods Eng* 29(2):921–940. <https://doi.org/10.1007/s11831-021-09597-4>
- Ismail SNA, Nayan NA, Jaafar R, May Z (2022) Recent advances in non-invasive blood pressure monitoring and prediction using a machine learning approach. *Sensors*. <https://doi.org/10.3390/s22166195>
- Pankaj, Kumar A, Komaragiri R, Kumar M (2023) Optimized deep neural network models for blood pressure classification using Fourier analysis-based time–frequency spectrogram of photoplethysmography signal. *Biomed Eng Lett*. <https://doi.org/10.1007/s13534-023-00296-6>
- Wang R, Jia W, Mao ZH, Sciabassi RJ, Sun M (2014) Cuff-free blood pressure estimation using pulse transit time and heart rate. In: *International conference on signal processing proceedings (ICSP)*. Institute of Electrical and Electronics Engineers Inc., pp 115–118. <https://doi.org/10.1109/ICOSP.2014.7014980>
- Ganti VG, Carek AM, Nevius BN, Heller JA, Etemadi M, Inan OT (2021) Wearable cuff-less blood pressure estimation at home via pulse transit time. *IEEE J Biomed Health Inform* 25(6):1926–1937. <https://doi.org/10.1109/JBHI.2020.3021532>
- Byfield R, Miller M, Miles J, Guidoboni G, Lin J (2022) Towards robust blood pressure estimation from pulse wave velocity measured by photoplethysmography sensors. *IEEE Sens J* 22(3):2475–2483. <https://doi.org/10.1109/JSEN.2021.3134890>
- Fotiadis DI et al (2018) Biomedical and health informatics and the body sensor networks conferences, 4–7 March 2018, Treasure Island Hotel, Las Vegas
- Liu W et al (2022) A wearable and flexible photoplethysmogram sensor patch for cuffless blood pressure estimation with high accuracy. *IEEE Sens J* 22(20):19818–19825. <https://doi.org/10.1109/JSEN.2022.3202803>
- Yang S, Sohn J, Lee S, Lee J, Kim HC (2021) Estimation and validation of arterial blood pressure using photoplethysmogram morphology features in conjunction with pulse arrival time in large open databases. *IEEE J Biomed Health Inform* 25(4):1018–1030. <https://doi.org/10.1109/JBHI.2020.3009658>
- Zhang Y, Zhang X, Cui P, Li S, Tang J (2021) Key feature selection and model analysis for blood pressure estimation from electrocardiogram, ballistocardiogram and photoplethysmogram. *IEEE Access* 9:54350–54359. <https://doi.org/10.1109/ACCESS.2021.3070636>
- Li P, Laleg-Kirati TM (2021) Central blood pressure estimation from distal PPG measurement using semiclassical signal analysis features. *IEEE Access* 9:44963–44973. <https://doi.org/10.1109/ACCESS.2021.3065576>
- Yao P et al (2022) Multi-dimensional feature combination method for continuous blood pressure measurement based on wrist PPG sensor. *IEEE J Biomed Health Inform* 26(8):3708–3719. <https://doi.org/10.1109/JBHI.2022.3167059>
- Gupta S, Singh A, Sharma A, Tripathy RK (2022) Higher order derivative-based integrated model for cuff-less blood pressure estimation and stratification using PPG signals. *IEEE Sens J* 22(22):22030–22039. <https://doi.org/10.1109/JSEN.2022.3211993>
- Dey J, Gaurav A, Tiwari VN (2018) InstaBP: cuff-less blood pressure monitoring on smartphone using single PPG sensor. In: *Annual international conference of the IEEE engineering in medicine and biology—proceedings*. <https://doi.org/10.1109/EMBC.2018.8513189>
- Chakraborty A, Goswami D, Mukhopadhyay J, Chakrabarti S (2021) Measurement of arterial blood pressure through single-site

- acquisition of photoplethysmograph signal. *IEEE Trans Instrum Meas.* <https://doi.org/10.1109/TIM.2020.3011304>
18. Cardoso GS, Lucas MG, Cardoso SS, Ruzicki JCM, Junior AAS (2022) Using PPG and machine learning to measure blood pressure. In: Bastos-Filho TF, de Oliveira Caldeira EM, Frizera-Neto A (eds) XXVII Brazilian congress on biomedical engineering. Springer, Cham, pp 1909–1915
  19. Wang D, Yang X, Liu X, Ma L, Li L, Wang W (2021) Photoplethysmography-based blood pressure estimation combining filter-wrapper collaborated feature selection with LASSO-LSTM Model. *IEEE Trans Instrum Meas.* <https://doi.org/10.1109/TIM.2021.3109986>
  20. Ruiz-Rodríguez JC et al (2013) Innovative continuous non-invasive cuffless blood pressure monitoring based on photoplethysmography technology. *Intensive Care Med.* <https://doi.org/10.1007/s00134-013-2964-2>
  21. Panwar M, Gautam A, Biswas D, Acharyya A (2020) PP-Net: a deep learning framework for PPG-based blood pressure and heart rate estimation. *IEEE Sens J* 20(17):10000–10011
  22. Esgalhado F, Fernandes B, Vassilenko V, Batista A, Russo S (2021) The application of deep learning algorithms for ppg signal processing and classification. *Computers.* <https://doi.org/10.3390/computers10120158>
  23. Chen Y, Zhang D, Karimi HR, Deng C, Yin W (2022) A new deep learning framework based on blood pressure range constraint for continuous cuffless BP estimation. *Neural Netw* 152:181–190
  24. Yen CT, Chang SN, Liao CH (2022) Estimation of Beat-by-beat blood pressure and heart rate from ECG and PPG Using a fine-tuned deep CNN model. *IEEE Access* 10:85459–85469. <https://doi.org/10.1109/ACCESS.2022.3195857>
  25. Leitner J, Chiang PH, Dey S (2022) Personalized blood pressure estimation using photoplethysmography: a transfer learning approach. *IEEE J Biomed Health Inform* 26(1):218–228. <https://doi.org/10.1109/JBHI.2021.3085526>
  26. Wang W, Mohseni P, Kilgore KL, Najafizadeh L (2022) Cuff-less blood pressure estimation from photoplethysmography via visibility graph and transfer learning. *IEEE J Biomed Health Inform* 26(5):2075–2085. <https://doi.org/10.1109/JBHI.2021.3128383>
  27. Song K, Chung KY, Chang JH (2020) Cuffless deep learning-based blood pressure estimation for smart wristwatches. *IEEE Trans Instrum Meas* 69(7):4292–4302. <https://doi.org/10.1109/TIM.2019.2947103>
  28. Yen CT, Liao JX, Huang YK (2022) Applying a deep learning network in continuous physiological parameter estimation based on photoplethysmography sensor signals. *IEEE Sens J* 22(1):385–392. <https://doi.org/10.1109/JSEN.2021.3126744>
  29. Athaya T, Choi S (2021) An estimation method of continuous non-invasive arterial blood pressure waveform using photoplethysmography: a u-net architecture-based approach. *Sensors* 21(5):1–18. <https://doi.org/10.3390/s21051867>
  30. Kim DK, Kim YT, Kim H, Kim DJ (2022) DeepCNAP: a deep learning approach for continuous non-invasive arterial blood pressure monitoring using photoplethysmography. *IEEE J Biomed Health Inform* 26(8):3697–3707. <https://doi.org/10.1109/JBHI.2022.3172514>
  31. Qiu S, Zhang YT, Lau SK, Zhao N (2022) Scenario adaptive cuffless blood pressure estimation by integrating cardiovascular coupling effects. *IEEE J Biomed Health Inform.* <https://doi.org/10.1109/JBHI.2022.3227235>
  32. Pankaj, Kumar A, Komaragiri R, Kumar M (2023) A novel CS-NET architecture based on the unification of CNN, SVM and super-resolution spectrogram to monitor and classify blood pressure using photoplethysmography. *Comput Methods Programs Biomed* 240:107716. <https://doi.org/10.1016/j.cmpb.2023.107716>
  33. Johnson AEW et al (2016) Data descriptor : MIMIC-III, a freely accessible critical care database. *Sci Data* 3:1–9
  34. Arunkumar KR, Bhasker M (2020) Heart rate estimation from wrist-type photoplethysmography signals during physical exercise. *Biomed Signal Process Control.* <https://doi.org/10.1016/j.bspc.2019.101790>
  35. Pankaj, Kumar A, Kumar M, Komaragiri R (2022) STSR: spectro-temporal super-resolution analysis of a reference signal less photoplethysmogram for heart rate estimation during physical activity. *IEEE Trans Instrum Meas* 71:1–10. <https://doi.org/10.1109/TIM.2022.3192831>
  36. Tanveer MS, Hasan MK (2019) Cuffless blood pressure estimation from electrocardiogram and photoplethysmogram using waveform based ANN-LSTM network. *Biomed Signal Process Control* 51:382–392
  37. Yan C et al (2019) Novel deep convolutional neural network for cuff-less blood pressure measurement using ECG and PPG signals. In: 2019 41st Annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, pp 1917–1920
  38. Slapničar G, Mlakar N, Luštrek M (2019) Blood pressure estimation from photoplethysmogram using a spectro-temporal deep neural network. *Sensors (Switzerland).* <https://doi.org/10.3390/s19153420>
  39. Baek S, Jang J, Yoon S (2019) End-to-end blood pressure prediction via fully convolutional networks. *IEEE Access* 7:185458–185468. <https://doi.org/10.1109/ACCESS.2019.2960844>
  40. Eom H et al (2020) End-to-end deep learning architecture for continuous blood pressure estimation using attention mechanism. *Sensors (Switzerland).* <https://doi.org/10.3390/s20082338>
  41. Hsu YC, Li YH, Chang CC, Harfiya LN (2020) Generalized deep neural network model for cuffless blood pressure estimation with photoplethysmogram signal only. *Sensors (Switzerland)* 20(19):1–18. <https://doi.org/10.3390/s20195668>
  42. Li YH, Harfiya LN, Purwandari K, der Lin Y (2020) Real-time cuffless continuous blood pressure estimation using deep learning model. *Sensors (Switzerland)* 20(19):1–19. <https://doi.org/10.3390/s20195606>
  43. Aguirre N, Grall-Maës E, Cymberknop LJ, Armentano RL (2021) Blood pressure morphology assessment from photoplethysmogram and demographic information using deep learning with attention mechanism. *Sensors* 21(6):1–19. <https://doi.org/10.3390/s21062167>
  44. Lee D et al (2021) Beat-to-beat continuous blood pressure estimation using bidirectional long short-term memory network. *Sensors (Switzerland)* 21(1):1–15. <https://doi.org/10.3390/s21010096>
  45. Harfiya LN, Chang CC, Li YH (2021) Continuous blood pressure estimation using exclusively photoplethysmography by lstm-based signal-to-signal translation. *Sensors.* <https://doi.org/10.3390/s21092952>
  46. Mahmud S et al (2022) A shallow U-Net architecture for reliably predicting blood pressure (BP) from photoplethysmogram (PPG) and electrocardiogram (ECG) signals. *Sensors* 22(3):919
  47. Rastegar S, Gholam Hosseini H, Lowe A (2023) Hybrid CNN-SVR blood pressure estimation model using ECG and PPG signals. *Sensors.* <https://doi.org/10.3390/s23031259>
  48. Nour M, Polat K, Şentürk Ü, Arıcan M (2023) A novel cuffless blood pressure prediction: uncovering new features and new hybrid ML models. *Diagnostics.* <https://doi.org/10.3390/diagnostics13071278>
  49. Qin C, Li Y, Liu C, Ma X (2023) Cuff-less blood pressure prediction based on photoplethysmography and modified ResNet. *Bioengineering* 10(4):400. <https://doi.org/10.3390/bioengineering10040400>
  50. Kachuee M, Kiani MM, Mohammadzade H, Shabany M (2017) Cuffless blood pressure estimation algorithms for continuous healthcare monitoring. *IEEE Trans Biomed Eng* 64(4):859–869. <https://doi.org/10.1109/TBME.2016.2580904>

51. El-Hajj C, Kyriacou PA (2021) Cuffless blood pressure estimation from PPG signals and its derivatives using deep learning models. *Biomed Signal Process Control*. <https://doi.org/10.1016/j.bspc.2021.102984>
52. Cheng J, Xu Y, Song R, Liu Y, Li C, Chen X (2021) Prediction of arterial blood pressure waveforms from photoplethysmogram signals via fully convolutional neural networks. *Comput Biol Med* 138:104877. <https://doi.org/10.1016/j.combiomed.2021.104877>
53. Ibtehaz N, Rahman MS (2020) PPG2ABP: translating photoplethysmogram (PPG) signals to arterial blood pressure (ABP) waveforms using fully convolutional neural networks. *ArXiv Preprint*. <https://arxiv.org/abs/2005.01669>
54. Treebupachatsakul T, Boosamalee A, Shinnakerdchoke S, Pechprasarn S, Thongpance N (2022) Cuff-less blood pressure

prediction from ECG and PPG signals using fourier transformation and amplitude randomization pre-processing for context aggregation network training. *Biosensors (Basel)*. <https://doi.org/10.3390/bios12030159>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/370984006>

# International Journal of Research Publication and Reviews Capabilities and Features Offered by SDN on the Cloud Network Infrastructure

Article · May 2023

CITATIONS

0

READS

16

3 authors, including:



[Zalmai Zormatai](#)

Marwadi Education Foundation's Group of Institutions

7 PUBLICATIONS 1 CITATION

[SEE PROFILE](#)



[Samiullah Mehraban](#)

Delhi Technological University

7 PUBLICATIONS 11 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Hybrid SDN Network [View project](#)



## **Capabilities and Features Offered by SDN on the Cloud Network Infrastructure**

***Khwaja Hedayetulla Sidiqi <sup>a</sup>, Zalmai Zormatai <sup>b</sup>, Samiullah Mehraban <sup>c</sup>***

<sup>a</sup> *Researcher, Bakhtar University*

<sup>b</sup> *Assistant Professor, Bakhtar University*

<sup>c</sup> *PhD scholar, Delhi Technological University*

---

### **ABSTRACT**

Cloud computing is an evolutionary approach of offering IT services to the industry and IT businesses. Alongside the advantages of cloud computing, it has raised many security, performance and network management concerns in the cloud datacenters. In this study, I present the implementation of SDN (software-defined networking) into the cloud network infrastructure to overcome issues related to network administration, monitoring and security. SDN is an approach to network virtualization which adds to the security and performance of the network and helps to automate and ease the network management by making the network flexible to changes and extensions.

The research methodology used is descriptive and analytical based on the recent research work done by other researchers in this domain using a systematic literature review approach. The implementation of SDN into the cloud network infrastructure is presented as a suitable proposed solution to overcome some (network security, network performance, and network management) of the current cloud network infrastructure issues.

Keywords: Cloud Computing, IT services, SDN, Network security

---

### **1. Introduction**

Cloud computing is considered as a new field of study for researchers and academicians, cloud computing is a revolutionary change the way IT services are provided in organizations. Cloud computing avoids the necessity of organizations for owning on-premises data centers and network infrastructure facilities; which reduces the cost, need of IT expert staff and brings elasticity. When plugging an electric machine into a power socket, we care neither how electric power is produced nor how it gets to that power socket. This is probable because electricity is virtualized; that is, it is already available from a wall socket that hides power generation stations and a massive power supply grid. When we think about information technologies, this idea means providing useful functions while hiding how their internals work. Computing itself, to be considered fully virtualized, must allow computers to be made from distributed components such as processing, storage, data, and software resources. Such concept describes a business model where the consumers pay to the providers based on the theory of 'Pay-as-you-go'. "Cloud computing is a techno-business disruptive model of using distributed large-scale data centers either private or public or hybrid offering customers a scalable virtualized infrastructure or an abstracted set of services qualified by service-level agreements (SLAs) and charged only by the abstracted IT resources consumed (Buyya, Broberg, & Goscinski, 2011)". ISO/IEC defines cloud computing as follows "Paradigm for enabling network access to a scalable and elastic pool of shareable physical or virtual resources with self-service provisioning and administration on-demand". The NIST (National Institute of Standards and Technology) definition of cloud according to the NIST Special Report 800-145 is: "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be quickly allocated and released with slight administrative effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models (Mell & Grance, 2011)".

---

### **2. Main Findings**

The outcome of my work will demonstrate that why the conventional network infrastructures cannot meet the performance, security and network management needs of the cloud computing and will highlight that there are many problems with the current cloud computing as well. To overcome this issue the capabilities and opportunities (high performance and better oversight of the network bandwidth, access control and applications' security) of the DSN will briefly be described, why should we use SDN to address the problems associated with the cloud network infrastructure and the impact of the SDN on the cloud computing network infrastructure is stated in details.



### 3. Cloud Computing Essential Characteristics

**a. On-demand self-service:** A client can individually deliver computing resources, such as server time and network storage, as required dynamically without demanding human interaction with each cloud service provider (CSP).

**b. Broad network access:** Competencies are accessible over the public internet and accessed through standard procedures that support the use by various customer platforms (e.g., mobile phones, tablets, laptops, and workstations).

**c. Resource pooling:** The cloud service provider's computing resources are shared to assist numerous clients using a multi-tenant architecture, with various physical and virtual resources automatically assigned and reassigned according to user request. The consumers don't have any idea about the where their data is actually stored or maybe they are only have information regarding the location of data at a higher level (for example country, province, or datacenter).

**d. Rapid elasticity:** Computing resources (e.g. CPU) can be elastically provided to the clients and freed from clients, in some cases dynamically, to measure rapidly outward and inward proportion with demand. To the customer, the capabilities offered for provisioning often seem to be unlimited and can be expected in any amount at any interval of time.

**e. Measured service:** Cloud systems dynamically govern and adjust resource use by measuring and metering competency at some level of abstraction suitable to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be observed, controlled, and reported, providing transparency for both the cloud service provider and client of the used service (Mell & Grance, 2011; Mogull et al., 2017).

### 4. Cloud Service Models

**SaaS (Software as a service)** The services or capabilities offered to the clients are that they can utilize the applications lying on a cloud hardware infrastructure. Clients usually access the data using a simple web browser, the customers cannot manage or control the underlying infrastructure of the cloud such as (Servers, storage, Operating Systems (OS), Network & etc...).

SaaS is multitenant applications running on PaaS or IaaS due to increased agility, resilience and economic benefits. Many providers offer the SaaS services using the public APIs to support the variety of clients, especially web browsers and mobile applications. Thus, the SaaS tends to have an application/logic layers and storage layer with an API on top, and there is one presentation layer which includes the web browsers, mobile applications and public APIs (Mogull et al., 2017).

**IaaS (Infrastructure as a service)** The competencies delivered to the clients are the provisioning processing, network, storage and other computing resources where the clients are able to set up their software applications. The clients will not have control over the cloud infrastructure but they can have limited control over their own network appliances (such as the firewall). The foundation of the IaaS is formed of the physical facility and hardware infrastructure. The resources (physical hardware, network & storage) are pooled using the abstract and orchestration. Abstracts free the resources from the physical constraints using virtualization to enable pooling.

Orchestration then (a set of core connectivity and delivery tools) ties these abstracted resources together to create pools and provide automation so that the resources are delivered to the customers. All these are facilitated using the APIs (Application programming interface). Most of the APIs these days use the REST (Representational State Transfer) which runs on the HTTP protocol making it very suitable for internet services. In most cases, these APIs are wrapped into a web-based interface for remote access which is called the cloud management plane (because the customers use it to manage and configure the resources). So the IaaS consists of a facility, hardware, abstract layer, and the orchestration layer. The IaaS design is simplified in Figure-1, which shows the storage and compute controllers for orchestration and the hypervisors for abstraction and the relationship between abstract and orchestration. "A series of physical servers each run two components: a hypervisor (for virtualization) and the management/orchestration software to tie in the servers and connect them to the compute controller. A customer asks for an instance (virtual server) of a particular size and the cloud controller determines which server has the capacity and allocates an instance of the requested size" (Mogull et al., 2017).

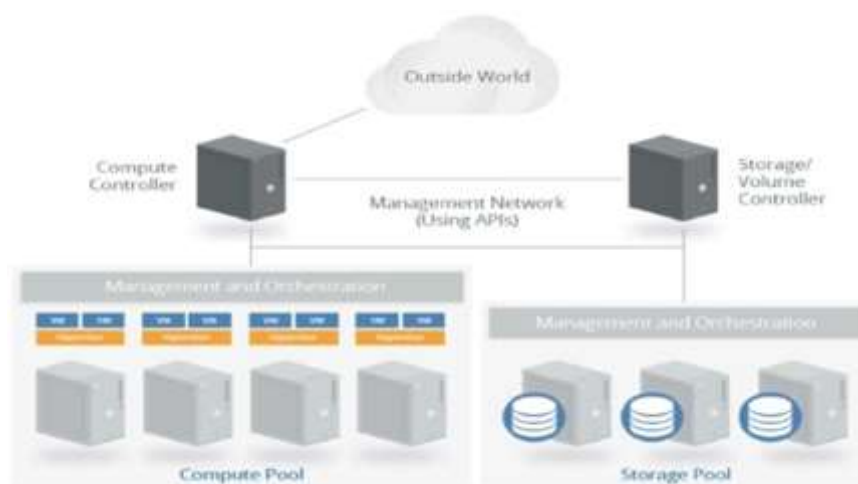


Figure 1 IaaS Platform Image source: Guidance, S. (2017). Security Guidance for Critical Areas of Focus for Cloud Computing.

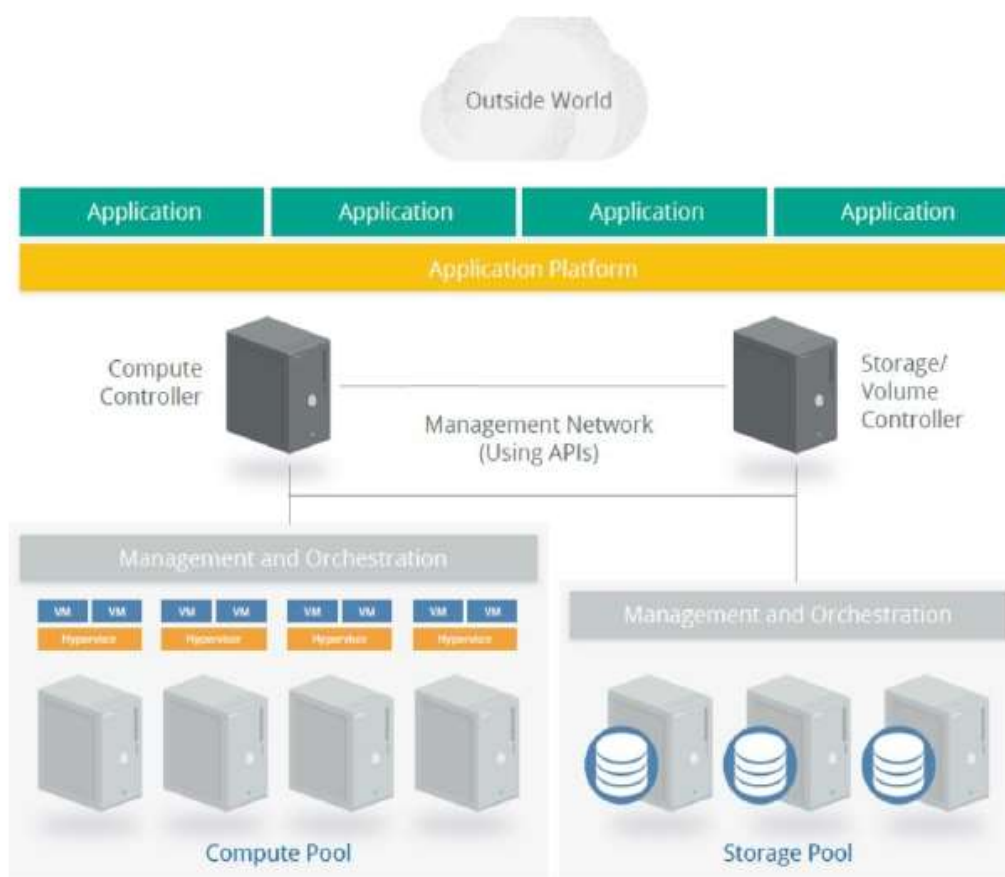


Figure 2 PaaS running on the IaaS infrastructure

Image source: Guidance, S. (2017). Security Guidance for Critical Areas of Focus for Cloud Comput

ing

#### PaaS (Platform as a service)

The competencies delivered to the consumers are the deployment of the customer's applications on the cloud infrastructure that are developed using the cloud-supported programming languages and libraries. Again, the consumers cannot control and manage the underlying cloud infrastructure. PaaS is a layer of integration and middleware built on the IaaS. It is then pooled together, orchestrated and made available to the customers using APIs. An example of PaaS is the application deployment platform where the developers can run the application codes without managing the underlying infrastructure (Mogull et al., 2017). Figure 2 shows the simplified diagram of PaaS built on the IaaS.

## 5. Cloud Deployment Models

**a. Private cloud:** The private cloud infrastructure is owned by a single organization. The organization itself or a third party will control and manage the cloud infrastructure, it may exist on or off the premises.

**b. Community cloud:** The cloud infrastructure is owned, managed and controlled by one or more organizations of a community and provides services to a specific community only. It may exist on or off the premises.

**c. Public cloud:** The public cloud provides services to everyone who is interested to use it. It is owned, managed and controlled by a business, academic or a governmental organization. The public cloud exists on the premises of the Cloud Service Provider (CSP).

**d. Hybrid cloud:** A combination of two or more of private, public and/or community cloud infrastructure that remains unique entities but are bound by a standard or proprietary technology which enables data and application portability (Mell & Grance, 2011; Mogull et al., 2017).

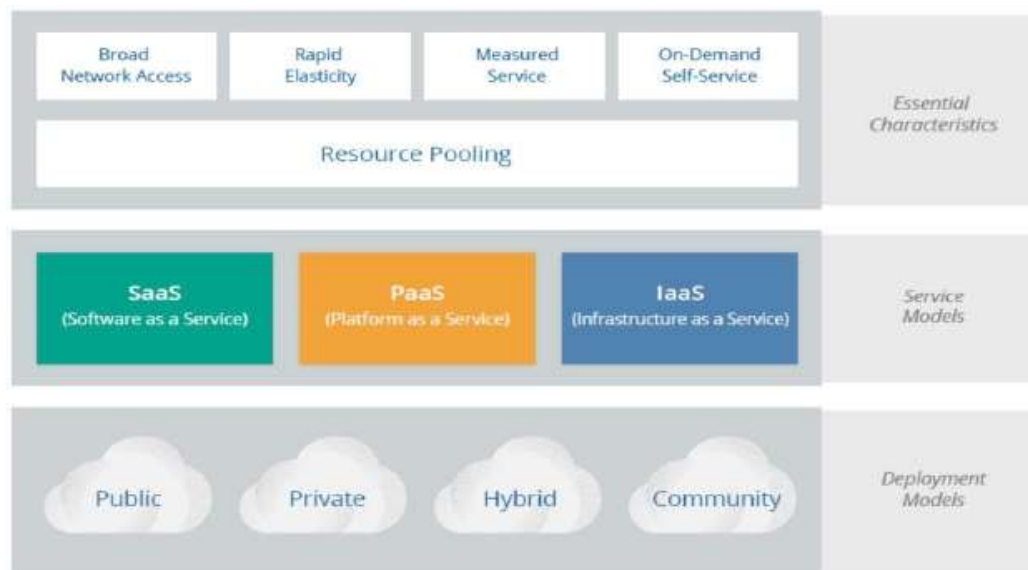


Figure 3 *Cloud Deployment Model, Service Model & Characteristics*

Image source: Guidance, S. (2017). Security Guidance for Critical Areas of Focus for Cloud Computing.

The cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort. Figure 3 shows the cloud characteristics, cloud deployment and services models, and Figure 4 illustrates the cloud deployment models with its corresponding owner.

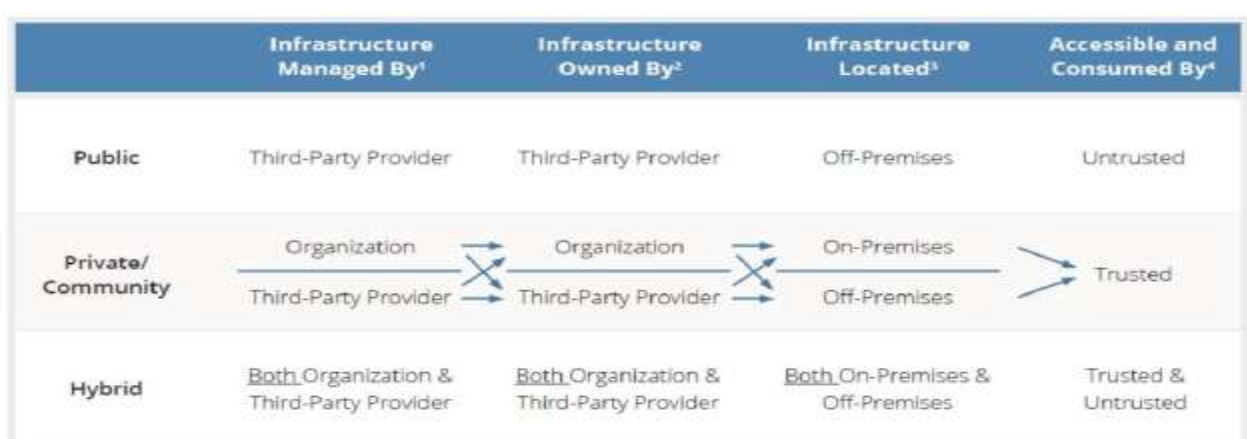


Figure 4 *Cloud Deployment Model*

Image source: Guidance, S. (2017). Security Guidance for Critical Areas of Focus for Cloud Computing.

Cloud computing is an emerging technology which can benefit the businesses by providing on-demand IT services for the payment based on the principle of pay-as-you-go. The traditional on-premises IT infrastructure requires a well-equipped facility, purchase of hardware devices, hiring professional staff, staff training, operational and maintenance costs. Such requirements cause a huge increase in overall service cost. The lack of network security, high

reliability, availability, network management, network extension, network monitoring, high network performance, traffic and network isolation, network flexibility and disaster recovery are the main problems with the on-premises network infrastructure. To overcome these problems, an organization has to use the cloud computing and rent IT infrastructure or services from the Cloud Service Providers. Cloud offers the IT services at low cost, with high reliability and availability, high performance, on-demand services, no operational or maintenance cost, and security measure at place. With all these benefits and opportunities, the cloud computing still experience the network security, performance, and network management issues.

## 6. Discussion

SDN on the other hand is a new technology which provides new opportunities and benefits for both the CSP and cloud users. SDN provides a virtualized environment and decouples the data plane from the control plane. The SDN controller is a logical centralized server which enforce policies on SDN-enabled switches, Routers, firewalls and access points. The switches usually communicate with the SDN controller using the OpenFlow protocol and update the policies into their flow tables. The benefits of SDN include but are not limited to packet filtering, network virtualization, traffic and network isolation, increased security, high performance, better network management, energy efficient, inexpensive, load balancing, fault tolerance and logical centralized control.

## 7. Conclusion

The research questions that were posted at the beginning of my thesis were: what are the weaknesses, vulnerabilities, data security, network infrastructure security and performance issues of the current cloud network infrastructure? What are the capabilities and features offered by SDN? What will be the impact of SDN implementation on current cloud network infrastructure? To answers to these questions I have conducted a systematic literature review and selected around 62 articles from IEEE, ACM and science direct journals based on the selection criteria mentioned in Chapter 6. I have reviewed and analyzed those articles and I arrived at the following findings:

- Cloud computing is vulnerable to various security threats, the main areas where the security vulnerabilities exist are the cloud network infrastructure and cloud virtualization. Cloud computing also experience the network management and network performance issues.
- SDN provides various opportunities to the cloud network infrastructure including packet filtering, network virtualization, traffic and network isolation, increased security, high performance, better network management, energy efficient, inexpensive, load balancing, fault tolerance and logical centralized control.
- The deployment of the SDN into the cloud computing can increase the cloud computing security, network performance, and network management. Based on the results of the systematic literature review, I conclude that the implementation of SDN into the cloud computing is a possible well-suited solution.

## 8. Limitation and Future work

There are still various open issues with SDN. To address these issues, further research is required. Two future research questions are summarized as follows: 1. How significant is the security in the SDN environment? As technology develops, with new trends new security vulnerabilities also emerges. The security of the SDN has to be researched in order to understand that how significant is the security with the SDN. In order to conduct further research for evaluation of the SDN security, a combination of the systematic literature review and experimental research would be better. 2. Dynamic load balancing for multiple SDN-controllers: For load balancing and fault tolerance in the SDN environment, the use of multiple SDN-Controllers is suggested. How these multiple controllers communicate with the switches and how they dynamically balance their load has to be further investigated. Experimental research approach is suggested in order to solve the problem of dynamic load balancing between multiple controllers in the SDN environment. In the near future, SDN will be adopted and deployed by various businesses and IT companies due to the benefits of SDN. Various vendors of the network devices are implementing SDN and are producing SDN-enabled network devices. Some organizations such as Open Networking Foundation (ONF) are working for the development of protocols, which communicate between the controller and the SDN-enabled switches. These protocols will be alternatives to the OpenFlow.

## References

1. Tuysuz, M. F., Ankarali, Z. K., & Gözüpek, D. (2016). A Survey on Energy Efficiency in Software Defined Networks. *Computer Networks*. <https://doi.org/10.1016/j.comnet.2016.12.012>.
2. Wang, B., Zheng, Y., Lou, W., & Hou, Y. T. (2015). DDoS attack protection in the era of cloud computing and Software-Defined Networking. *COMPUTER NETWORKS*, 81, 308–319. <https://doi.org/10.1016/j.comnet.2015.02.026>.
3. Watson, R. T., & Webster, J. (2002). Analysing The Past to Prepare for The Future: Writing Literatur Review. *MIS Quarterly Vol. 26 No. 2, Pp. Xiii-xxiii/June 2002*, 26(2). <https://doi.org/10.1.1.104.6570>.
4. Yan, Q., & Yu, F. R. (2015). Distributed Denial of Service Attacks in Software- Defined Networking with Cloud Computing, (April), 52–59.

5. Yan, Q., Yu, F. R., Member, S., Gong, Q., & Li, J. (2015). Software-Defined Networking ( SDN ) and Distributed Denial of Service ( DDoS ) Attacks in Cloud Computing Environments : A Survey , Some Research Issues , and Challenges, (c), 1–23. <https://doi.org/10.1109/COMST.2015.2487361>
6. Yang, B., Tan, F., & Dai, Y. S. (2013). Performance evaluation of cloud service considering fault recovery. In *Journal of Supercomputing* (Vol. 65, pp. 426–444). <https://doi.org/10.1007/s11227-011-0551-2>.
7. Yen, T. C., & Su, C. S. (2014). An SDN-based cloud computing architecture and its mathematical model. In *Proceedings - 2014 International Conference on Information Science, Electronics and Electrical Engineering, ISEEE 2014* (Vol. 3, pp. 1728–1731). <https://doi.org/10.1109/InfoSEEE.2014.6946218>.
8. Yigitbasi, N., Iosup, A., Epema, D., & Ostermann, S. (2009). C-Meter: A framework for performance analysis of computing clouds. In *2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, CCGRID 2009* (pp. 472–477). <https://doi.org/10.1109/CCGRID.2009.40>.
9. Yoon, C., Park, T., Lee, S., Kang, H., Shin, S., & Zhang, Z. (2015). Enabling security functions with SDN : A feasibility study. *Computer Networks*. <https://doi.org/10.1016/j.comnet.2015.05.005>.
10. Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud computing: State-of-the-art and research challenges. *Journal of Internet Services and Applications*, 1(1), 7–18. <https://doi.org/10.1007/s13174-010-0007-6>.
11. Zhang, Y., Cui, L., Wang, W., & Zhang, Y. (2018). A survey on software defined networking with multiple controllers. *Journal of Network and Computer Applications*. Elsevier Ltd. <https://doi.org/10.1016/j.jnca.2017.11.015>.
12. Zhu, G., Yin, Y., Cai, R., & Li, K. (2017). Detecting Virtualization Specific Vulnerabilities in Cloud Computing Environment. In *IEEE International Conference on Cloud Computing, CLOUD* (Vol. 2017–June, pp. 743–748). <https://doi.org/10.1109/CLOUD.2017.105>.
13. Zissis, D., & Lekkas, D. (2010). Addressing cloud computing security issues. *Future Generation Computer Systems*, 28(3), 583–592. <https://doi.org/10.1016/j.future.2010.12.006>

RESEARCH ARTICLE | SEPTEMBER 05 2023

## CFD analysis of 2x2 rod bundles at supercritical flow condition

Gaurav Kumar ; Raj Kumar Singh



AIP Conf. Proc. 2863, 020013 (2023)

<https://doi.org/10.1063/5.0155431>



Export  
Citation

CrossMark

### Articles You May Be Interested In

Stability analysis of a square rod bundle sub-channel in supercritical water reactor

*AIP Conference Proceedings* (July 2013)

Influence of corrosive environments on steels studied by positron annihilation spectroscopy

*AIP Conference Proceedings* (November 2021)

Experimental investigations on flow instabilities in a forced circulation loop at near-critical and supercritical pressures

*AIP Conference Proceedings* (July 2013)

500 kHz or 8.5 GHz?  
And all the ranges in between.

Lock-in Amplifiers for your periodic signal measurements



Find out more





# CFD Analysis of 2x2 Rod Bundles at Supercritical Flow Condition

Gaurav Kumar<sup>a)</sup> and Raj Kumar Singh

*Department of Mechanical Engineering, Delhi Technological University, Delhi-42, India*

<sup>a)</sup>Corresponding author: gauravkmr716@gmail.com

**Abstract.** Supercritical Water Reactors (SCWRs) are in consideration of many countries as it improves the thermal efficiency of the nuclear reactors as comparison to which are in regular use. The thermal-hydraulic behaviour of the water at supercritical condition are studied in past to take a step forward in the development of SCWRs. The main aim of this research is to enhance the heat transfer rate of the rod bundle with the implementation of a circular wire. CFD analysis with the help of Ansys Fluent has been done. It was seen that the wall adjacent temperature of the wire wrapped rod has been reduced. The wrapping of wire increased the transverse velocity and hence increases the heat transfer rate from rod to fluid.

## INTRODUCTION

Supercritical Water Reactor (SCWR) is one of the new concept of nuclear reactors identified by Generation-IV International Forum [1, 2]. These type of reactors have higher thermal efficiency and compact design in comparison to the general use nuclear reactors in market. Many countries are involved in doing the experiments regarding the development of SCWRs [3].

Shanghai Jiao Tong University in China is conducting continuous experiments regarding SCWRs. The main aim of the experiments is to reduce the adjacent wall temperature of the rod, so that the rod melting accident could be stopped. For that purpose various spacers are used in the rod. The wrapping of wire in the rod acts as a spacer which gives the transverse velocity to the fluid, which are helpful in increase in heat transfer from rod to fluid [3, 4, 5].

Various researchers have analyzed the heat transfer characteristics of supercritical water because the thermophysical properties of supercritical water varies quickly. It is very difficult to simulate the water at supercritical condition [6, 7].

The objective of this research is to numerically analyze the thermal hydraulic phenomena of 2x2 rod bundles with and without wires at supercritical water condition.

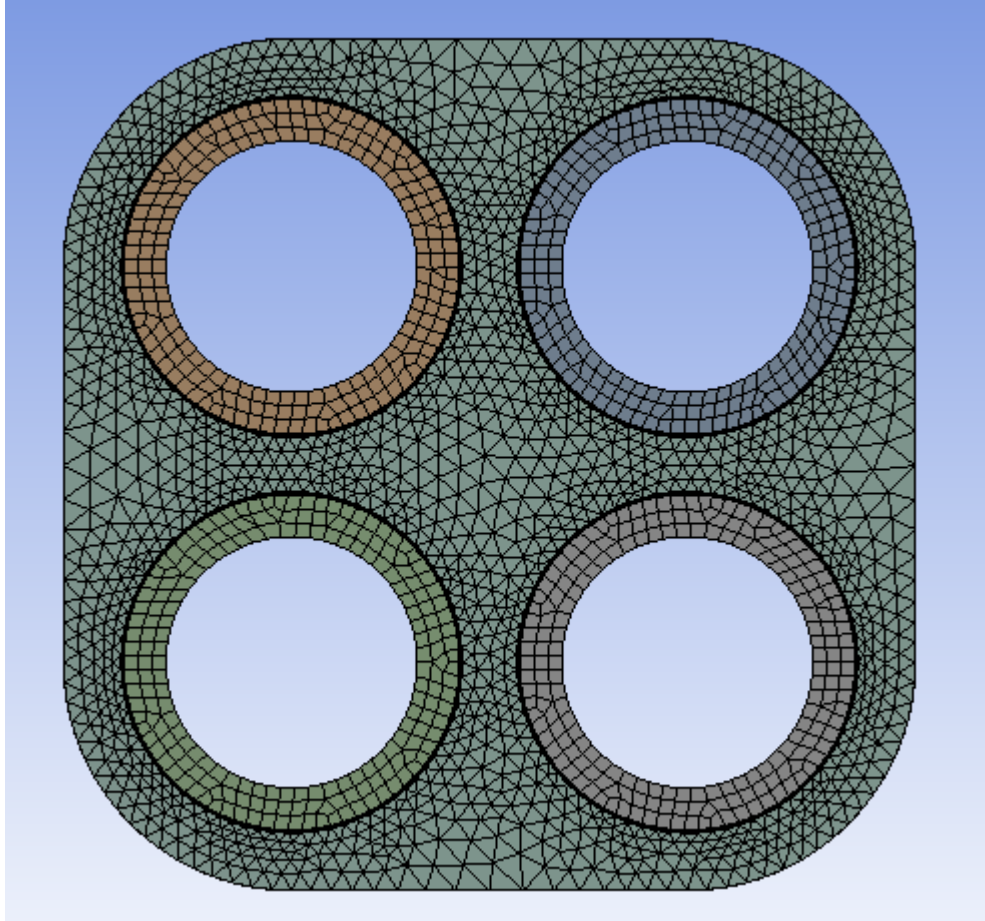
## MATHEMATICAL MODELLING

The governing equation of mass, momentum and energy was solved along with the SST K- $\omega$  model with the help of CFD software Ansys Fluent 2020 R1. The geometry was prepared with the help of dimensions given in the Table I and mesh was generated as can be seen in Fig. 1, Fig. 2 and Fig. 3. The fluid and solid part of the geometry was meshed with the help of ansys mesher and an inflation with first layer thickness of 2.96 mm was given to all the rod surfaces to capture the boundary layer phenomena.

The mass flux inlet boundary condition was used at inlet and outlet boundary condition was set to pressure outlet. A constant heat flux was applied at the inner surface of the rods. All the mathematical values of the boundary conditions

**TABLE I.** Geometrical dimension of the model.

Parameter (mm)	Value
Rod to wall corner gap	1.4
Thickness of heated rod	1.5
Heated length	600
Diameter of wire	1.2
Pitch of wire	200
Outer diameter of heated rod	8



**FIGURE 1.** Mesh of bare rod geometry

**TABLE II.** Geometrical dimension of the model.

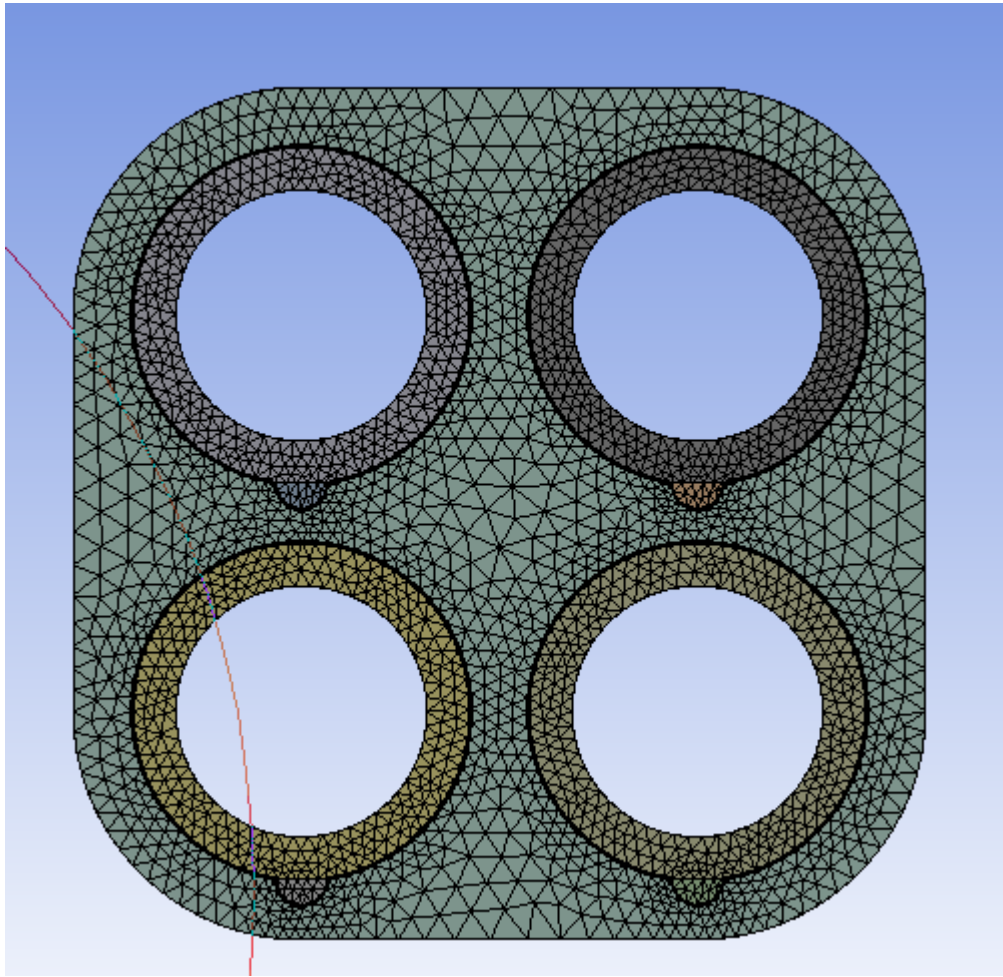
Parameter	Value
Mass Flux ( $\text{Kg/m}^2\text{s}$ )	1000
Heat Flux ( $\text{KW/m}^2$ )	600
Pressure (MPa)	25
Inlet Temperature ( $^{\circ}\text{C}$ )	397

are given in Table II

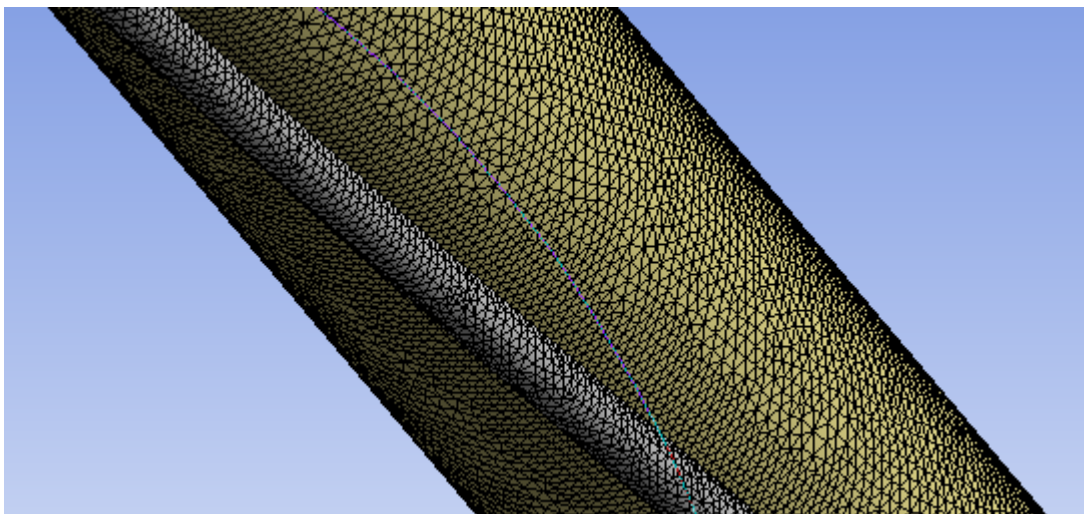
The thermophysical properties of supercritical water varies abruptly in the critical and pseudocritical temperature range. All the data of thermophysical properties of water has been fitted and a user defined function has been made which was invoked in the Ansys Fluent. The governing equations was solved with Semi Implicit Method for Pressure linked Equation(SIMPLE) type algorithm.

## RESULTS AND DISCUSSIONS

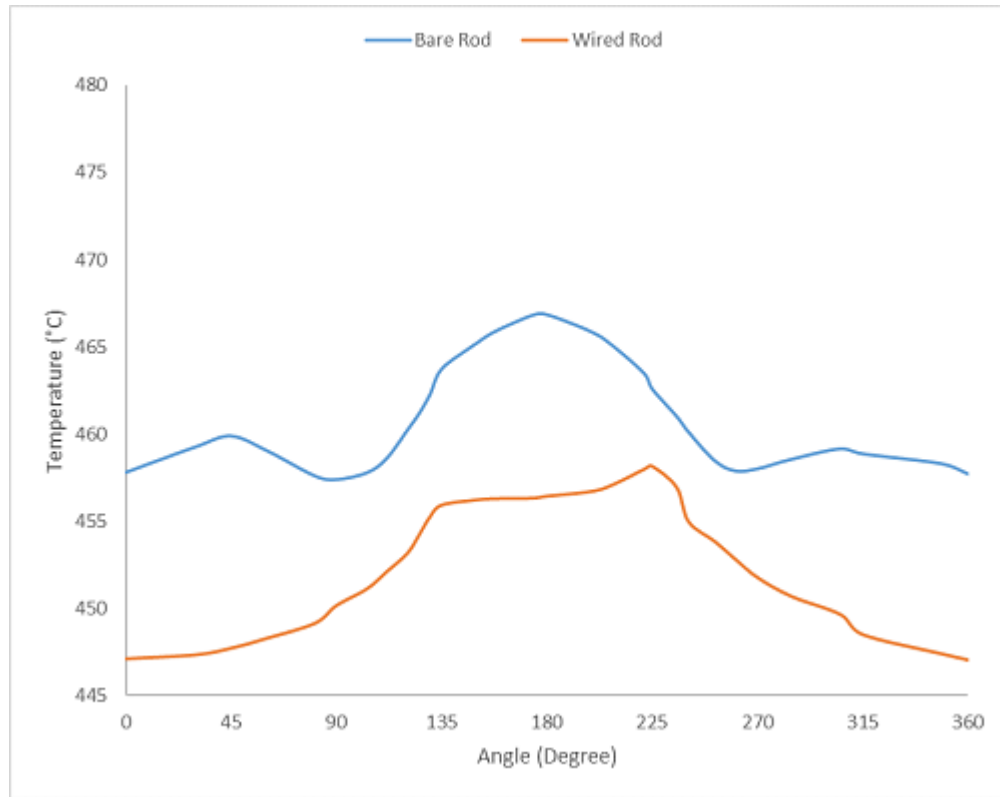
As it can be seen from Fig. 4 that the temperature of wall of the rod with wired spacer are less as compare to without wire. The installation of wire gives transverse velocity to the fluid and this process carry out more heat from rod.



**FIGURE 2.** Mesh of wired rod geometry



**FIGURE 3.** Wired rod figure



**FIGURE 4.** Graphical representation of wall temperature and Angle

## CONCLUSIONS AND RECOMMENDATIONS

The 2x2 rod bundles are numerically analysed and it can be noticed from results that the rod wall temperature at the required location was found to be less with wire in comparison to without wire. Thus the objective to reduce the peak temperature was achieved and can be seen clearly in result section. The reduction in peak temperature was confirmed because of perturbation created by wire in the fluid and a transverse velocity was seen at the location, which increases the heat transfer rate from rod to fluid.

Further the LES simulation can be done to produce more high fidelity solutions which can be validated from the experimental results.

## REFERENCES

1. T. Schulenberg and D. C. Visser, "Thermal-hydraulics and safety concepts of supercritical water cooled reactors," *Nuclear Engineering and Design* **264**, 231–237 (2013).
2. D. US, "GIF Annual Report 2014," Tech. Rep. (Generation IV International Forum, USA, 2014).
3. H. bo Li, M. Zhao, Z. xiao Hu, H. yang Gu, and D. hua Lu, "Experimental study on transient heat transfer across critical pressure in  $2 \times 2$  rod bundle with wire wraps," *International Journal of Heat and Mass Transfer* **110**, 68–79 (2017).
4. H. Wang, Q. Bi, and L. K. Leung, "Heat transfer from a  $2 \times 2$  wire-wrapped rod bundle to supercritical pressure water," *International Journal of Heat and Mass Transfer* **97**, 486–501 (2016).
5. M. Zhao and H. Y. Gu, "Experimental and numerical investigation on heat transfer of supercritical water flowing upward in  $2 \times 2$  rod bundles," *Nuclear Engineering and Design* **370** (2020), 10.1016/j.nucengdes.2020.110903.
6. Z. Shang and S. Lo, "CFD in supercritical water-cooled nuclear reactor (SCWR) with horizontal tube bundles," *Nuclear Engineering and Design* **241**, 4427–4433 (2011).
7. A. Kiss and B. Mervay, "Further details of a numerical analysis on the thermal hydraulic effect of wrapped wire spacers in fuel bundle," *Journal of Nuclear Engineering and Radiation Science* **6** (2020), 10.1115/1.4046842.

# Characterization and Steady State Analysis of Multiport Switched Boost Converter

Rishabh Bansal

CoE for Electric Vehicle and Related Technologies  
Department of Electrical Engineering  
Delhi Technological University  
New Delhi, India  
[bansalrishabh0@gmail.com](mailto:bansalrishabh0@gmail.com)

Rushiv Bansal

CoE for Electric Vehicle and Related Technologies  
Department of Electrical Engineering  
Delhi Technological University  
New Delhi, India  
[rushiv.bansal1719@gmail.com](mailto:rushiv.bansal1719@gmail.com)

Vaibhav Tokas

CoE for Electric Vehicle and Related Technologies  
Department of Electrical Engineering  
Delhi Technological University  
New Delhi, India  
[vaibhav.tokas199@gmail.com](mailto:vaibhav.tokas199@gmail.com)

Mayank Kumar (SM IEEE)

CoE for Electric Vehicle and Related Technologies  
Department of Electrical Engineering  
Delhi Technological University  
New Delhi, India  
[mayankkumar@dtu.ac.in](mailto:mayankkumar@dtu.ac.in)

**Abstract**—Multiport DC-DC converters are extensively used in electric vehicles (EVs), charging applications, UPS systems, and hybrid energy storage systems. In this paper, a multiport DC-DC converter is presented with switched boost action topology. The detailed steady state behaviour of multiport switched boost converter (MSBC) is presented. The scale factor with respect to duty ratio is defined for the characterization of MSBC. The mathematical relations are developed for gain analysis of both boost and buck output side of the converter. The steady-state characteristics of input/output current and voltage waveforms are analysed. The analytical relations are developed for voltage and current ripple and accordingly switching sequence of switches are applied. Simulation results are presented for the comparative analysis of theoretical results.

**Keywords**—Buck and boost output, converter gain, duty ratio, multiport converter, ripple analysis, scaling factor.

## I. INTRODUCTION

A dc-dc converter is used for its simple construction, efficient results and low cost (over other substitutes). With advancement in technology specially in the field of portable electronic devices [1-3], HVDC power systems, UPS systems, power supplies for consumer appliances [4-6], renewable energy applications for homes [7-8], Electric vehicles [9] and charging applications, medical devices; the requirement for a converter that can step-up and step-down dc voltage is increasing. With respect to isolations, the dc-dc converters are classified as isolated dc-dc converter and non-isolated dc-dc converters. The isolated dc-dc converters have a transformer between the input and output.

A multiport dc-dc converter (MPDC) has the advantage of more than two input/output port, which is used for more than one voltage level applications. The MPDCs may have the ability to step-up as well as step-down voltage from a single source for diversified applications from both the output ends. An MPDC can be of single input and multiple output (SIMO) type [10] or multi-input and multi output (MIMO) type [11-12]. Again, there are isolated and non-isolated type MPDC. Non-isolated type dc-dc converters are of different topologies like: buck-boost, which is commonly used in portable electronic devices; SEPIC (single-ended primary inductance converter), commonly used in LED lighting and battery charging applications; Cuk, which is used in LED lighting, telecommunication and computer systems; Isolated type dc-

dc converters are: flyback, which uses a transformer to convert voltages and used in lower power applications; full bridge used in higher power applications (motor drives, renewable energy applications); dual active bridge used in electric vehicle and renewable energy applications; isolated bidirectional in which current flow is possible in both directions and thus used in battery power applications; multi-level isolated, which uses multi-level topology and provides isolation between different input and output ports, used in high power applications.

In this paper, a non-isolated MPDC is presented, which employs switched boost topology [13]. A switched boost topology converter has an inductor coil, a capacitor and a switch. When the switch is closed, the inductor stores energy from the input sources and the output capacitors discharge to provide the load current. When the switch is open, the diode conducts and the inductor releases its energy to charge the output capacitors and maintain the output voltage. The detailed analysis of SIMO multiport switched boost converter (MSBC) is presented for single input and two outputs, one step-up and other step-down. Two IGBT switches have been employed to facilitate the switched boost action. The relationship between the duty ratio of the gate pulses of the two switches have been given by a factor defined as *scaling factor* ( $\lambda$ ). Complete characterization of the current waveforms in different arms of the converter during different modes have been presented. An instantaneous change in the output current waveform was recorded. Efforts have been made to minimize it. Thus, its variation with the voltage gain, scaling factor and duty ratio has been studied. Also, variation of the voltage gain, scaling factor, input current ripple and duty ratio with each other have been studied.

## II. MULTIPORT SWITCHED BOOST CONVERTER (MSBC)

The circuit diagram of MSBC is depicted in Fig. 1. Inductor ( $L_1$ ) is connected to the source voltage ( $V_s$ ). The parallel  $R_1$ - $C_1$  boost-load is then connected to  $L_1$  through a diode  $D$ . Two switches SW1 and SW2 are connected in series between the inductor  $L_1$  and ground. SW1 and SW2 are IGBT switches and SW2 must require anti-parallel diode. Inductor ( $L_2$ ) is connected between the two switches and with series to the parallel buck-load  $R_2$ - $C_2$ .



When both SW1 and SW2 are open, the inductor  $L_1$  discharges through  $R_1$  and  $C_1$ . When SW1 is closed and SW2 is open,  $L_1$  discharges through  $R_1-C_1$  and  $L_2-R_2-C_2$ . While  $L_2$  charges during this period. When both SW1 and SW2 is closed,  $L_1$  charges through  $V_s$ . During this period,  $L_2$  discharges through  $R_2-C_2$ . The step-up and step-down voltage outputs of the MSBC are appeared at  $V_{o1}$  and  $V_{o2}$ , respectively. The mathematical analysis and principle of operation is presented in the next section.

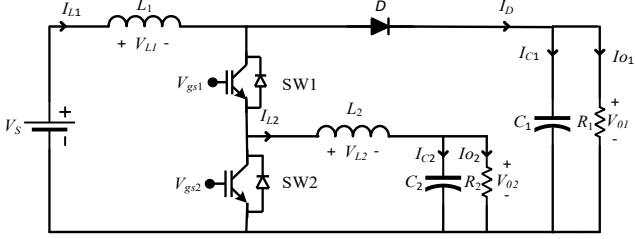


Fig. 1. Circuit diagram of multiport switched boost converter.

### III. MATHEMATICAL ANALYSIS OF MSBC

The gate pulse of SW2 (i.e.,  $V_{gs2}$ ) has a duty ratio  $D$ . The gain of the boost output voltage is dependent on this duty ratio alone (as derived in the next section). The gate pulse of SW1 (i.e.,  $V_{gs1}$ ) has a duty ratio  $\lambda D$ . Here,  $\lambda$  has been attributed as ‘scaling factor’. The duty ratio of SW2 is scaled with reference to this factor  $\lambda$ . Different voltage values and current waveform patterns are observed for different values of  $\lambda$ . These can be grouped into 3 sections for  $\lambda > 1$ ,  $\lambda = 1$  and  $\lambda < 1$  presented as follows:

#### A. Scaling Factor $\lambda > 1$

When the scaling factor ( $\lambda$ ) is greater than 1 (i.e.,  $\lambda DT > DT$ ) or the gate pulse to the switch SW1 is longer than to the gating pulse to the switch SW2. For this case, the current waveforms across both the inductors and capacitors are plotted in Fig. 2.

**A1. Switching Operation:** Based on the switching periods of the two switches SW1 and SW2, the output voltage and current waveforms, the inductor currents and the capacitor voltages is calculated. Also, the ripple in the input current and the instantaneous change in output current is also calculated. Based on the ON and OFF periods of SW1 and SW2, 3 modes can be identified as follows:

- Mode I: SW1 and SW2- ON:  $L_1$  charging,  $L_2$  discharging;
- Mode II: SW1-ON, SW2-OFF:  $L_2$  charging,  $L_1$  discharging;
- Mode III: SW1 and SW2-OFF:  $L_1$  and  $L_2$  discharging.

(i) Mode I:  $0 < t < t_1$ : During time period  $t_0 - t_1$ , switches SW1 and SW2 are closed and diode  $D$  is OFF due to reverse voltage, thus

$$V_{L1} = V_{in}; V_{L2} = -V_{o2}; V_D = -V_{o1} \quad (1)$$

$$i_{c1} = -I_{o1}; i_{c2} = i_{L2} - I_{o2}; i_1 = 0$$

(ii) Mode II:  $t_1 < t < t_2$ : During time period  $t_1 - t_2$ , switches SW1 and diode  $D$  is closed while switch SW2 is open, thus

$$V_{L1} = V_{in} - V_{o1}; V_{L2} = V_{o1} - V_{o2}; V_D = 0 \quad (2)$$

$$i_{c1} = i_{L1} - i_{L2} - I_{o1}; i_{c2} = i_{L2} - I_{o2}; i_1 = i_{L1} - i_{L2}$$

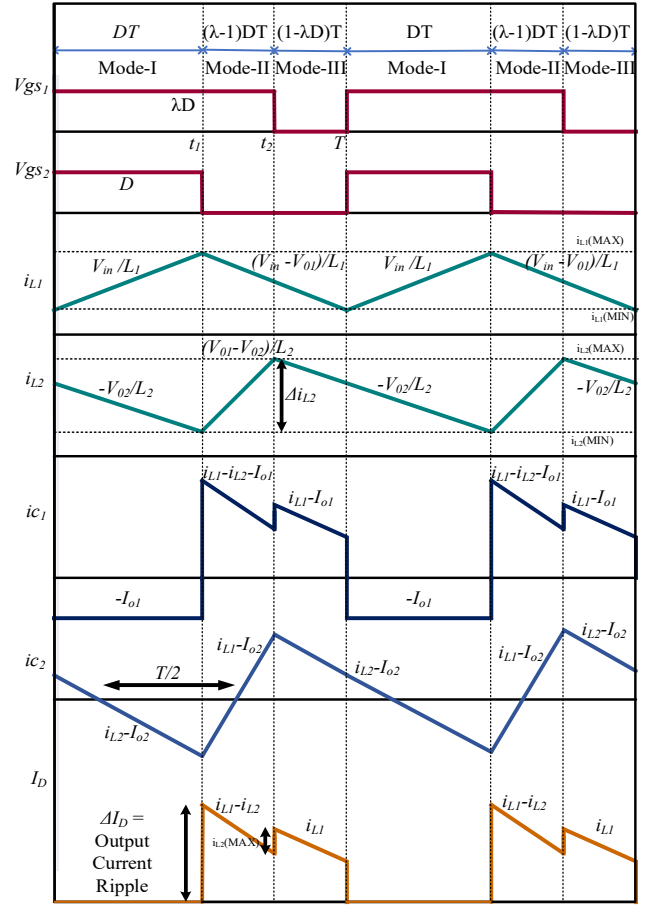


Fig. 2. Current waveforms of the MSBC for scaling factor  $\lambda > 1$ .  $i_{L1}$  and  $i_{L2}$  waveforms have slope mentioned for each mode.

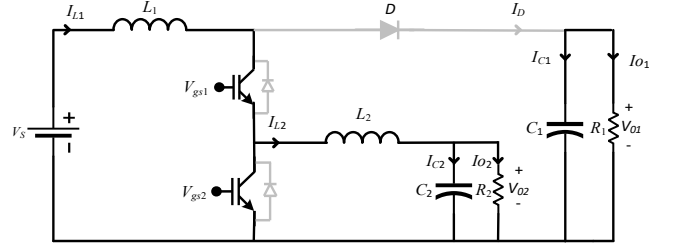


Fig. 3. Operation of MSBC during  $(0 - t_1)$ .

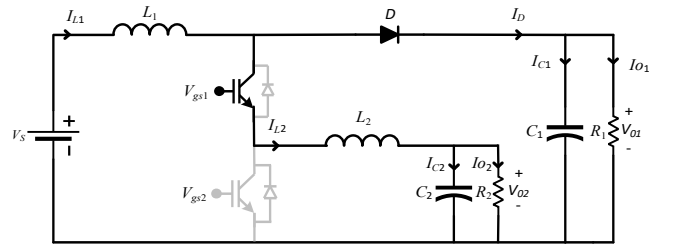


Fig. 4. Operation of MSBC during  $(t_1 - t_2)$ .

(iii) Mode III:  $t_2 < t < T$ : During time period  $t_2 - T$ , switches SW1 and SW2 are open while diode  $D$  is closed, also the anti-parallel diode of SW2 is conducting in this period, thus

$$V_{L1} = V_{in} - V_{o1}; V_{L2} = -V_{o2}; V_D = V_{D2} = 0 \quad (3)$$

$$i_{c1} = i_{L1} - I_{o1}; i_{c2} = i_{L2} - I_{o2}; i_1 = i_{L1}$$



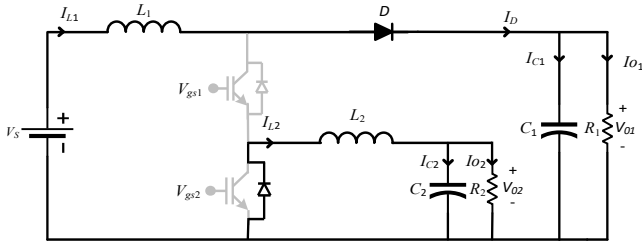


Fig. 5. Operation of MSBC during  $(t_2 - T)$ .

Now we can apply volt-sec balance to both of the inductors. Volt-sec balance says that average voltage across the inductor in a time period is zero. So,

$$\int_T V_{L1} dt = 0 \Rightarrow (V_{in} DT) + (V_{in} - V_{o1})(T - DT) = 0 \quad (4a)$$

$$V_{o1} = \frac{V_{in}}{1 - D} \quad (4b)$$

Similarly, zeroing average voltage across inductor  $L_2$ :

$$-V_{o2}DT + (V_{o1} - V_{o2})(\lambda - 1)DT - V_{o2}(1 - \lambda D)T = 0 \quad (5a)$$

$$V_{o2} = \frac{D(\lambda - 1)V_{in}}{1 - D} \quad (5b)$$

Now using ampere-sec balance to both of the capacitors, the average current across the capacitor in a switching period is zero. Therefore,

$$\int_T i_{C1} dt = 0 \quad (6a)$$

$$-I_{o1}DT + (i_{L1} - i_{L2} - I_{o1})(\lambda - 1)DT + (i_{L1} - I_{o1})(1 - \lambda D)T = 0$$

$$i_{L1}(1 - D) - i_{L2}(\lambda - 1)D = I_{o1} \quad (6b)$$

Similarly applying ampere-sec balance on the second capacitor. Therefore,

$$(i_{L2} - I_{o2})(T) = 0 ; i_{L2} = I_{o2} \quad (7a)$$

$$i_{L2_{avg}} = \frac{D(\lambda - 1)}{1 - D} \cdot \frac{V_{in}}{R_{o2}} \quad (7b)$$

Now, putting value of  $i_{L2}$  from (7b) into equation (6b),

$$i_{L1_{avg}} = \left( \frac{D(\lambda - 1)}{1 - D} \right)^2 \cdot \frac{V_{in}}{R_{o2}} + \left( \frac{1}{1 - D} \right)^2 \cdot \frac{V_{in}}{R_{o1}} \quad (8)$$

**A2. Calculation of ripple inductor currents, ripple capacitor voltages, critical inductances:** The instantaneous voltage across the inductor  $L_1$  during Mode I is given as follows:

$$V_{in} = L_1 \frac{di}{dt} \Rightarrow \Delta i_{L1} = \frac{DTV_{in}}{L_1} \quad (9)$$

Now, the maximum and minimum inductor current can be given as follows:

$$I_{L1_{max/min}} = I_{L1_{avg}} \pm \frac{\Delta i_{L1}}{2} \quad (10)$$

$$i_{L1_{max}} = V_s \left[ \frac{1}{(1 - D)^2} \left( \frac{D(\lambda - 1)^2}{R_{o1}} + \frac{1}{R_{o2}} \right) + \frac{DT}{2L_1} \right] \quad (11)$$

$$i_{L1_{min}} = V_s \left[ \frac{1}{(1 - D)^2} \left( \frac{D(\lambda - 1)^2}{R_{o1}} + \frac{1}{R_{o2}} \right) - \frac{DT}{2L_1} \right] \quad (12)$$

The instantaneous voltage across the second inductor  $L_2$  during Mode II is given as follows:

$$\left( \frac{V_{o2} - V_{o1}}{L_2} \right) = L_2 \frac{di}{dt} \quad (13a)$$

$$\Delta i_{L2} = \frac{D(\lambda - 1)T}{(1 - D)L_2} (1 - D\lambda + D)V_{in} \quad (13b)$$

Now, the maximum and minimum inductor current can be given as follows:

$$I_{L2_{max/min}} = I_{L2_{avg}} \pm \frac{\Delta i_{L2}}{2} \quad (14)$$

$$i_{L2_{max}} = V_s \left[ \left[ \frac{D(\lambda - 1)}{1 - D} \right] \left( \frac{1}{R_{o2}} + \frac{(1 - \lambda D + D)T}{2L_2} \right) \right] \quad (15)$$

$$i_{L2_{min}} = V_s \left[ \left[ \frac{D(\lambda - 1)}{1 - D} \right] \left( \frac{1}{R_{o2}} - \frac{(1 - \lambda D + D)T}{2L_2} \right) \right] \quad (16)$$

(15) also represents the instantaneous change in output current, as can be verified from Fig. 2. From the plot of capacitor currents, ripple in capacitor voltage calculated as follows:

$$C \frac{dV}{dt} = i_c ; \Delta V_c = \int \frac{i_c dt}{C} \quad (17)$$

$$\Rightarrow \Delta V_{c1} = \frac{I_{o1}DT}{C_1} = \frac{V_{o1}DT}{R_{o1}C_1} = \left( \frac{D}{1 - D} \right) \cdot \frac{T V_{in}}{R_{o1}C_1} \quad (18)$$

$$\Rightarrow \Delta V_{c2} = \frac{D[1 - D\lambda + D](\lambda - 1)}{8(1 - D)L_2} \left( \frac{T^2}{C_2} \right) V_{in} \quad (19)$$

Critical Inductance is the minimum value of inductance below which the inductor current becomes discontinuous. These are minimum value of inductors to be chosen for satisfied operation. For inductor current to be just continuous, the average value is put equal to half the ripple value:

$$i_{L_{avg}} = \frac{\Delta i_L}{2} \quad (20)$$

$$L_{1_{critical}} = \frac{D(1 - D)^2 TR_{o1}}{2 \left[ 1 + \frac{R_{o1}}{R_{o2}} \{ D(\lambda - 1) \}^2 \right]} \quad (21)$$

$$L_{2_{critical}} = \frac{DT R_{o2}}{2} \quad (22)$$

The following design parameters are used for the verification of derived results:  $D = 0.5$ ,  $\lambda = 1.5$ ,  $V_{in} = 24$  V,  $L_1 = L_2 = 5$  mH;  $C_1 = C_2 = 50$   $\mu$ F;  $R_{o1} = R_{o2} = 100$   $\Omega$ . The system outputs are calculated and mentioned as below:  $V_{o1} = 48$  V,  $V_{o2} = 12$  V,  $i_{L1_{avg}} = 1.02$  A;  $\Delta i_{L1} = 0.24$  A;  $i_{L1_{max}} = 1.14$  A;  $i_{L1_{min}} = 0.9$  A;  $i_{L2_{avg}} = 0.12$  A;  $\Delta i_{L2} = 0.18$  A;  $i_{L2_{max}} = 0.21$  A;  $i_{L2_{min}} = 0.03$  A;

$$\Delta V_{c1} = 0.48V ; \Delta V_{c2} = 0.045V ; L_{1critical} = 0.588mH$$

$$L_{2critical} = 3.75mH .$$

The above-mentioned voltage gain values at the specified duty ratio and scaling factor can be verified from the Fig. 8(e) and Fig. 8(h). The input current ripple at the specified duty ratio and scaling factor can be verified from the Fig. 8(f) and Fig. 8(i). The instantaneous change in output current at the specified duty ratio and scaling factor can be verified from the Fig. 8(g) and Fig. 8(j).

### B. Scaling Factor $\lambda < 1$

**B1. Switching Operation:** Based on the switching periods of the two switches SW1 and SW2, the output voltage and current waveforms, the inductor currents and the capacitor voltages is calculated. Based on the ON and OFF periods of SW1 and SW2, 3 modes can be identified as follows:

Mode I: SW1 and SW2- ON:  $L_1$  charging,  $L_2$  discharging;

Mode II: SW1-OFF, SW2-ON:  $L_1$  and  $L_2$  discharging;

Mode III: SW1 and SW2-OFF:  $L_1$  and  $L_2$  discharging.

Here  $L_2$  is discharging in all the modes implies  $V_{o2} = 0$  and  $I_{o2} = 0$ .

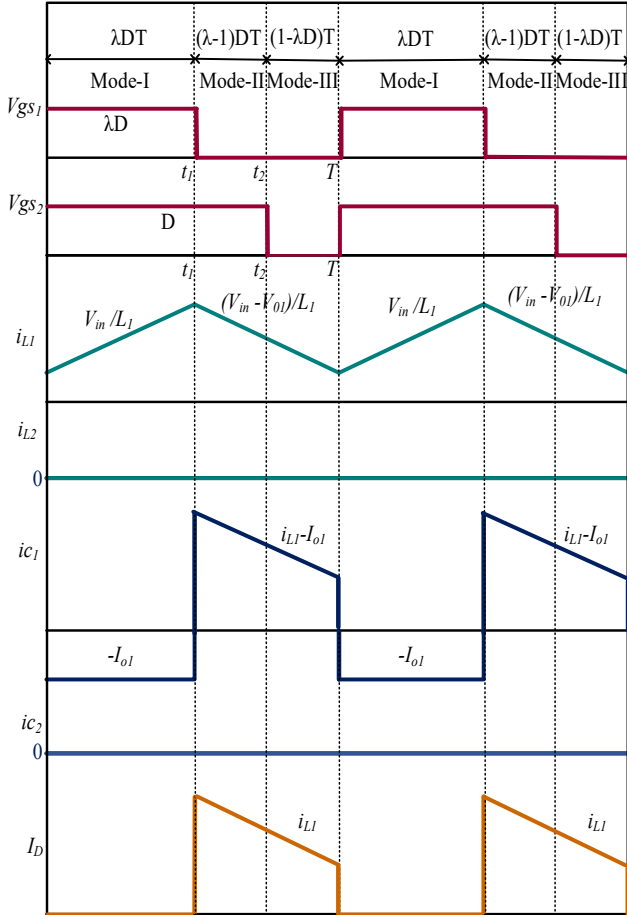


Fig. 6. Current waveforms of the MSBC for scaling factor  $\lambda < 1$ .

(i) Mode I ( $0 < t < t_1$ ) and Mode III ( $t_2 < t < T$ ): Both the modes (circuit diagram and mathematical relations) are similar to the cases discussed above for  $\lambda > 1$ .

(ii) Mode II:  $t_1 < t < t_2$ : During time period  $t_1$ - $t_2$ , switches SW2 and diode D is closed while switch SW1 is open, thus

$$V_{L1} = V_{in} - V_{o1} ; V_{L2} = -V_{o2} ; V_D = 0 ;$$

$$i_{c2} = i_{L2} - I_{o2} ; i_1 = i_{L1} ; i_{c1} = i_{L1} - i_{L2} - I_{o1}$$

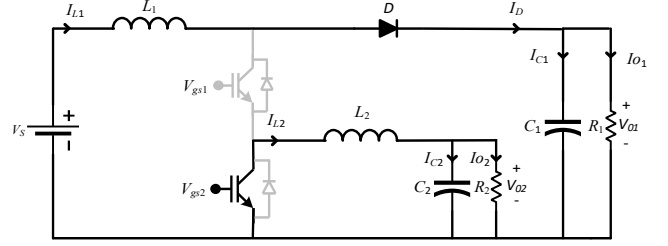


Fig. 7. Operation of MSBC during  $(t_1 - t_2)$ .

**B2. Analytical formulations:** For analysis of different modes of operation, It is assumed that  $i_{L2}$  has finite value. But we find that  $L_2$  is discharging in all modes. Hence  $i_{L2} = 0$ .

Similarly, by applying volt-sec and ampere-sec equations:

$$V_{o1} = \frac{V_{in}}{1 - \lambda D} \quad (24)$$

$$V_{o2} = 0 ; I_{o2} = 0 \quad (25)$$

$$i_{L1avg} = \frac{V_{in}}{R_{o1}(1 - \lambda D)^2} \quad (26)$$

$$\Delta i_{L1} = \frac{\lambda D T V_{in}}{L_1} \quad (27)$$

$$i_{L1(min)} = V_{in} \left[ \frac{1}{R_{o1}(1 - \lambda D)^2} - \frac{\lambda D T}{2L_1} \right] \quad (28)$$

$$i_{L1(max)} = V_{in} \left[ \frac{1}{R_{o1}(1 - \lambda D)^2} + \frac{\lambda D T}{2L_1} \right] \quad (29)$$

$$\Delta V_C = \frac{\lambda D T V_{in}}{(1 - \lambda D) R_{o1} C} \quad (30)$$

### C. Scaling factor $\lambda = 1$

When the scaling factor  $\lambda = 1$ , the buck converter gain is zero. Now other results including the voltage at the boost side can be calculated from either of above results by putting  $\lambda = 1$ . The derived results show that the converter works as MSBC for  $\lambda > 1$ , whereas the same converter is used as boost converter for  $\lambda \leq 1$ .

## IV. RELATIONSHIP OF VOLTAGE GAIN AND CURRENT RIPPLE WITH SCALING FACTOR AND DUTY RATIO

### A. Gain v/s Instantaneous change in output current:

From Fig. 2, it can be observed that the instantaneous change in output current ( $I_D$ ) i.e.  $i_{L2(max)}$  is given by (15). Its relationship with the gain of boost and buck outputs are plotted in Figs. 8 (a). and 8 (b). The graph shown in Fig. 8 (a) has been plotted for  $\lambda D = 0.75$ , where  $\lambda > 1$ . For the value of  $\lambda \leq 1$ , the instantaneous change in output current will be equal to zero. It can be seen that the boost-gain is inversely proportional to the instantaneous change for a certain period while the buck-gain is directly proportional to it in the same

period. After a certain time, after a maximum instantaneous change value is reached, there is no further increase in it with fall in boost-gain and rise in buck-gain. In Fig. 8 (b), same graph has been plotted for multiple values of  $\lambda D$ . It can be seen that for the same value of instantaneous change in output current, greater value of the product  $\lambda D$  yields more gain for boost and buck side of the output decreases slightly.

### B. Gain v/s Input Current Ripple:

From Fig. 2, it can be observed that the input current ( $i_{L1}$ ) ripple i.e.,  $\Delta i_{L1}$  is given by (9). Its relationship with the gain of

boost and buck outputs are plotted in Figs. 8(c) and 8(d). The graph shown in Fig. 8(c) has been plotted for  $\lambda D = 0.75$  where  $\lambda > 1$ . With increase in boost-gain, the ripple in the input current also increases. The buck-gain decreases slowly as the ripple increases and finally becomes 0. From Fig. 8(d), it can be observed that the variation of boost-gain with ripple is same for all values of  $\lambda D$ . It can also be seen that for the same value of ripple, the gain of buck converter increases with  $\lambda D$ . Thus, depending on the voltage required at the buck side, value of  $\lambda D$  can be so chosen such that the ripple is minimum and in accordance to the duty ratio required for the boosted output.

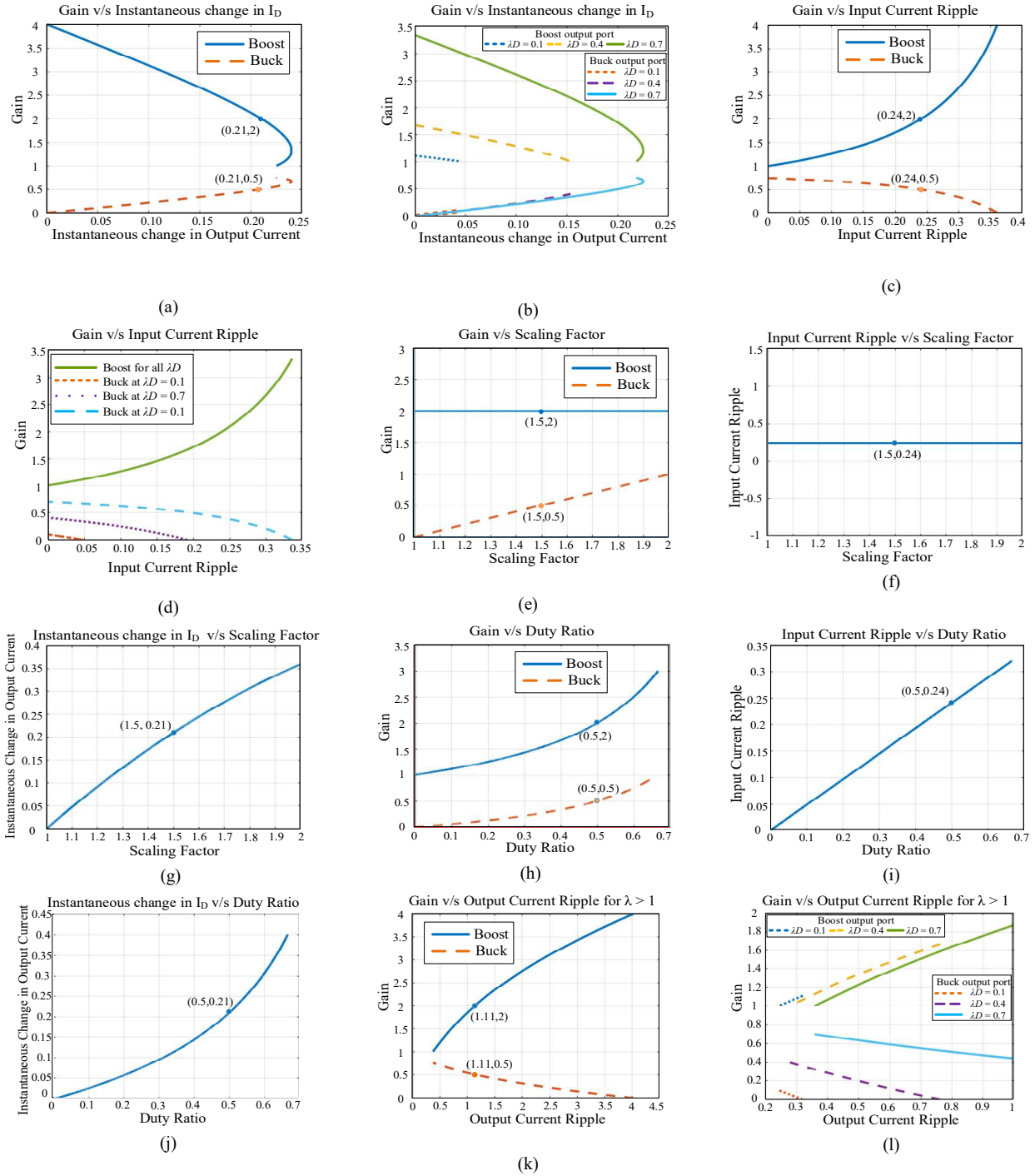


Fig. 8. Relationship of voltage gain with instantaneous change in output current at: (a)  $\lambda D = 0.75$ , (b) different values of  $\lambda D$ ; relationship of voltage gain with input current ripple at (c)  $\lambda D = 0.75$ , and at (d) different values of  $\lambda D$ ; (e), (f), (g) behavior of gain, input current ripple and instantaneous change in  $I_D$  at different values of  $\lambda$  with constant duty  $D = 0.5$ , respectively; (h), (i), (j) relationship of voltage gain, input current ripple, instantaneous change in output current at different duty ratio for constant scaling factor  $\lambda = 1.5$ , respectively; (k), (l) relationship of voltage gain with output current ripple for  $\lambda > 1$  and at (k)  $\lambda D = 0.75$ , (l) different values of  $\lambda D$ .

### C. Gain, Instantaneous change in Output Current, Input Current Ripple v/s Scaling Factor for a fixed Duty Ratio:

As  $D$  is fixed, boost output is same for varying  $\lambda$  because it depends only on  $D$  (refer (4b)) and buck output linearly increases with increase in  $\lambda$  (refer (5b)). Boost and buck voltage gains are shown in Fig. 8 (e) where  $\lambda$  is varying from 1 to 2 and  $D$  is fixed at 0.5. Input current ripple is independent of  $\lambda$  (refer (9)) hence current ripple value is constant at 0.25 in Fig. 8 (f). Relation of instantaneous change in output current can be seen from (15). Thus, at fixed  $D$ , the instantaneous change in output current increases as the scaling factor is increased as shown in Fig. 8 (g).

### D. Gain, Instantaneous change in Output Current, Input Current Ripple v/s Duty Ratio for a fixed Scaling Factor:

As  $\lambda$  is fixed, buck and boost output are increasing for increasing  $D$  (refer (4b) and (5b)). Buck and boost voltage gains are shown in Fig. 8 (h) where  $D$  is varying from 0.5 to 0.9 and scaling factor  $\lambda$  is fixed at 1.5. Input current ripple is directly proportional to  $D$  (refer (9)) hence current ripple value is linearly varying with  $D$  in Fig. 8 (i). Relation of instantaneous change in output current can be seen from (15). Thus, at fixed  $\lambda$ , instantaneous change in output current increases as the duty ratio is increased as shown in Fig. 8 (j).

### E. Gain v/s Output Current Ripple:

From Fig. 2, it can be observed that the output current ripple ( $\Delta I_D = i_{L1(max)} - i_{L2(min)}$ ). Its relationship with the gain of boost and buck outputs are plotted in Figs. 8 (k). and 8 (l). The graph shown in Fig. 8 (k) has been plotted for  $\lambda D = 0.75$ , where  $\lambda > 1$ . It can be seen that the boost-gain is directly proportional to the ripple while the buck-gain is inversely proportional to it. In Fig. 8 (l), same graph has been plotted for multiple values of  $\lambda D$ . It can be seen that for the same value of output current ripple, greater value of the product  $\lambda D$  yields less gain for boost and more gain at the buck side of the output.

## V. CONCLUSION

The characterization of a SIMO type MSBC with switched boost topology has been proposed in this paper. The proposed converter is running on two IGBT switches whose gate signals have been related through a factor attributed as the scaling factor ( $\lambda$ ). The behavior of the converter with the value of  $\lambda$  with respect to unity has been proposed. The MSBC gain analysis along with current waveforms and theoretical formulations have been presented. The variation of gain with the instantaneous change in output current yielded: for the same value of instantaneous change in output current, greater value of the product  $\lambda D$  yields more gain for boost and buck side of the output decreases slightly.

The variation of gain with input current ripple concluded that depending on the voltage required at the buck side, value of  $\lambda D$  can be so chosen such that the ripple is minimum and in accordance to the duty ratio required for the boosted output. It was also concluded that at a fixed duty ratio (i.e., fixed boost-gain), the buck-gain varies linearly with  $(\lambda-1)$ . Input current ripple is found to be independent of the scaling factor. instantaneous change in output current depends on the scaling factor and increases as  $\lambda$  increases.

Both buck and boost voltage gain of MSBC varies with duty ratio when the scaling factor was kept constant. Input current ripple varied linearly with duty ratio irrespective of the value of scaling factor. The instantaneous change in output current increases with duty ratio at a fixed scaling factor. It is concluded from the relationship of gain v/s output current ripple that, for the same value of output current ripple, greater value of the product  $\lambda D$  yields less gain for boost and more gain at the buck side of the output.

## ACKNOWLEDGMENT

This research supported by the Science and Engineering Research Board (SERB), Department of Science & Technology, Government of India., under the SERB sanction order number SRG/2021/001640.

## REFERENCES

- [1] R. -J. Wai and J. -J. Liaw, "High-Efficiency-Isolated Single-Input Multiple-Output Bidirectional Converter," in IEEE Transactions on Power Electronics, vol. 30, no. 9, pp. 4914-4930, Sept. 2015, doi: 10.1109/TPEL.2014.2364817.
- [2] P. Patra, A. Patra, and N. Mishra, "Single-Inductor Multiple-Output Switcher With Simultaneous Buck, Boost, and Inverted Outputs," in IEEE Transactions on Power Electronics, Vol. 27, No. 4, April 2012, pp. 1936- 1951.
- [3] D. Ma, W.-H. Ki, C. Y. Tsui, and P. K. T. Mok, "Single-inductor multiple output switching converters with time-multiplexing control in discontinuous conduction mode," IEEE J. Solid-State Circuits, vol. 38, pp. 89-100, Jan. 2003.
- [4] Hyun-Chang Kim, Chang Soo Yoon, Deog-Kyoon Jeong, and Jaeha Kim, "A Single-Inductor, Multiple-Channel Current-Balancing LED Driver for Display Backlight Applications," in IEEE Transactions On Industry Applications, Vol. 50, No. 6, Nov. 2014, pp. 4077-4081..
- [5] Kumar Modepalli and Leila Parsa, "A Scalable N-Color LED Driver Using Single Inductor Multiple Current Output Topology," in IEEE Transactions On Power Electronics, Vol. 31, No. 5, May 2016, pp. 3773-3783. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
- [6] Rajiv Damodaran Prabha and Gabriel A. Rincón-Mora, "Battery-assisted and Photovoltaic-sourced Switched-inductor CMOS Harvesting Charger- Supply," in IEEE International Symposium on Circuits and Systems (ISCAS2013), pp. 253-256.
- [7] H. Shao, X. Li, C. -Y. Tsui and W. -H. Ki, "A Novel Single-Inductor Dual-Input Dual-Output DC-DC Converter With PWM Control for Solar Energy Harvesting System," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 22, no. 8, pp. 1693-1704, Aug. 2014, doi: 10.1109/TVLSI.2013.2278785.
- [8] Y. Qian, H. Zhang, Y. Chen, Y. Qin, D. Lu, and Z. Hong, "A SIDIDO DC- DC converter with dual-mode and programmable-capacitor-array MPPT control for thermoelectric energy harvesting," in IEEE Trans. Circuits and Systems II: Express Briefs
- [9] Nahavandi A, Hagh M T, Sharifian M B B, et al. "A Non-isolated multiinput multioutput DC-DC boost converter for electric vehicle applications," IEEE Trans. Power Electron., vol. 30, no. 4, pp. 1818-1835, 2015.
- [10] S. K. Mishra, K. K. Nayak, M. S. Rana and V. Dharmarajan, "Switched-Boost Action Based Multiport Converter," in IEEE Transactions on Industry Applications, vol. 55, no. 1, pp. 964-975, Jan.-Feb. 2019, doi: 10.1109/TIA.2018.2869098.
- [11] H. Behjati and A. Davoudi, "A multiple-input multiple-output DC-DC converter," IEEE Transactions on Industry Applications, vol. 49, no. 3, pp. 1464-1479, May/Jun. 2013.
- [12] S. D. Saman, P. Zhang, X. N. Lu, and M. Hamzeh, "Mutual interactions and stability analysis of bipolar DC microgrids," CSEE Journal of Power and Energy Systems, vol. 5, no. 4, pp. 444-453, Dec. 2019.
- [13] Olive Ray and Santanu Mishra, "Switched-boost action: a phenomenon for achieving time-division-multiplexed multi-port power transfer for nanogrid applications," in Sadhana, Vol. 42, No. 8, pp. 1227-1238, August 2017.

# Circular Polarization Division Multiplexing in Visible Light Communication System by Incorporating QPSK and Distortion Compensation Enabled DSP/IDSP

**Hunny Pahuja**

KIET Group of Institutions, Delhi-NCR

**Monika Verma**

Software Engineer, Delhi Technological University

**Shippu Sachdeva** (✉ [sachdeva.shippu@gmail.com](mailto:sachdeva.shippu@gmail.com))

Lovely Professional University

**Simarpreet Kaur**

Chandigarh University

**Manoj Sindhwani**

Lovely Professional University

**Manoj Kumar Shukla**

Symbiosis Institute of Technology

---

## Research Article

**Keywords:** VLC, CPDM, QPSK, GMOS, DSP

**Posted Date:** September 18th, 2023

**DOI:** <https://doi.org/10.21203/rs.3.rs-3349565/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Circular Polarization Division Multiplexing in Visible Light Communication System by Incorporating QPSK and Distortion Compensation Enabled DSP/IDSP

Hunny Pahuja<sup>1</sup>, Monika Verma<sup>2</sup>, Shippu Sachdeva<sup>3,\*</sup>, Simarpreet Kaur<sup>4</sup>, Manoj Sindhwani<sup>5</sup>, Manoj Kumar Shukla<sup>6</sup>

<sup>1</sup>Department of Electronics and Communication Engineering, KIET Group of Institutions, Delhi-NCR, Ghaziabad, India

<sup>2</sup>Software Engineer, EE, Delhi Technological University, Delhi, India

<sup>3,5</sup>School of Electronics and Electrical Engineering, Lovely Professional University, Jalandhar, India

<sup>4</sup>Department of Electronics and Communication, Chandigarh University, Gharuan, India

<sup>6</sup>Department of Robotics and Automation, Symbiosis Institute of Technology, Pune, India

\*Correspondence: Sachdeva.shippu@gmail.com

## Abstract

Hybrid polarization division multiplexing (PDM) and Visible Light Communication (VLC) have fostered speedy data transmission in the last few years and emerged as the strong candidate that enables users to leverage the pervasive illumination/communication infrastructure. Circular PDM (CPDM) is replacing the linear PDM (LPDM) variant in wired/wireless systems due to the even scattered light distributions and elimination of polarization axis alignment requirements. In this research work, a 1.6 Tbps multi-wavelength line of sight (LoS) based VLC system is presented and data modulation is performed by employing Quadrature phase shift keying (QPSK). The conventional DSP algorithms such as Viterbi Phase Estimation (VPE), Blind Phase Search (BPS), and Constant Modulus Algorithm (CMA) algorithms are replaced with Gram-Schmidt orthogonalization procedure (GSOP), time-domain equalization algorithm (TEDA), improved Viterbi algorithm (IVA), and least mean square (LMS) algorithm in proposed IDSP. Three different systems are compared such as LPDM-VLC-DSP, CPDM-VLC-DSP, and CPDM-VLC-IDSP at different VLC link ranges, transmitter half angles (THA), incidence half angles (IHA), and optical concentrator areas (OCAs) in terms of error vector magnitude percentage (EVM%), log symbol error rate (log SER), and Q factor. After doing the extensive comprehensive literature survey, it is discerned that the presented CPDM-VLC-IDSP system has covered the longest distance i.e. 14 cm at 1.6 Tbps capacity under the acceptable bit error rate (BER) limits.

**Keywords-** VLC, CPDM, QPSK, GMOS, DSP

## 1. Introduction

Visible Light Communication (VLC) and fifth-generation (5G) technology serve different but complementary roles in the wireless communication landscape. A hybrid combination of VLC and 5G can enhance connectivity and enable a wide range of applications in various domains (Lu et al. 2021). The 5G is the fifth generation of wireless technology for cellular networks. It brings significant improvements over previous generations (Fourth Generation (4G), Third Generation (3G), etc.) and is designed to enable faster downloads and upload speeds (<sup>a</sup>Kaur et al. 2022; <sup>a</sup>Sachdeva et al. 2022), making it ideal for high-definition video streaming, online gaming, and other data-intensive mobile applications (Yang et al. 2019) with low latency and high reliability (Kishore et al. 2022), enable autonomous vehicles to communicate with each other and with infrastructure, improving safety and traffic management, smart traffic management (Yoo et al. 2016), energy-efficient street lighting, environmental monitoring, and enhance the user experience for Virtual / Augmented Reality (VR/AR) applications, making them more immersive and responsive (Angurala et al. 2022).

An integration of 5 G-enabled systems and VLC technology can open up a new window for high-speed networks. Using 380 nm to 780 nm wavelength window, a wireless transmission technology is feasible with



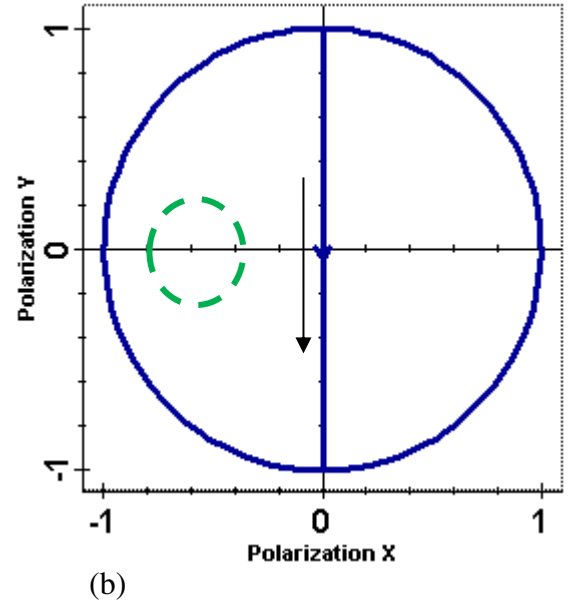
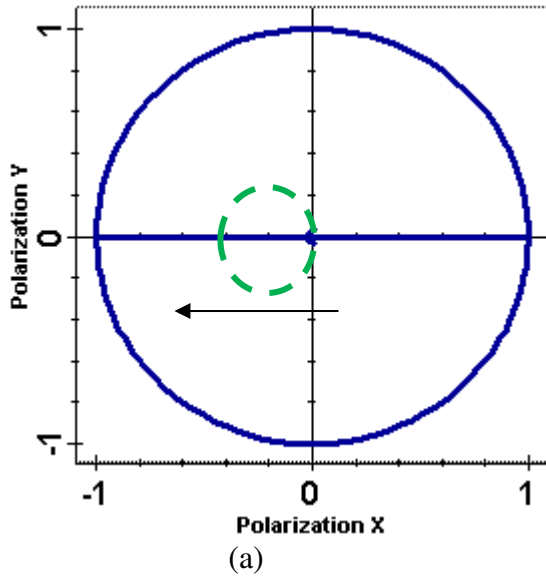
the deployment of VLC (<sup>b</sup>Kaur 2022; Kaur 2021; Kaur et al. 2019). It can be used for navigation systems and indoor positioning. Wireless data transmission using VLC is preferred in hospitals, aircraft cabins, or areas sensitive to Radio Frequency (RF) interference. Moreover, Light-Fidelity (LiFi) technology is one of the futuristic technologies offered by VLC and can provide secure and high-speed wireless internet access in specific areas, like offices and homes employing Light Emitting Diodes (LEDs) ((Miras et al. 2018). Underwater communication is also supported by VLC for underwater robotics, exploration, and monitoring (Shawky et al. 2023). Nowadays, the use of LED is replaced by Lasers in extended-reach VLC systems to support higher data traffic and to provide better performance (Retamal et al. 2015).

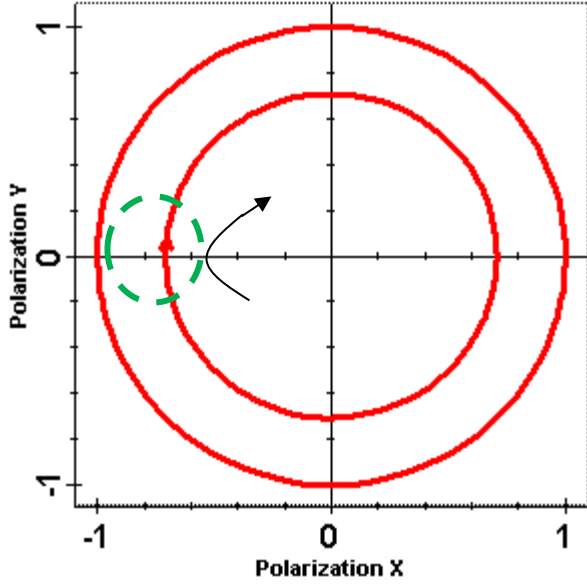
Numerical simulations and experiments were performed using the VLC channel and data encoding was performed through on-off keying particularly Non-Return-to-Zero (NRZ) modulation format due to better spectral efficiency and ease of implementation (Zhang et al. 2012). Further, a VLC link length of 0.54 meters was achieved by driving NRZ-modulated data into LED (Jia et al. 2011). Reach enhancement was noticed in VLC by incorporating Green LED at 400 Mbps data streams (Yeh et al. 2018). In NRZ modulated VLC data, dispersion effects are performance deteriorating and there are other spectral efficient modulations available for enhanced performance. Similarly, other modulation formats like Pulse amplitude modulation (PAM) suffer from low receiver sensitivity (Bachtiar et al. 2019), Pulse position modulation (PPM) affected from the service interruption under full-light (Bai et al. 2010), and Pulse width modulation (PWM) has the issue of low data rate carrying potential (Choi et al. 2010). For high-capacity systems, WDM is a robust technology and integrated in 20 channels and a 3 m VLC link serving 7.2 Gbps (Rahman et al. 2020). Nowadays, for enhanced performance and high capacity, VLC data is encoded with multi-level modulation such as QPSK, Orthogonal frequency division multiplexing (OFDM), and Quadrature amplitude modulation (QAM), due to their narrow carrier and potential to tolerate the intersymbol interference. As QAM comes in different variants, 16-QAM modulation with OFDM was investigated at 4 Gbps using a blue laser and filter over a VLC link length of 53 cm (Retamal et al. 2015). VLC systems were also analyzed by incorporating 64-QAM-OFDM and successfully covered the distance of 4.8 m between transmitter and receiver at 1.45 Gbps data speed (Nakamura et al. 2015), a 128-QAM time domain hybrid modulation scheme presented at 0.40 Gbps (Zou et al. 2020), and a 343 Mbps orthogonal circulant matrix transform precoding enabled a bit-interleaved polar-coded modulation was performed in 256-QAM-OFDM VLC link over 80 cm (Wu et al. 2018).

Laser light consists of electromagnetic waves that vibrate in various directions and the orientation of these vibrations is known as the polarization of light. Out of the four properties of Laser light such as frequency ( $f$ ), power, phase, and polarization, only polarization is not fully explored (Kaur et al. 2017), and here comes the need for Polarization Division Multiplexing (PDM) in VLC systems (Xiang-Peng et al. 2021). The capacity improvement is introduced with PDM by leveraging the polarization properties of light to transmit multiple independent data streams over VLC. In PDM, two or more independent data signals are modulated onto light waves with different polarization states (<sup>b</sup>Sachdeva et al. 2022). These signals are typically orthogonal (perpendicular) to each other to avoid interference. The modulated light signals with different polarization states are sent into the optical medium simultaneously. Since the optical wired/wireless mediums preserve the polarization of light, the signals remain separated as they propagate. One of the greatest advantages of the PDM is its ability to be compatible with available multiplexing techniques such as Wavelength division multiplexing (WDM) for further increase data transmission capacity. Scope of PDM is prominent in prolonged reach data transmission and multi-channel fiber optical data communication systems, One common application of PDM is in optical coherent communication systems, where both the phase and amplitude of the optical signals are present for modulation, and the polarization states are carefully controlled to achieve high data rates and signal quality.

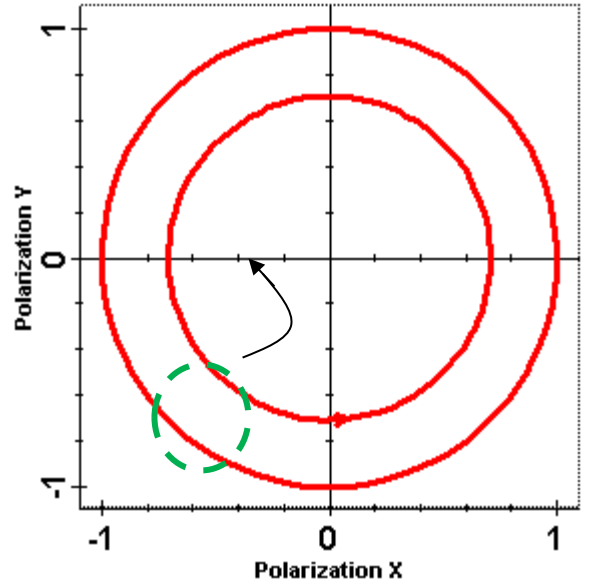
PDM comes in two variants Linear PDM (LPDM) and Circular PDM (CPDM). In wireless optical systems, LPDM is employed extensively in optical communication systems, or free space optics due to horizontal and vertical orthogonal polarizations. LPDM can be generated by varying the electrical field vector component of the electromagnetic wave in one particular direction. However, recently, CPDM has come out to be a better alternative to LPDM due to the elimination of the requirements for polarization axis alignments and provides even scattered light distributions. In CPDM, transmission capacity is tripled and one-fourth reduction in symbol rate can be discerned. Right and left circular polarization states can be formed when the electrical field vector in an electromagnetic wave rotates in either a clockwise or in anti-clockwise direction. Figure 1 represents the two-dimensional visualization for LPDM in horizontal in (a), vertical polarization in (b), right CPDM in (c), and left CPDM in (d). As compared to LPDM, the CPDM is considered a better performance-enhancing technique in wireless optical communications. A high speed 40.665 Gbit/s WDM-VLC link with laser source was demonstrated using Bit-and-power-loading based OFDM (Wei et al. 2019). A 16 channels WDM and 100 Gbps/f enabled LPDM-VLC-QPSK system was investigated over an 8 m link length by employing DSP at the receiver for the eradication of dispersion, nonlinear effects, phase errors, and filtering (<sup>c</sup>Kaur et al. 2022).

Further, a 100 Gbps optical code division multiplexing (OCDMA) dependent 6 m VLC link with the QPSK modulation and DSP was presented in (<sup>d</sup>Kaur et al. 2022).





(c)



(d)

Figure 1 Illustration of 2D depiction of polarization analyzer for (a) Horizontal LPDM (b) Vertical LPDM (c) Right CPDM (d) Left CPDM

A five-user and 100 Gbps per user weight managed OCDMA code was tested over a 5 m VLC link using Laguerre-Gaussian (LG) and Hermite Gaussian (HG) modes (Kaur et al. 2023). An ultra-high capacity 160 Gbps and 22 m VLC link was presented using QPSK (<sup>b</sup>Kaur 2022).

In this work, a performance-enhanced CPDM-based QPSK modulated and IDSP-supported VLC system is presented at 1.6 Tbps. Performance comparison of LPDM-VLC-DSP, CPDM-VLC-DSP, and CPDM- VLC-IDSP has been performed at varied distances, THA, IHA, and different OCA in terms of Q factor, log SER and EVM%. Phase errors, dispersion eradication, nonlinear compensation, equalization, normalization, and signal filtering have been carried out with the incorporation of MATLAB-based IDSP at the coherent QPSK receiver by employing GSOP, LMS, TDEA, and IVA algorithms.

The remaining paper is structured as: Section 2 elaborates the two major types of VLC channels i.e. LoS and non-LoS link, and Section 3 discusses the proposed IDSP module. Section 4 covers the simulation system and parameters considered for the system design in Optisystem. Section 5 elaborates on the investigation of the presented work at varied distances, IHA, THA, and OCA in terms of EVM%, Q factor, log SER and BER. The conclusion and future of the presented work are discussed in Section 6.

## 2. VLC Channel: LoS and NLoS

A VLC can be deployed in both LoS and NLoS scenarios, but the choice depends on the specific application requirements. A LoS-VLC is preferable for high-speed, reliable communication over short distances with a clear line of sight. On the other hand, NLoS-VLC is useful when obstacles or a wider coverage area are a concern, even though it may come with some trade-offs such as susceptibility to interference and data transmission speed. The former can achieve lower bit errors and also offer low latency in the binary biot transmission. However, external light sources, like sunlight or other artificial lighting, can interfere with communication. The NLoS-VLC can work in scenarios where transmitter and receiver have no direct line of

sight but still has reflective surfaces like walls or ceilings. The light signals bounce off these surfaces to reach the receiver. It can extend coverage to areas with obstacles or where direct line of sight is not possible. Moreover, the bouncing or scattering of light can make the communication link less susceptible to interruptions. However, NLoS is susceptible to interference and comes with lower data rates compared to LoS VLC due to the increased complexity of the signal path. VLC channel with LoS link receives power (RP) as expressed in equation (1) (Kaur *et al.* 2022)

$$R_p = T_p H(0) = T_p R(\theta, \phi) G(\psi_c) S_{RX} / D^2 \cos \psi \quad (1)$$

Where,  $R_p$  and  $T_p$  are received and transmitted powers respectively,  $D$  is the separation between the receiver and transmitter,  $H(0)$  shows the channel characteristics considering the one transmitter and receiver. In case of LoS channel, DC channel gain  $H(0)$  is equals to  $H(f)$  and as a results there is no change in the spectrum throughout the VLC channel.  $R(\theta, \phi)$  represents the transmitted light distribution/solid angle,  $\phi$  is azimuth angle in  $xy$ - plane, and  $\theta$  is elevation angle in the direction following negative to positive  $z$ -axis. Receiver photodetector surface is  $S_{RX}$ , gain is  $G(\psi_c)$ ,  $\cos \psi$  denotes the transmitter and receiver path inclination, field of view (FOV) is  $\psi_c < \pi/2$ ,  $n$  is refractive index in the light concentrating lens. Relationship between  $n$ ,  $\psi_c$  and  $G(\psi_c)$ , is given as

$$G(\psi_c) = n^2 / \sin^2 \psi_c \quad (2)$$

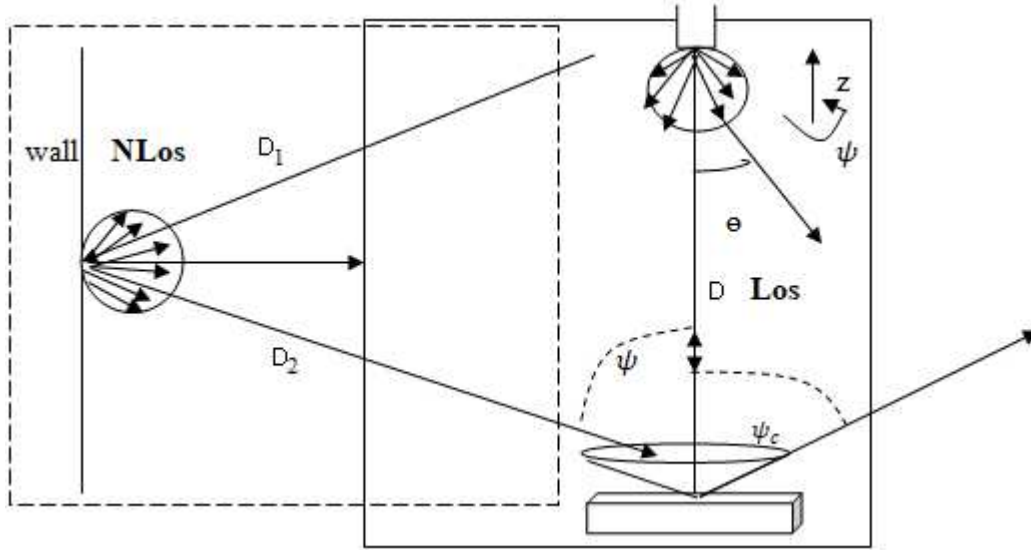


Figure 2 A VLC structure representing LoS and NLoS models

In NLoS variant of VLC channel, light coming from the transmitter to the receiver by striking the sidewalls ( $w$ ), is also calculated using the equ. (3). The nature of the reflected light coefficient ( $\rho_i$ ) depends on the quality of the light reflecting surface and the reflections can be of three types such as Diffuse, Specular or both diffuse+specular.  $h^{w-1}$  is the DC gain of channel due to reflection from walls. Reflections models are Phong and Lambertian models, and the total  $R_p$  is the sum of multiple reflected rays ( $M_r$ ) as expressed in equ. (3). Also, it is assumed that phase deviation is null between rays and therefore phase errors of the source are ignored. Highest power of reflected light is assumed from 1<sup>st</sup> reflection  $w = 1$  and multiple reflections are denoted by  $M_r$  (Kaur *et al.* 2022).

$$R_P = T_P(H(0)) + H^{Mr} = T_P \left( H^0 + \sum_{i=1}^w (\rho_i R(\phi) G(\psi_c) \frac{S_{RX}}{D_1^2 D_2^2} \cos \psi) + h^{w-1} \right) \quad (3)$$

### 3. Proposed MATLAB coded IDSP for Distortion Compensation

The polarization of the optical signal fluctuates erratically following long-distance transmission. The transmission can also be hampered by the local oscillator (LO) laser's  $f$  difference from the transmitter laser. High-speed DSP technology is the industry's go-to method for eradicating polarization mode dispersion (PMD), chromatic dispersion (CD), various distortion correction, and  $f$ , phase, as well as other distortions. The original data is finally restored in a symbolic move. Multi or bi-polarization QPSK encoded data transmission has attracted much attention due to the drastic fall in the inter-symbol interference and offers two-fold bandwidth spectrum efficiency (Kaur et al 2022). In CPDM-based QPSK transmission, two halves of a single laser are enabled with circular polarization right and left. A CPDM-modulated QPSK signal is shown as

$$Q_{CPDM-QPSK, CP_R} = Q_{CP_R}(t) \exp(j2\pi f_c t) \exp[P_{m,CP_R}(t)] \quad (4)$$

$$Q_{CPDM-QPSK, CP_L} = Q_{CP_L}(t) \exp(j2\pi f_c t) \exp[P_{m,CP_L}(t)] \quad (5)$$

Where,  $f_c$  is carrier wave freq.,  $Q_{RCP}(t)$  and  $Q_{LCP}(t)$  are the amplitudes of  $CP_R$  and  $CP_L$  polarization states, and phase of these polarizations are  $P_{m,RCP}$  and  $P_{m,LCP}$  respectively. The received signals at the receiver for RCP are expressed in equ. (6) and (7) for I and Q signals and for LCP in (8) and (9).

$$CP_{R_I} = R\sqrt{2L_{LO}}Q(t)[\cos(2\pi(f_c - f_{LO})t) + P_{m,CP_R}(t) - \phi_{LO}(t)] \quad (6)$$

$$CP_{R_Q} = R\sqrt{2L_{LO}}Q(t)[\cos(2\pi(f_c - f_{LO})t) + P_{m,CP_R}(t) - \phi_{LO}(t)] \quad (7)$$

$$CP_{L_I} = R\sqrt{2L_{LO}}Q(t)[\cos(2\pi(f_c - f_{LO})t) + P_{m,CP_L}(t) - \phi_{LO}(t)] \quad (6)$$

$$CP_{L_Q} = R\sqrt{2L_{LO}}Q(t)[\cos(2\pi(f_c - f_{LO})t) + P_{m,CP_L}(t) - \phi_{LO}(t)] \quad (7)$$

Where,  $\phi_{LO}$ ,  $L_{LO}$ ,  $f_{LO}$  are the phase, launched power and  $f$  of the local oscillator,  $CP_{R_I}$ ,  $CP_{R_Q}$  are in-phase and quadrature phase components of Right circular polarization and  $CP_{L_I}$ ,  $CP_{L_Q}$  for left circularly polarized signals respectively. Irrespective of the synchronization and balanced detection of symbols, there is still I/Q mismatch in orthogonality due to photo-diode mismatch, offset bias points and other issues. GSOP algorithm removes the fluctuations in the I/Q symbols and make them near to perfectly orthogonal (Chang *et al.* 2009). Time-varying effects on the optical transmission connection include PMD, CD, and polarization rotation. For CD eradication in the system, equalization is carried out by employing fixed coefficient based two FIR filters and improved electric signal ability to offer better suppression of CD.

The remaining CD damage, full polarisation demultiplexing, and PMD are then compensated with minimum tap based butterfly adaptive filter (Savory *et al.* 2008). The adaptive filter decide the tap coefficient by considering the LMS method (Zhang *et al.* 2014). The conventional M times approach of estimation feed-forward all digital phase  $f$  offset estimation is commonly incorporated technique in the coherent optical transmission system receiver. To eliminate signal phase modulation, phase modulation information is implemented M times. M is 2 for CPDM-QPSK (Zhang *et al.* 2021). In conventional DSP modules, following algorithms are used such as BPS, VPE, and CMA algorithms (Sachdeva et al. 2023; Kakati et al. 2018; Chauhan et al. 2021). Figure 3 depicts the proposed IDSP for signal compensation using GSOP, LMS, TDEA, and IVA algorithms.

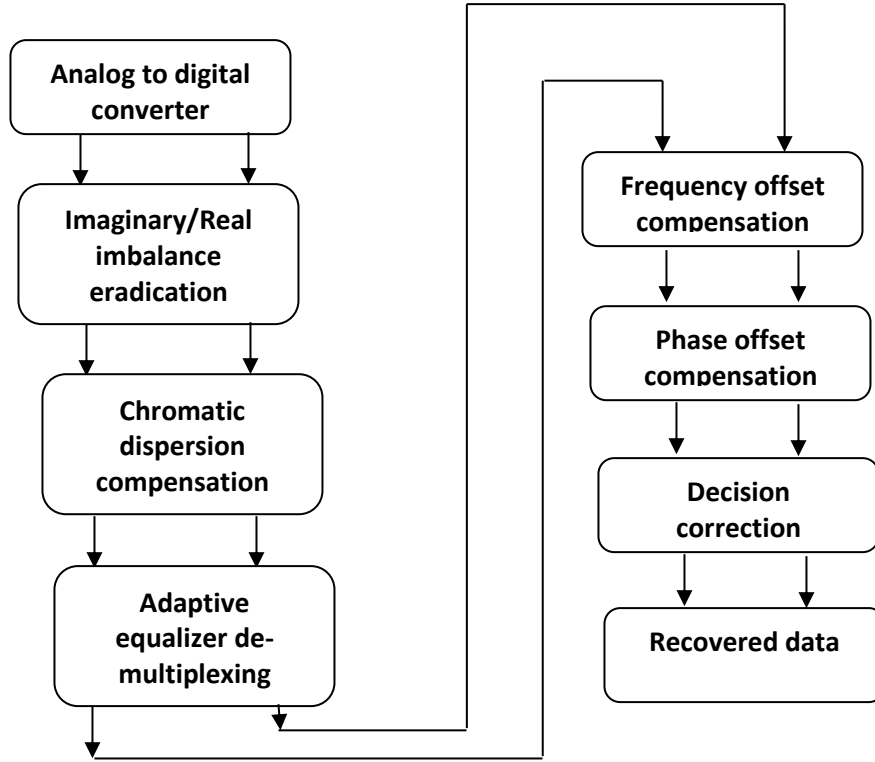


Figure 3 Block diagram of proposed IDSP using GSOP, LMS, TDEA, and IVA algorithms

For the implementation of proposed IDSP, all the algorithms are simulated in MATLAB version 2019 and further called in Optisystem software version 2020. The  $I$  and  $Q$  symbols stabilization is the first stage of DSP module and the fluctuations in these symbols are compensated using GSOP algorithm. Further, in order to mitigate the CD in the symbols, TDEA algorithm is incorporated in the IDSP second stage. In the third stage, polarization de-multiplexing is performed for the eradication of PMD dispersion and IVA algorithm is employed for the compensation of phase errors as well as  $f$  offset. In final stage, LMS algorithm does the operation of directing symbols towards ideal symbol points.

#### 4. Simulation Design of Proposed CPDM-VLC System

The presented system as illustrated in Figure 4 is constructed in the latest version of Optisystem software and MATLAB code for IDSP module is called in Optisystem from MATLAB 2019.

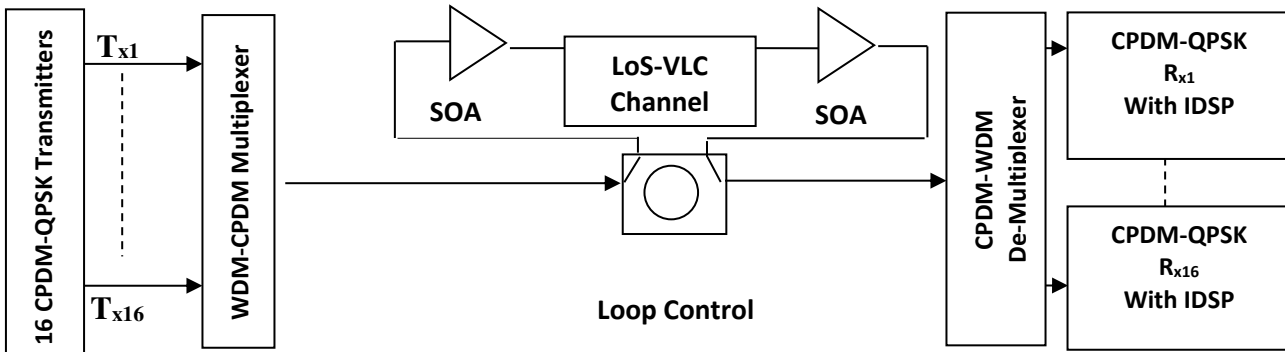


Figure 4 Presented 16 Tx/Rx WDM-CPDM-VLC system employing IDSP



A multi-channel CPDM-VLC system at 100 Gbps per channel is simulated using 16 WDM channels each at 100 GHz  $f$  spacing. Visible light yellow frequencies starting from 540 THz and upto 541.5 THz covering 16 frequencies are taken at launched power of 30dBm/transmitter. All the WDM channels are modulated with multi-level spectral efficient QPSK modulations loaded with CPDMs. Each CPDM-QPSK transmitter enables with the 100G binary data generator (BDG), serial to parallel data converter (S/P), QPSK modulators such as  $M_1$  and  $M_2$ , electrical to optical (E/O) modulators, RCP and LCP polarization components divided from single laser source using polarization beam splitter (PBS), and polarization combiner (PC), semiconductor optical amplifiers (SOAs), and LoS-VLC channel as illustrated in Figure 5 (a).

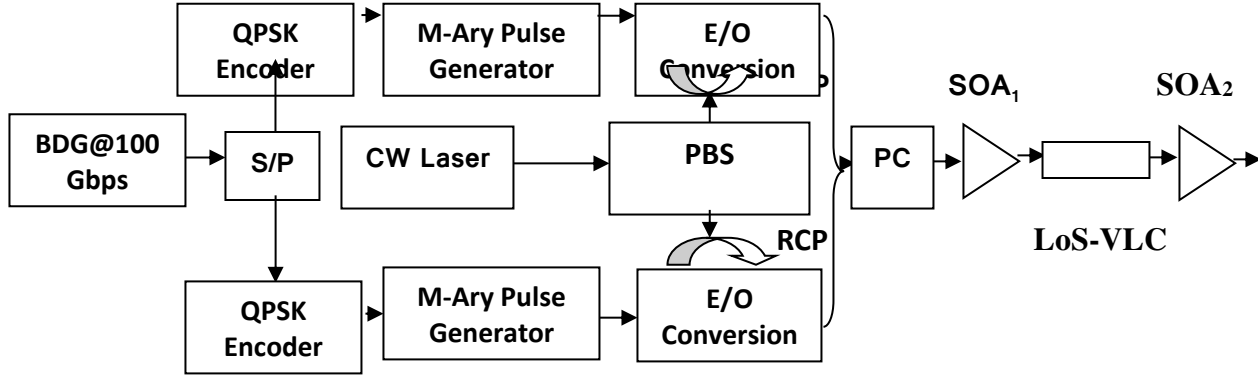


Figure 5 (a) Proposed single channel CPDM-QPSK transmitter

Table 1 Simulation parameters of presented CPDM-QPSK-VLC system

Simulation Parameters	Values Selected
BDG speed/ $f$	100 Gbps ( <sup>c</sup> Kaur <i>et al.</i> 2022)
Input power/ $f$ and overall capacity	30 dBm and 1.6 Tbps ( <sup>c</sup> Kaur <i>et al.</i> 2022)
$f$ range and spacing	540-541.5 THz and 100 GHz ( <sup>c</sup> Kaur <i>et al.</i> 2022)
Modulation	CPDM (Ghatwal <i>et al.</i> 2023; <sup>a</sup> Sachdeva <i>et al.</i> 2023) –QPSK with IDSP
IDSP algorithms	GSOP, LMS, TDEA, and IVA algorithms
Symbol rate	25 Gsymbols/sec
SOPs	2, RCP/LCP (Ghatwal <i>et al.</i> 2023; <sup>a</sup> Sachdeva <i>et al.</i> 2023)
Bits Per Symbol	2
Amplifiers	SOAs ( <sup>b</sup> Kaur <i>et al.</i> 2022)
Length of LoS-VLC channel	2-30 m
THA	40-80 degrees ( <sup>c</sup> Kaur <i>et al.</i> 2022)
IHA	10-50 degrees ( <sup>c</sup> Kaur <i>et al.</i> 2022)
OCA	0.5-2.5 cm <sup>2</sup> ( <sup>c</sup> Kaur <i>et al.</i> 2022)

The LoS-VLC channel has different parameters such as IHA, THA and OCA that purely decides the final performance of the overall system and simulation parameters are discussed in the Table 1. There are two SOA amplifiers placed in the transmission loop such as  $SOA_1$  as a post amplifier to the LoS-VLC channel and  $SOA_2$  placed as a post amplifier (<sup>b</sup>Kaur *et al.* 2022). Signals after travelling through the VLC channel reaches at the receiver unit and for the respective  $f$  demodulation, a  $1 \times 16$  channel de-multiplexer is employed. Further, each receiver has comprised of local oscillator, CPDM decoders, PBS, couplers, balanced photo-detection and combiners. All the de-modulated signals then electrically amplified using 20 dB electrical amplifiers (EAs) followed by low pass Bessel filtering. An IDSP is unit as discussed in Section 3 is employed in the receiver for Q/I stabilization, CD eradication, time recovery, re-sampling, adaptive equalization, phase error reversal, nonlinear compensation and  $f$  offset. Constellation analyzers are placed after IDSP for EVM%, Q factor, and SER calculations. M-Ary threshold detectors extract the different amplitude-phases and further QPSK decoders provide the parallel decoded data. At the final stages, parallel data converted into serial and errors are calculated using bit error rate test set (BERT) as shown in Figure 5 (b).

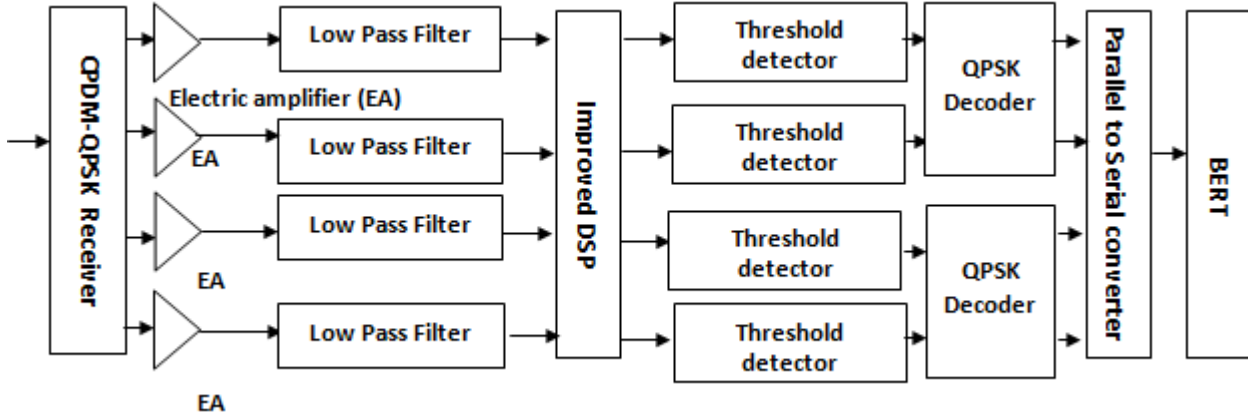


Figure 5 (b) CPDM-QPSK receiver and structure of complete receiver unit/ $f$

## 5. Performance Investigation of Proposed CPDM-QPSK System

The presented system has a narrower optical carrier spectrum due to the integration of RCP and LCP that offers the least signal blocking as well as provides even distribution of the polarization as compared to the LPDM, where more power losses occurs and synchronization of the axis is required. Figure 6 (a) depicts the RCP/LCP modulated narrow optical carrier spectrum for channel 1 and Figure 6 (b) depicts the multiplexed carrier spectrums at 100 GHz channel spacings. Single CPDM-QPSK spectrum shows higher power as compared to the WDM-CPDM due to higher insertion losses.

Figure 7 represents the detailed comparison of three different cases in LoS-VLC system such as LPDM-VLC-DSP, CPDM-VLC-DSP and CPDM-VLC-IDSP in terms of (a) log SER, (b) Q factor and (c) EVM%. From Figure 7 (a), it is discerned that log SER increases as the VLC range prolongs from 2 m to 16 m for all investigated cases. There are multiple factors affecting the VLC performance such as open air attenuation, scatterings, ambient noise sources and inter-symbol interferences. For the signal strengthening, cost effective, compact, and VLC channel mounted pre SOA and post SOA are employed.

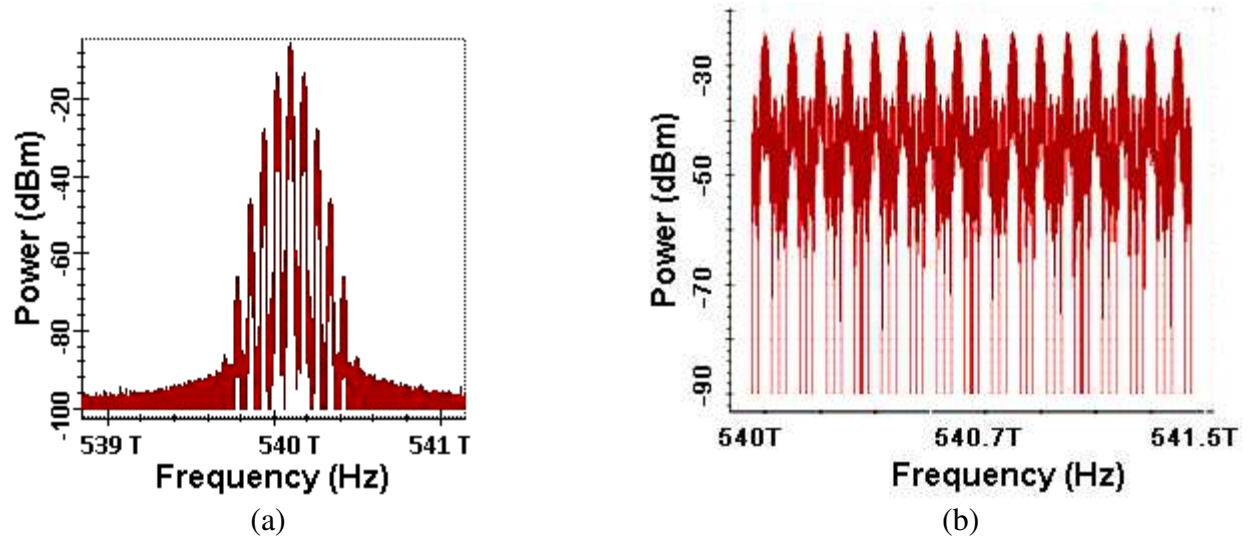
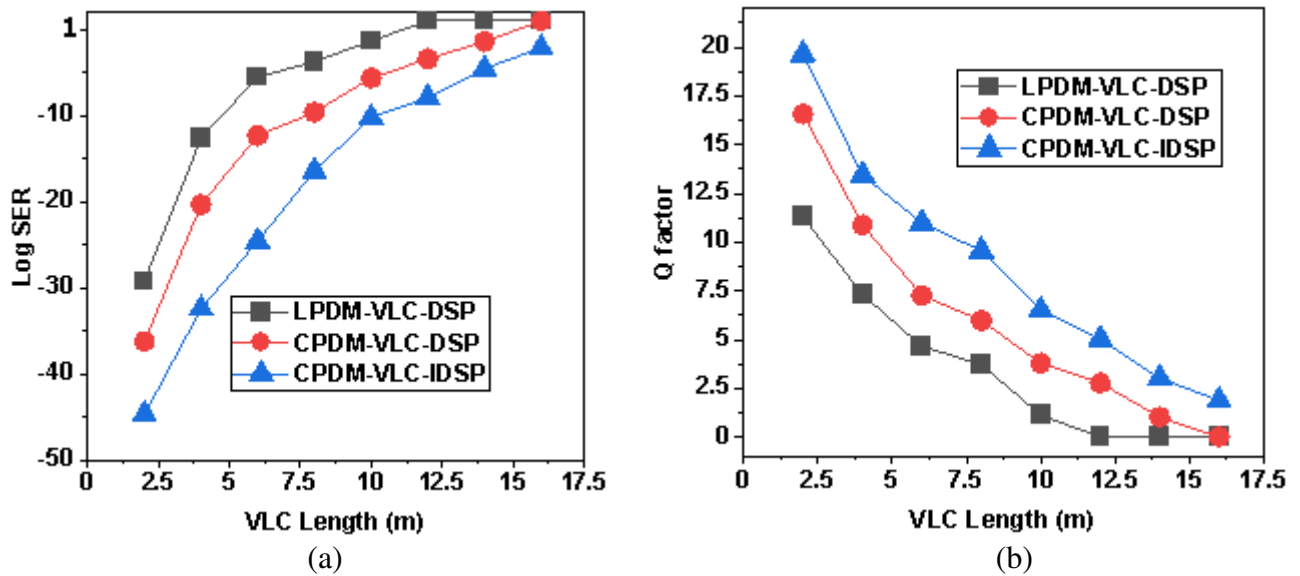
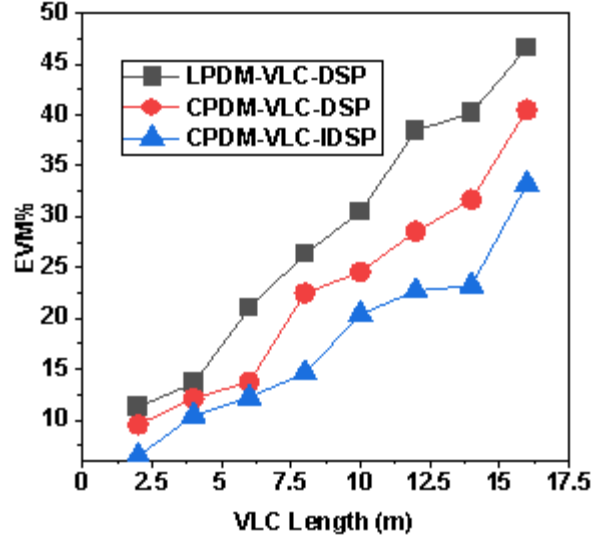


Figure 6 CPDM modulated optical carrier spectrums for (a) single channel (b) 16 WDM channels

Rest of the performance deteriorating issues like CD, nonlinear effects, time delays and  $I/Q$  symbol instabilities are eradicated by DSP modules. Scattered light uneven distribution, need for the synchronization of axis alignments, and inability to carry high input powers are the various reasons for the highest log SER (1) and EVM% (40.2%) (Figure 7(c)) in case of LPDM-VLC-DSP at 14 m. On the other hand, the CPDM-VLC-DSP system offers slight better performance in terms of log SER (-1.4) and EVM% (31.65%) (Figure 7(c)) at 14 m due to the uniformity in the scattered light, no need for axis alignment synchronizations and has potential to support higher power levels. However, results revealed that both the cases have employed conventional DSP for the signal conditioning which are not upto the mark. Therefore, in third case, DSP is replaced with IDSP by incorporating better algorithms such as GSOP, LMS, TDEA, and IVA as compared to BPS, VPE and CMA algorithms. The least EVM% (23.2%) (Figure 7(c)) and log SER (-4.53) is offered by CPDM-VLC-IDSP at 14 m. Moreover, in Figure 7 (b), the variations of Q factor with LoS-VLC link range is given and the highest Q factor is observed in case of CPDM-VLC-IDSP and least Q factor is portrayed in LPDM-VLC-DSP.

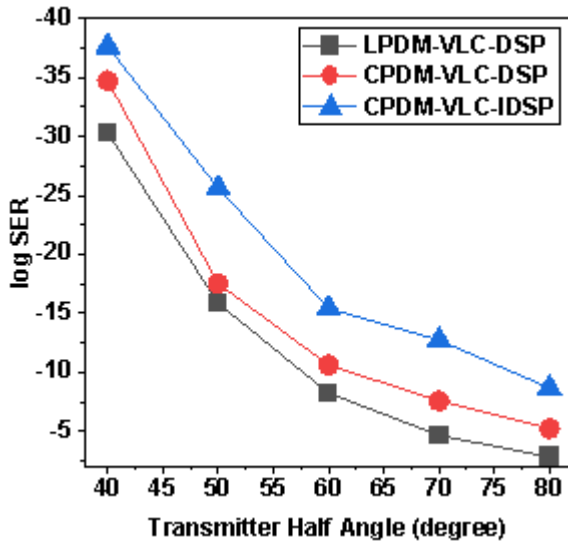




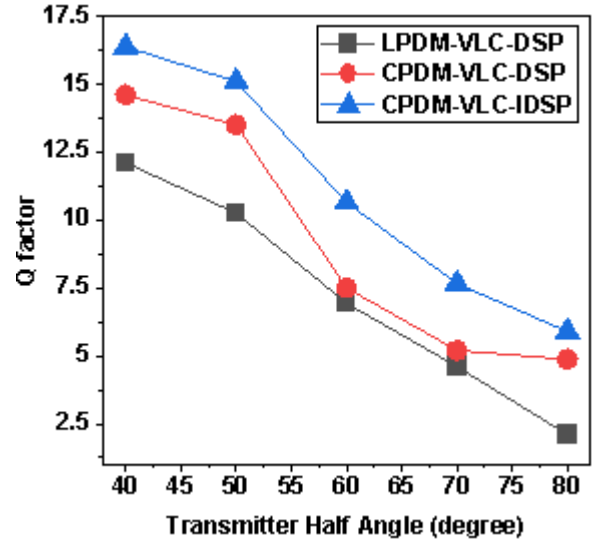
(c)

Figure 7 Performance comparison of three different cases at different LoS-VLC link ranges in terms of (a) log SER (b) Q factor and (c) EVM%

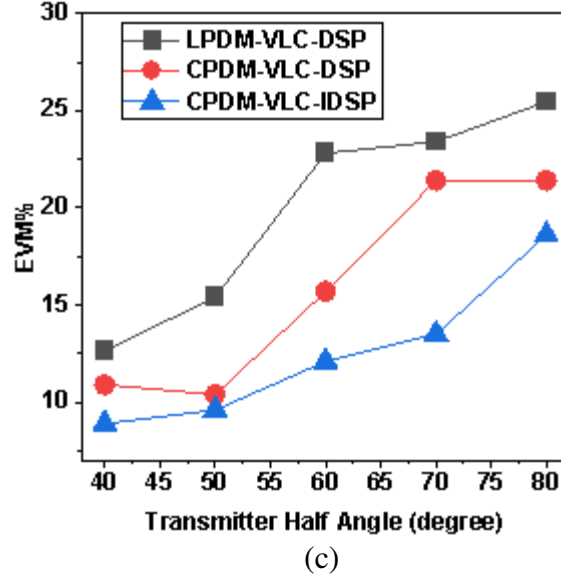
In LoS-VLC systems, receiver average power (RAP) and fluctuation in power are the important parameters that are affected to the fullest by THA of VLC. With the increase in the THA angles, reduced RAP values are calculated and therefore lower Q factor values are exhibited in all three investigated cases. On contrary, higher values are seen in case of EVM% and log SER at higher THA angles. Figure 8 depicts the log SER versus LoS-VLC link range for all the three cases as shown in (a), Q factor versus LoS-VLC in (b) and EVM% versus LoS-VLC link range in (c). At lowest considered THA angle  $40^\circ$ , the least log SER and EVM% values are obtained such as -37.6 and 8.9% at 2 m LoS-VLC range in case of CPDM-VLC-IDSP.



(a)



(b)

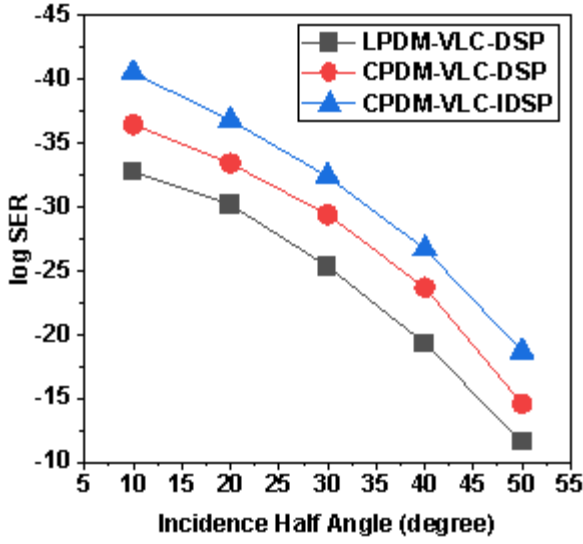


(c)  
Figure 8 Effects of THA on the three different VLC systems in terms of (a) log SER (b) Q factor and (c) EVM%

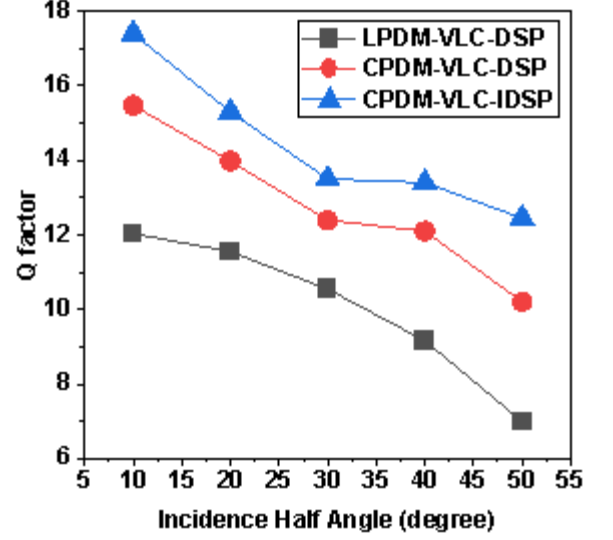
With the increase in the THA from  $50^\circ$  to  $80^\circ$ , the values of log SER and EVM% changes from -25.6 to -8.65 and 9.6% to 18.6% respectively for CPDM-VLC-IDSP as shown in Figure 8 (a) and (c). Similarly, values of Q factor are 16.37 to 5.89 that are received for THA angles  $40^\circ$  and  $80^\circ$  in case of PDM-VLC-IDSP as illustrated in Figure 8(b). The highest EVM%, log SER and least Q factor has been checked for LPDM-VLC-DSP.

We often assume that receivers and transmitters are positioned in parallel when localizing particular receivers in LoS-VLC systems. This assumption is challenging to prove, though, because in reality, receivers move arbitrarily. Considering the situation when the transmitter and receiver are not exactly parallel and the slanted angles and separations of transmitter/receiver from the illuminators can assist you spot variations in optical gain. For getting the higher transmission rate and better performance, optimal values of receiver IHA are required under illumination constraints. Figure 9 (a) and 9 (c) represents the IHA angle variations for VLC systems and validated the outcomes in terms of log SER and EVM%. Starting values of IHA were fixed to  $10^\circ$  and increased till  $50^\circ$  and it is cleared that higher the IHA angles, more will be the EVM% and log SER. CPDM-VLC-IDSP showed lowest log SER (-40.5) and EVM% (6.45%) values for  $10^\circ$  IHA and highest for  $50^\circ$  in term of log SER (-11.6) and EVM% (14.19%) using LPDM-VLC-DSP. The Q factor values increases at lower IHA angles and tend to reduce as we go towards higher IHAs.

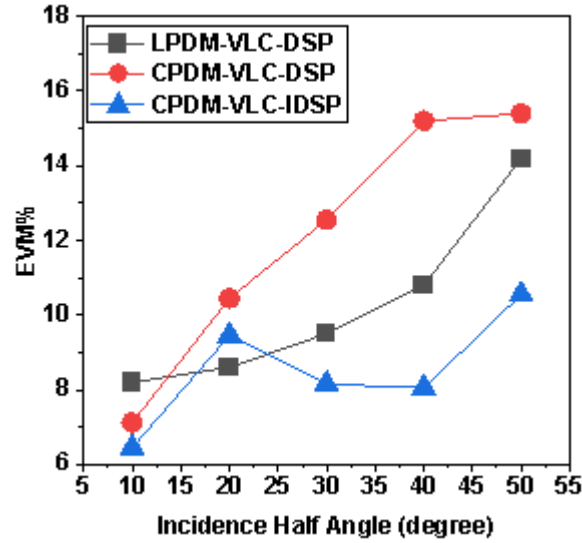
In VLC systems, at the receiver, FOV needs to be narrower for achieving large concentration because flux gathering potential becomes weaker in optical concentrators showing wide FOV. The compound parabolic concentrator consists of two parabolic concentrators directing the light towards the each other's bottom corners. The side walls reflect the light entered towards aperture and divert it towards exit plane. Concentrator in this work is made up of dielectric material and having shape of solid paraboloid. Area of the optical concentrator i.e. OCA plays an important role in deciding the performance of the VLC systems and therefore, in Figure 10, performance of proposed system and conventional systems (<sup>c</sup>Kaur *et al.* 2022; <sup>b</sup>Kaur *et al.* 2022) is compared at different OCAs. The range of OCA is varied from  $0.5 \text{ cm}^2$  to  $2.5 \text{ cm}^2$  and performance is analyzed in terms of EVM%, SER, and Q factor.



(a)



(b)



(c)

Figure 9 Variations of (a) log SER (b) Q factor and (c) EVM% with respect to different IHA angles

Figure 10 (a) and (c) represents that wider OCA increase the log SER and EVM% in case of CPDM-VLC-DSP/IDSP but poor power handling incompetency in LPDM-VLC-DSP experience nonlinear effects and therefore a sharp peak (maximum log SER and EVM%) is observed at 1 cm<sup>2</sup> OCA. Q factor decrease has been observed for wider FOV and least values is seen at 1 cm<sup>2</sup> OCA for LPDM-VLC-DSP due to kerr's effects as shown in Figure 10 (b). Therefore, CPDM-VLC-IDSP system offers best performance at lower OCAs due to potential of carrying high power signals.

A received test pattern or received data sequence in the BERT component is compared with the sequence of transmitted binary bits or predetermined stress patterns in the form of logical 1's and 0's and final BER is calculated from this component. Figure 11 illustrates the BER values at 14 cm VLC link range for LPDM-VLC-DSP, CPDM-VLC-DSP and CPDM-VLC-IDSP. Results disclosed that -4 BER is observed for CPDM-DSP-IDSP, -1 for CPDM-VLC-DSP and 100% errors bits are seen for LPDM-VLC-DSP. The successful data



transmission for maximum reach 14 cm is witnessed in case of CPDM-VLC-IDSP, 12 cm for CPDM-VLC-DSP, and 8 cm for LPDM-VLC-DSP.

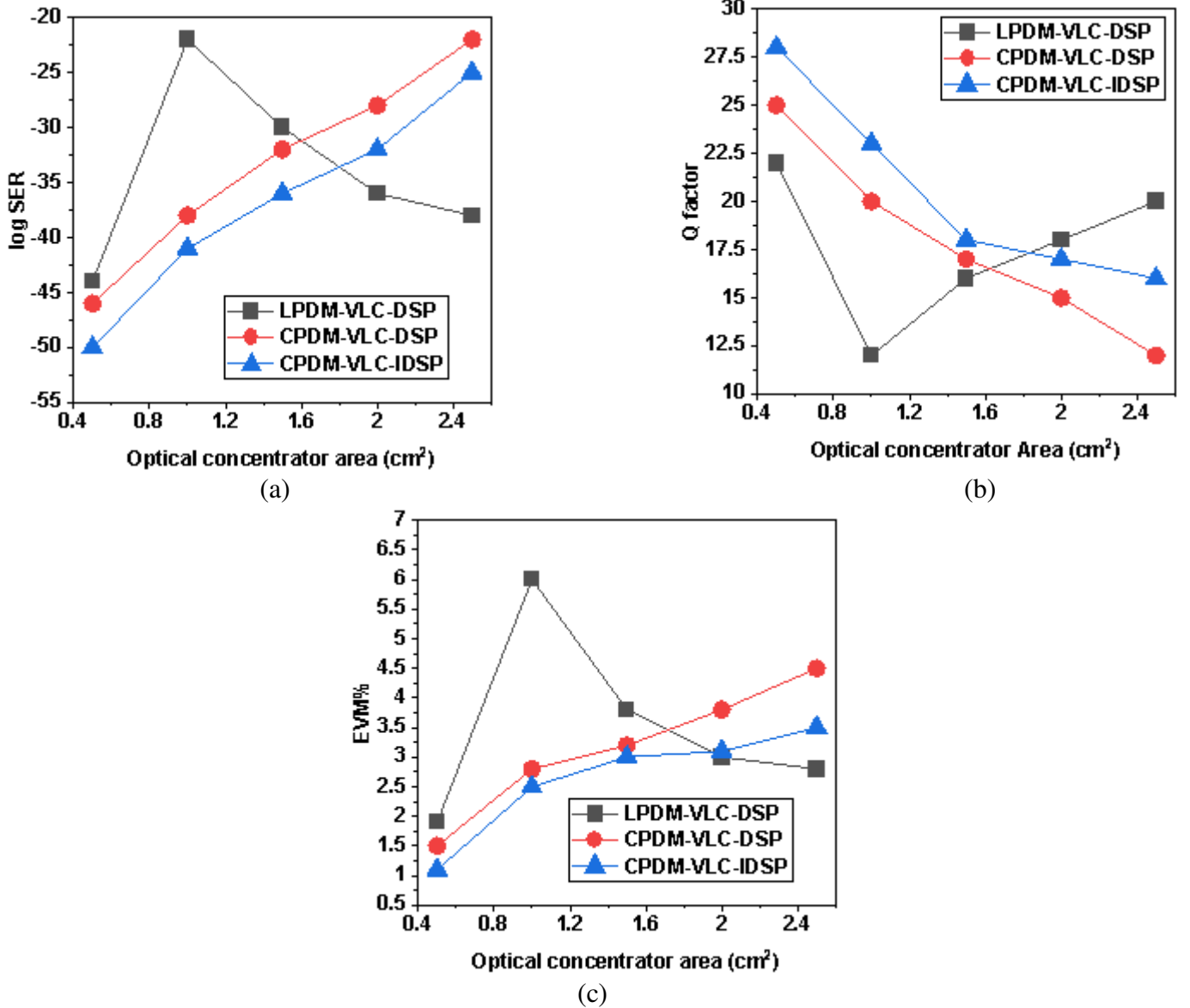


Figure 10 Comparison of investigated configurations at different OCA values in terms of (a) log SER (b) EVM% and (c) Q factor

Figure 12 shows the constellation diagrams for (a) LPDM-VLC-DSP (b) CPDM-VLC-DSP and CPDM-VLC-IDSP at 14 cm VLC link range. From Figure (a), it is discerned that IDSP provides the best symbol placements in the constellation quadrants and effectively compensates the effects of phase errors, CD, nonlinear effects, time delays, inter-symbol interferences, noises, and symbol shape distortions. In case of the LPDM-VLC-DSP system, both the LPDM and DSP are not competent to carry high power and effective eradication of phase errors, CD, nonlinear effects, time delays, inter-symbol interferences, noises, symbol shape distortions, and therefore shows 100% bit errors at 14 cm VLC link range.

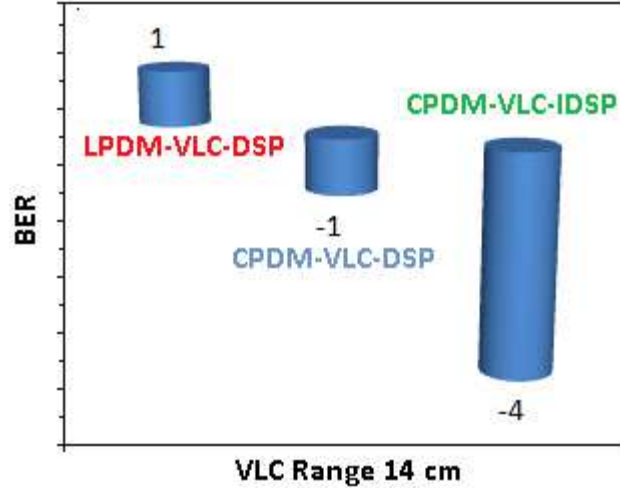


Figure 11 BER values at BERT at 14 cm VLC link range

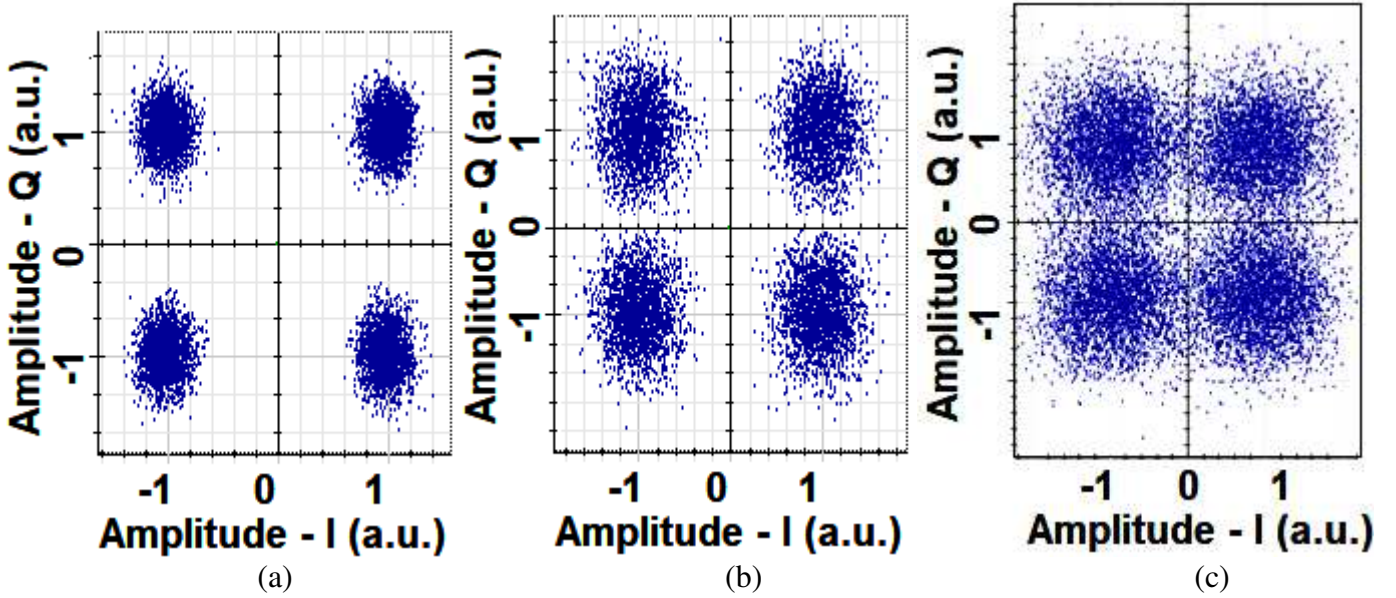


Figure 12 Constellation diagrams at 14 cm VLC link in case of (a) CPDM-VLC-IDSP (b) CPDM-VLC-DSP and LPDM-VLC-DSP

Our presented system has different enhancements as compared to the reported VLC systems such as (1) CPDM is introduced for the first time in 1.6 Tbps VLC systems (2) DSP is replaced with IDSP by using enhanced algorithms (3) different VLC parameters are investigated and outcomes revealed about the use of lower THA, IHA, and OCA for better results and (4) SOA is used in the place of EDFA due to cost effectiveness, compact size, and VLC mounted operations. In the comprehensive literature survey, different research studies were presented in the field of VLC (<sup>c</sup>Kaur *et al.* 2022; <sup>b</sup>Kaur *et al.* 2022; <sup>d</sup>Kaur *et al.* 2022; Kaur *et al.* 2023) and these researches are compared with presented work in terms of output and input parameters as shown in Table 2. It is discerned that presented system has maximum VLC range covered at 1.6 Tbps capacity using CPDM-IDSP. In the near future, proposed study can be enhanced by using narrower carrier spectrum based modulations, and by incorporating mode division multiplexing.

Table 2 Comparison of presented CPDM-VLC-IDSP system with existing VLC systems

Parameters	<sup>b</sup> Kaur <i>et al.</i> 2022	<sup>c</sup> Kaur <i>et al.</i> 2022	<sup>d</sup> Kaur <i>et al.</i> 2022	Presented Work
Data Rate	10 Gbps	100 Gbps	100 Gbps	100 Gbps
Channels/Users	16	16	5	16
Capacity	160 Gbps	1.6 Tbps	500 Gbps	1.6 Tbps
VLC Range Achieved	22 m	8 m	6 m	14 m
Modulation	QPSK	QPSK	QPSK	QPSK
Multiplexing	WDM-LPDM	WDM-LPDM	WDM-OCDMA-LPDM	WDM-CPDM
DSP Algorithms	BPS, VPE, and CMA	BPS, VPE, and CMA	BPS, VPE, and CMA	GSOP, LMS, TDEA, and IVA
Amplifier	EDFA, SOA	EDFA	EDFA	SOA
BER @ VLC range	-3.5 at @ 22 m using SOA	-2.42 @ 7 m	-3.42 @ 6 m	-4 @ 14 m

## 6. Conclusion

In the presented simulation investigation, a yellow light based WDM-VLC system is designed to unleash an immense amount of underutilized visible light region. The potential of CPDM and IDSP has been manifested in VLC system at 100 Gbps data transfer rate over 14 m link range. A performance comparison of LPDM-VLC-DSP, CPDM-VLC-DSP and CPDM-VLC-IDSP has been performed at varied distances, THA, IHA, and different OCA in terms of log SER, EVM%, Q factor, and log BER. Phase errors, CD eradication, nonlinear compensation, equalization, normalization, and signal filtering has been carried out with the incorporation of MATLAB based IDSP at the coherent QPSK receiver by employing GSOP, LMS, TDEA, and IVA algorithms. Our presented system has different enhancements as compared to the reported VLC systems such as (1) CPDM is introduced in literature for 1.6 Tbps VLC systems (2) DSP is replaced with IDSP by using enhanced algorithms (3) different VLC parameters are investigated and outcomes revealed about the use of lower THA, IHA, and OCA for better results and (4) SOA is used in the place of EDFA due to cost effectiveness, compact size, and VLC mounted operations. The presented CPDM-VLC-IDSP system provides least log SER (-4.53), EVM% (23.2%) at 14 m, log SER -37.6 and EVM% (8.9%) for 40<sup>0</sup> THA at 2 m, log SER (-40.5) and EVM% (6.45%) for 10<sup>0</sup> IHA at 2m, and log SER (-50), EVM% (1.1%) for 0.5 cm<sup>2</sup> at 2m. Results disclosed that -4 BER is observed for CPDM-DSP-IDSP, -1 for CPDM-VLC-DSP and 100% errors bits are seen for LPDM-VLC-DSP. The successful data transmission for maximum reach 14 cm is witnessed in case of CPDM-VLC-IDSP, 12 cm for CPDM-VLC-DSP, and 8 cm for LPDM-VLC-DSP. In the near future, proposed study can be enhanced by using narrower carrier spectrum based modulations, and by incorporating mode division multiplexing.

## Ethical Approval

not applicable

## Competing interests

Work is performed in Lovely Professional University, Punjab, India

## Authors' contributions

H.P., M.V., S.S., S.K., M.S., and M.K.S. have directly participated in the planning, execution and analysis of this study. H.P. drafted the manuscript. All authors have read and approved the final version of manuscript.

**Funding**

Not applicable

**Availability of data and materials**

Not applicable

**References**

- Angurala, M., Singh, H., Anupriya et al. Testing Solar-MAODV energy efficient model on various modulation techniques in wireless sensor and optical networks. *Wireless Netw* 28, 413–425 (2022). <https://doi.org/10.1007/s11276-021-02861-2>.
- Armstrong, J. OFDM for Optical Communications. *Journal of Lightwave Technology*. 27, 189-204, (2019). <https://doi.org/10.1109/JLT.2008.2010061>.
- Bachtiar, Y.A., Adiono, T.: PAM-4 Modulator-Demodulator Design Modeling for Visible Light Communication (VLC). In: International Symposium on Electronics and Smart Devices (ISESD), pp. 1-5. IEEE, Badung, Indonesia (2019)
- Bai, B., Xu, Z., Fan, Y.: Joint LED dimming and high capacity visible light communication by overlapping PPM. In: 19th Wireless and Optical Communications Conference (WOCC 2010), pp. 71–75 (2010).
- Chauhan, A., Yadav, I., Dhawan, P., Kaur, S. and Verma, A. Nonlinear/dispersion compensation in dual polarization 128-QAM system incorporating optical backpropagation. *Journal of Optical Communications*, 000010151520200282 (2021). <https://doi.org/10.1515/joc-2020-0282>
- Choi, J.-H., Cho, E.-B., Kang, T.-G., Lee, C.G.: Pulse width modulation based signal format for visible light communications. In: 15th Optoelectronics and Communication Conference (OECC 2010), pp. 276–277 (2010)
- Chang, S. H., Chung, H. S., Kim, K. Impact of Quadrature Imbalance in Optical Coherent QPSK Receiver. *IEEE Photonics Technology Letters* 21(11), 709 – 711 (2009).
- Ghatwal, S., Saini, H. A high power 320 Gbps CPDM-256-QAM based Ro-FSO system enabling 80 GHz under rain and haze effects. *Optical and Quantum Electronics* 55, 80 (2023).
- Jia, YP., Ye, WL., Tian, CW. et al. Numerical analysis on an OOK-NRZ visible light communication system based on a single white LED. *Optoelectron. Lett.* 7, 376 (2011). <https://doi.org/10.1007/s11801-011-1050-7>.
- Kaur, H., Singh, N. Security enhancement of an integrated mode division multiplexed VLC system using two-dimensional WMZCC codes. *Journal of Optics*, (2023). [doi.org/10.1007/s12596-023-01216-8](https://doi.org/10.1007/s12596-023-01216-8).
- <sup>a</sup>Kaur, S., Kumar, M., Verma, A. An Integrated High-Speed Full Duplex Coherent OFDM-PON and Visible-Light Communication System. *Journal of Optical Communications* 43(3), 379-383 (2022).
- <sup>b</sup>Kaur, S. Performance Investigation of Visible Light Communication System Employing PDM-QPSK and EDFA/SOA Amplifier. International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON), Bengaluru, India, 22-24 December, 2022.
- <sup>c</sup>Kaur, H., Singh, N. Ultra high-speed VLC system using polarization division multiplexed QPSK, DSP, and matched filters. *Optical and Quantum Electronics* 54(636), 1-17 (2022).

<sup>d</sup>Kaur, H., Singh, N. Performance enhancement of visible light communication (VLC) system incorporating WMZCC-OCDMA codes and PDM-QPSK-DSP data encoding. *International Journal of Communication systems* 36(1), 1-15 (2022).

<sup>e</sup>Kaur, S., Sachdeva, S., Sindhwani, M. 400 Gb/s free space optical communication (FSOC) system using OAM multiplexing and PDM-QPSK with DSP. *Journal of Optical Communications*, (2022). <https://doi.org/10.1515/joc-2022-0111>

Kishore, N., Senapati, A. 5G smart antenna for IoT application: A review. *Int. Jour. Comm. Sys.* 35(13), e5241 (2022).

Kaur, S., Verma, A.: From Communication to Illumination: Visible Light Communication, Pros and Cons, Applications, Current and Future Trends, State-of-the-art Discussion. In: SSRN, Available at SSRN: <https://ssrn.com/abstract=3868998> or <http://dx.doi.org/10.2139/ssrn.3868998>, (2021)

Kaur, S., Kumar, M., Verma, A. Visible light communication employing multiplexing of different colours. *IJRAR* 6, 300-304 (2019).

Kaur, S., Kaur, G., Singh, G., Verma, A., Julka, N. Polarization Crosstalk Suppression in Wavelength Division Multiplexed Free Space Optical System Incorporating Polarization Diversity. *IJCRT* 5(3), 384-390 (2017).

Kakati, D., Arya, S. C. A full-duplex optical fiber/wireless coherent communication system with digital signal processing at the receiver. *Optik* 171, 190-199 (2018).

Lu, H-H, et al. A 400-Gb/s WDM-PAM4 OWC system through the free-space transmission with a water–air–water link. *Nature, Scientific reports* 11, 21431(1-9) (2021).

Miras, D., Maret, L., Maman, M., Laugeois, M., Popon, X., Ktenas, D. A High Data Rate LiFi Integrated System with Inter-cell Interference Management. *IEEE Wireless Communications and Networking Conference (WCNC)*, IEEE, Barcelona, Spain, 15-18 April, 2018.

Nakamura, K., Mizukoshi, I., and Hanawa, M. Optical wireless transmission of 405 nm, 1.45 Gbit/s optical IM/DD-OFDM signals through a 4.8 m underwater channel. *Opt. Express* 23, 1558-1566 (2015).

Rahman, M. T., Parthiban, R. Modeling and analysis of multi-channel gigabit class CWDM-VLC system. *Optics Communications* 460, 125141 (2020).

Retamal, J. R. D., et al. 4-Gbit/s visible light communication link based on 16-QAM OFDM transmission over remote phosphor-film converted white light by using blue laser diode. *Opt. Express* 23, 33656-33666 (2015).

Shawky, E., Aly, M., El-Shimy, M. Underwater VLC channel estimation based on Kalman filtering for direct current optical- and asymmetrically clipping optical- orthogonal frequency division multiplexing techniques. *Optical and Quantum Electronics* 55(386), 1-16 (2023).

<sup>a</sup>Sachdeva, S., et al. Simulation of an ultrahigh capacity free space optical (FSO) communication system incorporating hybrid WDM-CPDM techniques under disturbed weather. *Journal of optics*, (2023). <https://doi.org/10.1007/s12596-023-01255-1>.

<sup>b</sup>Sachdeva, S., et al. Ultra-High Capacity Optical Satellite Communication System Using PDM-256-QAM and Optical Angular Momentum Beams. *Sensors* 23(786), 1-18 (2023).

<sup>a</sup>Sachdeva, S., Sindhwani, M., Kaur, S., Kumar, A. and Adhikari, M.S. Hybrid OCSS/RF system employing wavelength division multiplexing and modulation transmitter diversity. *Journal of Optical Communications*, (2022). <https://doi.org/10.1515/joc-2022-0314>.

<sup>b</sup>Sachdeva, S., Sindhwani, M., Kaur, S., Kumar, A. and Adhikari, M.S. A novel approach towards the designing of WDM-FSO system incorporating three SOPs and its performance analysis under different geographical regions of India. *Journal of Optical Communications*, (2022). <https://doi.org/10.1515/joc-2022-0315>.

Savory, S. J. Digital filters for coherent optical receivers. *Optics express* 16(2), 804-817, (2008).

Wu, K., He, J., Ma, J., Wei, Y. A BIPCM Scheme Based on OCT Precoding for a 256-QAM OFDM-VLC System. *IEEE Photonics Technology Letters* 30(21), 1866 – 1869, (2018).

Xiang-Peng, C. A Cost-Efficient RGB Laser-Based Visible Light Communication System by Incorporating Hybrid Wavelength and Polarization Division Multiplexing Schemes. *Front. Phys.* 9,731405 (2021). doi: 10.3389/fphy.2021.731405

Yang, P., Xiao, Y., Xiao, M., and Li, S. 6G Wireless Communications: Vision and Potential Techniques. *IEEE Network* 33, 70-75 (2019).

Yeh, C.H., et al. 400 Mbit/s OOK green-LED visible light communication with low illumination. *Opt Quant Electron* 50(430), (2018). <https://doi.org/10.1007/s11082-018-1672-0>

Yoo, Y. H., Jang, J. S., Kwon, H. C., Song, D. W., Jung, S. Y. Demonstration of vehicular visible light communication based on LED headlamp. *International Journal of Automotive Technology* 17, 347–352 (2016).

Zhang, Y., Li, Y. Modulation Schemes of 6PolSK-QPSK in repeaterless transmission system. *Asia Communications and Photonics Conference 2021, Technical Digest Series* (Optica Publishing Group, 2021), Shanghai, China, paper T4A.46, 24–27 October, 2021.

Zou, P., Zhao, Y., Hu, F., Chi, N. Square geometrical shaping 128QAM based time domain hybrid modulation in visible light communication system. *China Communications* 17(1), 163 – 173 (2020).

Zhang, J., Yu, J., Chi, N., Chien, HC. Time-domain digital pre-equalization for band-limited signals based on receiver-side adaptive equalizers. *Opt. Express* 22, 20515-20529 (2014).

Zhang, S. Design and experiment of post-equalization for OOK-NRZ visible light communication system. *Optoelectron. Lett.* 8, 142–145 (2012).



RESEARCH ARTICLE | SEPTEMBER 05 2023

## Classification of toxic comments unified through diverse internet forums

Gull Kaur; Aakash Kumar; Aarjav Chauhan ; Abhishek Babbar



AIP Conf. Proc. 2754, 020013 (2023)

<https://doi.org/10.1063/5.0169608>



Export  
Citation

CrossMark

### Articles You May Be Interested In

Research of techniques used in toxicity detection

*AIP Conference Proceedings* (June 2023)

The United Kingdom Nuclear Science Forum

*AIP Conference Proceedings* (May 2005)

Reframing in neuro-linguistics programming to improve communication skills in scientific forums

*AIP Conference Proceedings* (January 2023)

500 kHz or 8.5 GHz?  
And all the ranges in between.

Lock-in Amplifiers for your periodic signal measurements



Find out more



# Classification of Toxic Comments Unified Through Diverse Internet Forums

Gull Kaur,<sup>1, a)</sup> Aakash Kumar,<sup>2, b)</sup> Aarjav Chauhan,<sup>2, c)</sup> and Abhishek Babbar<sup>2, d)</sup>

<sup>1)</sup> Department of Computer Engineering, Faculty of Computer Engineering, Delhi Technological University, Shahabad Daulatpur Village, Rohini, New Delhi, India.

<sup>2)</sup> Department of Computer Engineering, Delhi Technological University, Shahbad Daulatpur Village, Rohini, New Delhi, India.

<sup>a)</sup> Electronic mail: gullkaur@dtu.ac.in

<sup>b)</sup> Electronic mail: aakashkumar\_2k18co001@dtu.ac.in

<sup>c)</sup> Corresponding author: aarjavchauhan\_2k18co002@dtu.ac.in

<sup>d)</sup> Electronic mail: abhishekbabbar\_2k18co018@dtu.ac.in

**Abstract.** In the last half-decade, India has seen exponential growth in the Internet and social media. This huge growth resulted in better communication among friends and families and freely spread information, content, opinions, and ideas. Some users misuse this freedom and make social media platforms intolerable. The magnitude of detrimental content online, such as toxic comments or content, is not manageable by humans. This study creates a homogeneous dataset by manually labelling comments taken from social platforms and combining them with some publicly available datasets. We have classified them into two category labels, toxic and non-toxic. This work presents our unified dataset, including a wide spectrum of comments and an approach to classify Hinglish comments using the BERT transformer model. The study also includes training baseline models and depicting their performance based on selected evaluation criteria. The BERT model outperformed the baseline and other models trained on the unified dataset. This study gives importance to Hinglish Comments and provides an implementation for classifying them to make internet platform much more secure and friendly for regional language users.

## INTRODUCTION

The amount of toxic content on the internet has been growing rapidly in recent years, and the issue is now getting out of control. Toxic comments in social media have led to violence in the real world and has caused many unforeseen circumstances [1]. The review process is not automated in most social media platforms and goes through a human reviewer who reviews the content. Still, these reviewers cannot keep up with the sheer amount of content generated every hour, affecting their mental stature. This spam content has drastically affected users' browser experience on the social platform and has created an unhealthy environment. Thus it is essential to develop an automatic system to identify this spam content to filter the internet environment.

This study looks at detecting toxic comments or posts written in Hinglish. Prior studies have either worked with the English dataset or used toxic comments only from one social media platform, which don't provide a wider spectrum of data to work with. We found a solution to this problem by creating a unified dataset from major social platforms used in India like Facebook, Twitter, YouTube, Instagram and Reddit. We have curated the data from Reddit, YouTube and Instagram. We also used the comments from existing datasets and labelled them into two categories: toxic and non-toxic. Discretely these datasets are not big enough to create a reliable system. We create one large corpus of toxic comments in Hinglish by combining these datasets into one.

While most existing works have focused on English, we propose a model for the Hinglish language, leading to a more concise and accurate model. The paper is organized in the following manner. The Related Work section surveys the existing Hindi-English datasets and describes the methods used to train a model on this dataset. The methodology section provides the approach and challenges we faced while creating the unified dataset. In the Experimentation section, we describe steps in feature extraction, the architecture of the models and parameters used for training the model. Finally, the result section analyses the model's performance and discusses their drawbacks.

## RELATED WORK

In this section, we briefly discuss the existing toxic comment datasets, toxic comment classification models and different techniques applied to solve the problem of classifying toxic comments.

## Existing Datasets

**Kumar et al.** [2] worked with the dataset of Facebook posts and comments which a human-annotated into three major categories, namely (OAG),(CAG) and (NAG). There is a mix of around 11600 Hindi, English and Roman Comments from Facebook.

**Mathur et al.** [3] worked with the dataset consisting of tweets of toxic nature and Hinglish tweets. There are 3.2K tweets in the dataset. They have been labelled manually into three labels: abusive, hateful, and general tweets.

**Bohra A et al.** [4] worked with the dataset of tweets related to various topics like politics, public protest, riots etc. The tweets were labelled by humans into hate or not hate categories. There are a total of 4.5K tweets in this dataset.

**HASOC2019** [5] The dataset offered here shares a task on identifying abusive content in English and Hindi languages and contains around 13665 posts, with a fair share of both languages. There are three sub-tasks in this task. Whether or not the tweet wording is hateful, offensive, or both. Whether the text of a tweet is hateful, offensive, or obscene. Whether or not a tweet is identity hate towards a specific group or community.

## Existing Work and Methods

**Arkaitz et al.** [6] researched on Hinglish dataset by integrating numerous Convolutional Neural Networks. The BiLSTM layer was used above the CNN layer, capturing sequential and low-level textual representation. With the addition of CNN layers, the simple contextual BiLSTM classification was improved. The CNN-LSTM model showed mediocre performance, although it could be improved.

**Lynn D. et al** [7] worked on bidirectional LSTM structure and used it as the foundation model framework. Iteration on several models by modifying three aspects using foundation model structure was done. To use transfer learning, the initial consideration was taken to train the embedding layer from scratch.

**Sayar et al.** [8] worked on transformer-based masked language models to construct semantic embeddings for cleaned twitter text. The experiments were conducted using XLM-Roberta. XLM-Roberta has outperformed similar multilingual Transformers. As a result, for the shared work, the chosen XLMR model formed the foundation transformer model to fine-tune the weights of the XLMR Transformers.

**Saurabh et al** [9] worked on the evaluation and comparison of several strong baseline models, including methods like LSTM, Attention networks, Pre-trained LSTMs, Hierarchical ConvNet, CNN-multifilter, Bi-LSTM with xgboost. The suggested CapsNet architecture outperformed the other baseline algorithms. The combining focus loss with CapsNet resulted in an increase in the ROC-AUC and F1 score.

## METHODOLOGY

In this section, we represent method used for data collection, annotation and classes of dataset and steps related to pre-processing of the dataset. The proposed approach of implementing the system to filter out toxic comments is shown in Figure 1.

### Data Collection

Creating a unified dataset comes with many challenges. Firstly, the comments need to be fetched from different social platforms. Secondly, The different datasets have different class labels and structures. The Table 1 shows the bifurcation of comments taken from different internet websites to make the unified dataset.

### Data Labelling

The unified dataset consists of two classes. The comments from different social media were labelled and integrated into the dataset. We normalised the heterogeneous data and labelled them into two classes to ensure uniformity in our dataset. The comments taken from these diverse platforms, were labelled manually into two categories. For the existing dataset, we imported all the comments from them and labelled them manually into two classes.

TABLE 1. Dataset Distribution

Social Media	Number
Facebook	9200
Twitter	2100
YouTube	900
Instagram	400
Reddit	400

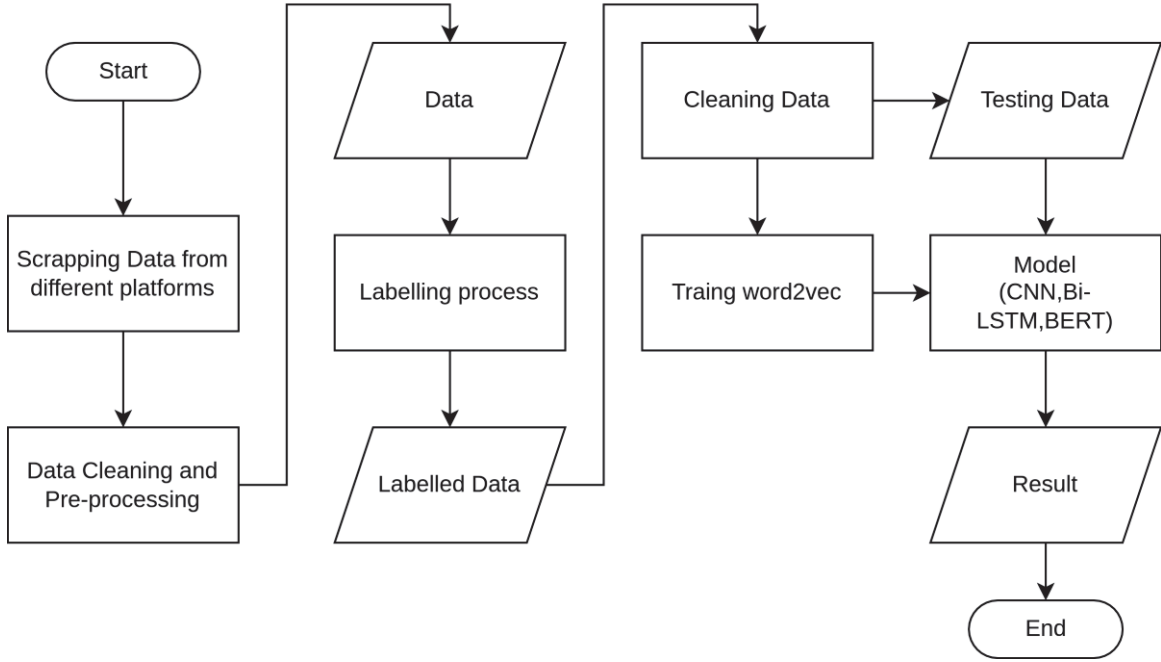


FIGURE 1. Proposed Approach

## Classes

The 1 represents that comment is toxic, whereas the 0 represents that the comment is non-toxic. The two classes which we used were toxic and non-toxic. The comments judged on the basis of the sentiment they created, if any abusive or harmful sentiment is produced by the comment, it was labelled toxic otherwise it was regarded as general (non-toxic) comment.

## Data Pre-processing

The comments extracted from social media platforms contain a lot of noise like links, mentions, hashtags and emojis. All characters were converted into lower case. Number and white spaces were also removed since they don't provide much context to the comment [10]. We also removed the rare word (which occurred less than 20 times). This resulted in data uniformity and reduced irrelevant noise from the dataset.

## Models

We trained five models namely Logistic Regression, SVM, CNN, Bi-LSTM and BERT based transformer model. The embedding of the texts was done using the word2Vec technique which was feeded to the models. The detailed implementation of these models along with their respective architecture is discussed in Experimentation section below.

## EXPERIMENTATION

We trained five different models. The Logistic regression formed the baseline models for classification. We then experimented with SVM, Convolutional Neural Network, Recurrent Neural Network and BERT.

We trained Logistic Regression model as our baseline model [11]. To accommodate large data and converge our model, we trained a logistic regression model for a max iteration of 10000.

After implementing our baseline model, we explored the SVM model for our classification task. We used Radial Basis Kernel Function and gamma as auto for our SVM model.

For textual classification, words in datasets must be represented in a meaningful manner that a computer can understand. The words are given meaning by representing them in a vector. To have quality feature vectors, we used a pre-trained model trained on 250K code-mixed tweets [12]. Using a pre-trained model helped in getting better results. In this part of the experiment, we used word2vec to re-train the model on our dataset to learn word embeddings. We trained the model for ten epochs. We limited the number of features to 200K words for our embedding model.

Taking some inspiration from **Spiros V at el** [13] we decided to experiment with CNN for text classification. The vector obtained from our embedding layer was used to input our CNN model. The architecture of our CNN model consisted of Spatial Dropout, 1D Convolutional layer with 100 filter and kernel size of 4; two dense layers and batch normalization. For input, we considered a maximum sequence length of 512 with a batch size of 150.

To have the best result on our dataset, we decided to work with state of the art model in the NLP domain such as Bi-LSTM and BERT. Bi-LSTM model had architecture similar to our CNN architecture here instead of 1D Convolutional Layer; we used a Bi-LSTM layer. This model's maximum sequence length and batch size are the same as CNN. We have used binary cross-entropy for both models as our loss function. The learning rate for both models is 1e-5.

With the use of a pre-trained embedding model, CNN performed well. However, the time consumed in training the CNN models was comparatively higher. CNN model can be further improved with much better feature extraction approaches and providing a large amount of data to work with. CNN Bi-LSTM provided better results than only CNN based model. The computational cost was similar for both of them.

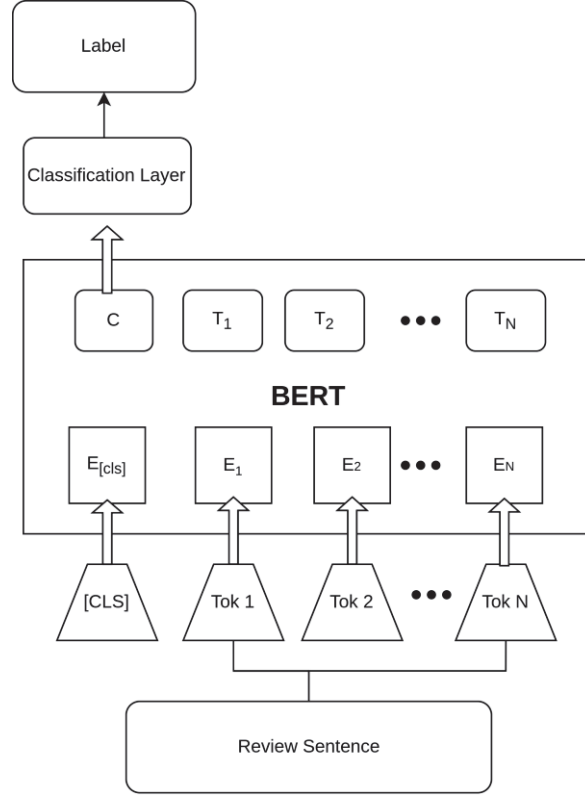
We also used the heavily trained model Hugging Face's BERT (distilbert-base-multilingual-cased) [14] which is trained on a Hinglish dataset. We decided to re-train this model on our dataset with batch size of 150. The proposed DistilBERT architecture is same as the common BERT. **V. Sanh at el.** [14] have majorly focused on reducing the number of layers. The BERT based model made use of transformers and formed a relation between textual words. The BERT model however, took a large amount of time for training on GPU and CPU, but training on TPU resulted in much faster training. We also experimented with keeping the mentions and hashtags from the dataset to provide context to the text, but it performed significantly worse. The visualisation of proposed BERT model is depicted in Figure 2.

## RESULTS

In this section, we discuss the performance of our trained models on the test set and compare them using the evaluation criteria we selected to evaluate the different models. Comparison of the models was made based on evaluation criteria, namely Recall, precision and F1 Score. The comparison draws some important points and helps evaluate the performance of models.

The SVM and Logistic Regression gave similar results on the model. Time to train these models were comparatively quick. Logistic Regression and SVM gave decent accuracy compared to all models, but they gave inadequate Recall and F1 Scores.

The BERT model (distilbert-base-multilingual-cased) [14] performed best with the highest accuracy of 82.5% and F1-Score. It forms a context of the entire comment at once in contrast to direction models which helped in achieving high accuracy. In Table 2, we present the results for all the approaches we have used.



**FIGURE 2.** Proposed Model Architecture

**TABLE 2.** Comparison of Result

Model	Accuracy	Recall	F1-Score
Logistic Regression	0.779	0.779	0.682
SVM	0.778	0.778	0.683
CNN	0.78	0.9679	0.8692
CNN Bi-LSTM	0.805	0.9818	0.8899
BERT(distilbert-base-multilingual-cased) [14]	0.825	0.9594	0.8955

## CONCLUSION

This paper proposes that toxicity and aggression in comments be detected automatically. This research presents an implementation of models on a unified dataset of comments collected from diverse online platforms. We preprocessed the data dealing with emoji characters, hashtags, mentions, and other special characters in this work.

We demonstrate the efficacy of our suggested model by comparing it to benchmark algorithms and demonstrating that it outperforms other baseline models. We depicted that our trained models gave better results than other relevant models in classifying Hinglish comments. The model achieves competitive results on combined data and demonstrates that it can classify toxic comments with much better precision. We also found that pre-trained embeddings resulted in better results. Further work can be done to improve the model architecture and performance - other approaches for feature extraction can be done. The use of unsupervised learning can be done to fine-tune our trained model.

As the comments belonging to toxic categories are diverse, the sentiment caused by toxic comments is different, so the labels among them can be classified further, making models provide more information regarding toxic comments. This study regards this as futuristic improvement in the system of classifying toxic comments.



## REFERENCES

1. L. Hanu, “How ai is learning to identify toxic online content,” (2021).
2. R. Kumar, A. N. Reganti, A. Bhatia, and T. Maheshwari, “Aggression-annotated corpus of hindi-english code-mixed data,” (2018), arXiv:[1803.09402](#) [cs.CL].
3. P. Mathur, R. Sawhney, M. Ayyar, and R. Shah, “Did you offend me? classification of offensive tweets in Hinglish language,” in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)* (Association for Computational Linguistics, Brussels, Belgium, 2018) pp. 138–148.
4. A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava, “A dataset of Hindi-English code-mixed social media text for hate speech detection,” in *Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media* (Association for Computational Linguistics, New Orleans, Louisiana, USA, 2018) pp. 36–41.
5. T. Mandla, S. Modha, G. K. Shahi, A. K. Jaiswal, D. Nandini, D. Patel, P. Majumder, and J. Schäfer, “Overview of the hasoc track at fire 2020: Hate speech and offensive content identification in indo-european languages,” (2021), arXiv:[2108.05927](#) [cs.CL].
6. N. Vashistha and A. Zubiaga, “Online multilingual hate speech detection: Experimenting with hindi and english social media,” *Information* **12** (2021), 10.3390/info12010005.
7. L. D. Kong, “Multilingual toxic comment classification,” (2020).
8. S. G. Roy, U. Narayan, T. Raha, Z. Abid, and V. Varma, “Leveraging multilingual transformers for hate speech detection,” (2021), arXiv:[2101.03207](#) [cs.CL].
9. S. Srivastava, P. Khurana, and V. Tewari, “Identifying aggression and toxicity in comments using capsule network,” in *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)* (Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018) pp. 98–105.
10. T. T. Sasidhar, P. B, and S. K. P, *Procedia Computer Science*, (2020), Vol. **171**, 1346–1352.
11. M. A. Saif, A. N. Medvedev, M. A. Medvedev, and T. Atanasova, “Classification of online toxic comments using the logistic regression and neural networks models,” *AIP Conference Proceedings* **2048**, 060011 (2018), <https://aip.scitation.org/doi/pdf/10.1063/1.5082126>.
12. S. Kamble and A. Joshi, “Hate speech detection from code-mixed hindi-english tweets using deep learning models,” (2018), arXiv:[1811.05145](#) [cs.CL].
13. S. V. Georgakopoulos, S. K. Tasoulis, A. G. Vrahatis, and V. P. Plagianakos, “Convolutional neural networks for toxic comment classification,” (2018), arXiv:[1802.09957](#) [cs.CL].
14. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter,” ArXiv abs/1910.01108 (2019).

# Collective Behavior based Slime Mold Optimization with application to UAV Energy conversion device analysis

Monika Verma (✉ [monikaverma\\_phd2k17@dtu.ac.in](mailto:monikaverma_phd2k17@dtu.ac.in))

Delhi Technological University, Delhi Technological University

Mini Sreejeth

Delhi Technological University, Delhi Technological University

Madhusudan Singh

Delhi Technological University, Delhi Technological University

NA NA

Delhi Technological University, Delhi Technological University

NA NA

---

## Research Article

**Keywords:** Collective behavior disturbances (CB), OR-PMSM-IPIM, slime mold inspired optimization (SMO), traction device, unmanned aerial vehicle (UAV)

**Posted Date:** September 18th, 2023

**DOI:** <https://doi.org/10.21203/rs.3.rs-3348357/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Collective Behavior based Slime Mold Optimization with application to UAV Energy conversion device analysis

Monika Verma\*, Mini Sreejeth, and Madhusudan Singh  
Electrical Engineering Department, Delhi Technological University, Delhi, India  
\*Corresponding Author: monikaverma\_phd2k17@dtu.ac.in

**Abstract**—As an extensively used traction component of unmanned aerial vehicles (UAV), outer-rotor-permanent-magnet-synchronous-motor with inner-periphery-inset-magnet structure (OR-PMSM-IPIM) has eminent advantages with respect to its electromagnetic performance. In order to achieve light weight, high efficiency, efficient speed-torque and low turbulence in UAV, a novel method for optimizing design parameters of OR-PMSM-IPIM is proposed in this paper. A modified slime mold inspired optimization (SMO) algorithm is designed by introducing its collective behavior and swarm intelligence in the environment. In order to avoid collapsing for local optimal spots and enhance diversity in the population, the modified framework of SMO is proposed in this work by introducing complex group interaction patterns in the traditional SMO algorithm. In order to verify the effectiveness and convergence speed of proposed methodology, few comparative experiments are administered. The application of proposed method is tested by developing design-parameter-optimization (DPO) problem for OR-PMSM-IPIM. The analysis results show that the performance of optimized model of OR-PMSM-IPIM has been improved through proposed SMO design optimization.

**Index Terms**—Collective behavior disturbances (CB), OR-PMSM-IPIM, slime mold inspired optimization (SMO), traction device, unmanned aerial vehicle (UAV).

**Statements and Declaration-** This work has been supported by Centre of Excellence for Electric Vehicles and Related Technologies, Electrical Engineering Department, Delhi Technological University, Delhi, India, 110042.

**Competing interest-** The authors have no Competing Interest.

---

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, "This work was supported in part by the U.S. Department of Commerce under Grant BS123456."

The next few paragraphs should contain the authors' current affiliations, including current address and e-mail. For example, F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: author@boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar.colostate.edu).

T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp).

## I. Introduction

The attention of researchers is grabbed by the energy conversion field employing unmanned aerial vehicles (UAV) in which electric motor dominates its energy conversion system. In order to fuel traction power system of UAV such as in electric aircraft, quadcopter and multi-copters etc., the employment of OR-PMSM-IPIM is practiced to enhance its endurance flight time with high efficiency.

Among various structures of PMSM, such as interior type, surface type, multilayer type, interior-exterior type; one of the majorly utilized configuration has been OR-PMSM with inner-periphery-inset-magnet structure. This design is endurance-focused motor configuration, specifically designed for numerous surveying applications [1-2]. It is also used for aerial photography in various fields. This configuration has been chosen as reference configuration for the energy conversion system in order to obtain high performance of the motor [3-4]. As can be seen in Fig. 1, as a part of UAV-Quadcopter, OR-PMSM-IPIM structure is shown in dissembled form as well as cross-sectional view.

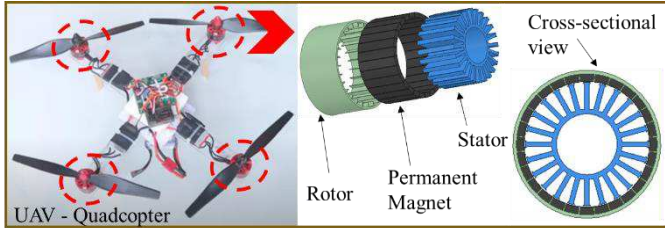


Fig. 1 UAV - Quadcopter Energy Conversion system Design outline

The design optimization helps in investigating the viability of accomplishing highly efficient, fully electric motor traction system for quadcopter UAV which is capable of performing Vertical Take Off and Landing (VTOL) operation [5].

Until now, only small UAV exhibits fully electrical system as its propulsion system, whereas large quadcopter UAV have dependence on traditional engines. The main reasons are low efficiency of electric motors along with the lack of high energy density of batteries. Due to this, the short operational range and short flight time of UAV traction system is obtained. Hence, to achieve smooth VTOL operation of UAV, it is necessary that the electric energy conversion device perform distortion-less operation. Since the variation in design variables of OR-PMSM showcases its capability of delivering smooth and efficient operation. In other words, the search of optimized design parameters of the motor has direct influence over the VTOL operation of UAV [6-7]. Therefore, the design optimization of OR-PMSM-IPIM is carried out in this paper by introducing novel optimization technique.

Table I shows the attempts and their respective research gaps related to design optimization of OR-PMSM.

Table I  
Literature Related to Prior Work of Design Optimization of OR-PMSM

Authors, Years	Application	Pros	Cons
M. Mutluer, 2021 [8]	Direct drive mixer	Cost reduction and efficiency improvement is obtained	Multi-objective optimization increases complexity
E. Mbadiwe et. al., 2021 [9]	Automotive industry	PM flux is increased to enhance torque performance	No definite method of optimization is used
C. Guerroudj et. al., 2021 [10]	Wind turbine	Doubly salient PM generator is optimized	Large cogging torque
H. Taha et. al., 2019 [11]	Electric bicycle	Inner and outer configurations are compared	Method of optimization is absent
Z. Shi et. al., 2020 [12]	Low speed campus patrol Electric Vehicle (EV)	Efficiency and torque density is improved	Limited working area of EV
K. T. Chau et. al., 2007 [13]	In-wheel EV	Torque performance is improved	Speed is reduced, ripples are more
M. Ahmad et. al., 2015 [14]	Direct drive EV	OR-PMSM gives better performance than inner rotor PMSM	Low speed operation is considered
I. Boldea et. al., [15]	Home applications use	FEA analysis is embedded within the optimization process	High variation of efficiency at different ratings of load due to inaccurate modeling of material properties
J. M. Ahn et. al., [16]	UAV high endurance	Geometry of PMSM is optimized with reduced cogging	Introduces more stochastic behavior in conventional PSO

Based on the “survival of the fittest” concept, the bio-inspired optimization methodologies having swarm intelligent algorithms are being progressively used in diverse fields. Out of the streak of such tools like modified artificial bee colony [17], particle swarm optimization [18], genetic algorithm [19], wind driven optimization [20], brain storm optimization [21], the bio-inspired optimization techniques can be governed by environmental as well as evolutionary mediums. The discontinuous and non-linear type of optimization problems can be efficiently solved using characteristic interactive behavior of such tools. The bio-inspired algorithms are strongly robust since the overall optimization problem is not influenced by variation in individual characteristic behavior. *The implementation of modified version of slime mold algorithm, in this work, aims to recognize how singular-level behaviors may prevail upon complicated faction-level patterns.*

Primarily, the study of collective behavior (CB) has been carried out in animal organizations such as bird flocking, fish schooling and insect colonies. As per the studies, the disturbance due to CB is also encountered in micro-organisms. *This work argues about slime molds for being dominant model schemes to deal with several outstanding problems in CB patterns.* Particularly, the clue of behaviors of slime mold leading to the linkage of singular-level structures to group-level mechanisms has been followed.

Prior to this article, many researchers have proposed modifications in the SMO algorithm. Table II exhibits the till-today versions of modified SMO algorithm. From the discussed literature, it is quite noticeable that the inclusion of CB and utilization outline are not addressed in any of them. Therefore, a novel CB based SMO technique called “CB-SMO” is developed in this paper. This proposed method is tested by solving various Benchmark functions. Thereafter, CB-SMO is utilized for its implementation in the application to UAV Energy conversion device analysis. At last, the performance of model, having optimized design parameters, is compared with that of the model without optimized parameters.

Organization of the paper is as given: Section I shows the introductory literature survey. Section II shows the technical details of proposed CB-SMO technique. Section III presents the design of DPO. Section IV contains all the results and the related discussion. It also presents the achievements obtained by implementing the novel CB-SMO method to the developed DPO. Section V shows the concluding remarks and future scope of this work.

### PROPOSED COLLECTIVE BEHAVIOR BASED SMO

The slime molds, of class Myxogastria, are multinucleate unicells, can grow up to 900 squared centimeters during plasmodium stage. The morphological life cycle structure of Physarum Polycephalum consisting of 1: germination, 2: flagellation, 3: amoeboid, 4: fusion of haploid, 5: micro-plasmodium, 6: mature plasmodium, 7: sporangium, 8: dispersal is shown in Fig. 2-(a). The food engulfing stage of slime mold during search of source is presented in Fig. 2-(b).

*The CB possesses conceptual origin in the investigation of coherent physical systems that involves the essential exchange of individual particles.* The relatively state-of-the-art understanding of cell behavior, driven by intercellular operations in slime molds, make them, conceivably, powerful model systems for implementation of CB [34].

The upcoming subsections describe the observed CB behaviors of Physarum Polycephalum (Slim mold), that are considered to develop novel CB-SMO technique. The corresponding display-chart is shown in Fig. 3.

Table II  
Literature Showing Historical Background of modified SMO

Authors, Year	Strategy
A.D. Tang et. al. [22], 2021	Introduced chaotic-opposition strategy, spiral search technique and adaptive parameter control schemes in the traditional SMO for global optimization [19]
M. K. Naik et. al. [23], 2021	Infused adaptive opposition based learning concept
Y. Xiao et. al. [24], 2021	Tent Chaotic map functions and inertial weights are inculcated
A. Hamed et. al. [25], 2021	The application of traditional SMO to lay best size and location of DSTATCOM and PV to improve its voltage and current profile
J. Jones et. al. [26], 2015	The multi-agent based approach has been introduced to comment on the performance of SMO in material computation
Y. Liu et. al. [27], 2021	Integration of chaotic maps and Nelder-Mead simplex technique and implemented it for photovoltaic (PV) parameters selection
H. Jia et. al. [28], 2021	Merged mutation and restart schemes with SMO for increasing accuracy and aid the feature selection
M. A. Basset et. al. [29], 2021	A hybrid version of whale optimization technique with SMO was implemented to solve the problem of image segmentation in paramedics
Z. Cui et. al. [30], 2021	The search space was improved using levy flight distribution with SMO
H. Gao. Et. al. [31], 2021	Combined with support vector machine technique for predicting stability in postgraduate employment
M. K. Naik et. al. [32], 2021	The concept of three best leaders inspired by Grey Wolf Optimizer has been introduced
L. Liu et. al. [33], 2021	The hybridization of differential evolution along with SMO has been proposed for performance improvement in multilevel image segmentation field

### A. Synchronization

The first behavior is *Synchronization*. Plasmodium slime molds are collective of oscillators. The food discovery and interaction between neighboring oscillators govern the frequency of oscillations.

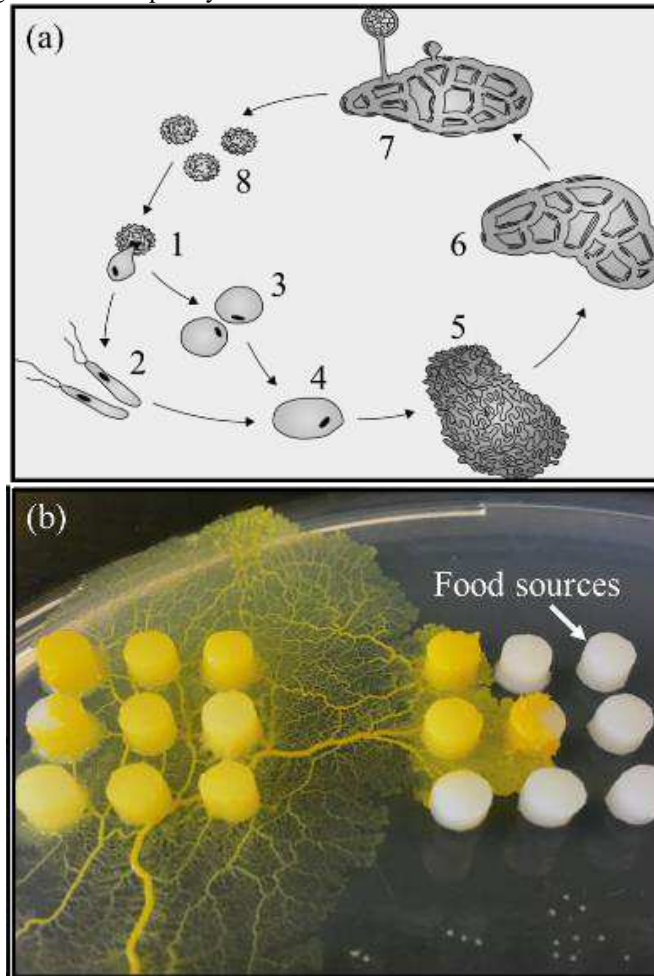


Fig. 2 (a) Life cycle of Physarum Polycephalum; (b) Food engulfing by slime mold as problem solving segment

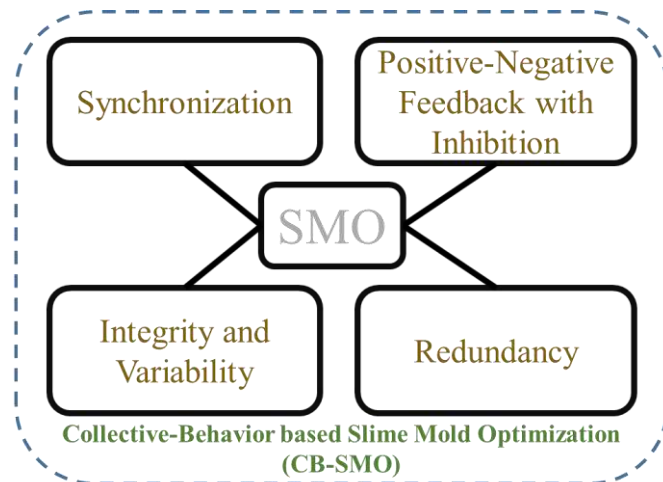


Fig. 3 The display-chart of the proposed CB-SMO technique, consisting of four CBs combined with traditional SMO

The *synchronization* behavior is carried out by slime through cytoplasmic flow. This CB follows simple steps:

- Food Source is recognized (by surface receptors of cell).
- Frequency of the oscillators (nearest to the food attractant) increases.
- The cytoplasmic flow leads to amoeboid migration in the direction of food attractant.
- Neighboring oscillators gets informed (due to existence of physical coupling) about the qualitative aspect of local environment.

➤ This encoded information is dispatched to the secluded cell parts, causing *Synchronization*.

Thus, the new solution  $x^{syn}(i+1)$ , at  $i^{th}$  index, is updated using the randomly selected neighbor,  $x^{ran}(i)$  (see Equation (1)). This process is “recruitment”. Then comes the “communication”.

The information is shared with the coupled oscillators which select and update its position based on the probability data. This probability data is computed from the fitness of the previous solution,  $fitness(x^{syn}(i))$ . The probability,  $p_i^{syn}$ , is computed using Equation (2). During “communication”, if  $ran < p_i^{syn}$ , the solution is updated using Equation (1) again. The solutions which are not improved by *synchronization* behavior or they exceed pre-determined *limit* (called abandoned oscillators), the solution is updated using Equation (3). They are thus determined to search new food source randomly.

$$x^{syn}(i+1) = x^{syn}(i) + r^{ran}(x^{syn}(i) - x^{ran}(i)) \quad (1)$$

$$p_i^{syn} = \frac{fitness(x^{syn}(i))}{\sum_{i=1}^N fitness(x^{syn}(i))} \quad (2)$$

$$x^{syn}(i+1) = lb + ran(ub - lb) \quad (3)$$

Where  $r^{ran}$  is random number from  $-1$  to  $1$ , to ensure randomness in the recruitment of neighbor. Here,  $ran \in [0,1]$  is random number,  $lb$  is the lower bound of solution and  $ub$  is upper bound of solution.

### B. Positive-Negative feedback with inhibition

The second behavior is *Positive-Negative feedback with Inhibition*. The traditional SMO makes use of weights in order to simulate positive-negative feedback produced during foraging. The weights of Slime Mold are computed using Equation (4), [where,  $bestF$  = best fitness,  $worstF$  = worst fitness during current iteration,  $so(i)$  = first half of sorted population,  $sx(i)$  = index of sorted synchronized fitness of population,  $N$  = population size].

So, due to positive-negative feedback, the updated position of oscillators,  $x^{pn}(i+1)$ , is computed using Equation (5). [where,  $z$  = constant to control exploration and exploitation,  $bestS(i)$  = The so far best solution,  $x^a(i)$  and  $x^b(i)$  = two randomly selected oscillators,  $x^{pn}(i)$  = old position of oscillator]. Factors  $v_b$  and  $v_c$  = vectors consisting of random numbers, are specified by Equation (6) and Equation (7) respectively. [where,  $t$  = current iteration,  $T$  = total iteration,  $p$  = control parameter,  $gbestF$  = global optimal fitness value (see Equation (8))].

However, in Physarum, Poly-L-malate like polyanions can have interaction with polymerase and other histones of Physarum resulting in the “inhibition” of activities of nuclei of Physarum plasmodia. It is observed that the severity of inhibition depends upon the distance between the number of cleft atoms [35]. To model this CB, three randomly selected fragments ( $x^i(i)$ , where *superscript i* indicates the  $i^{th}$  fragment) are chosen to form *inhibiting* solution vector.

Depending upon the inhibiting probability factor,  $p_i$ , the position of inhibited oscillator is computed using Equation (9). The inhibiting solution vector,  $x^{inh}(i+1)$ , is calculated using Equation (10), where  $f^i$  indicates the inhibition factor, set by the designer.

$$W^{pn}(sx(i)) = \begin{cases} 1 + ran \log\left(\frac{bestF - so(i)}{bestF - worstF} + 1\right); i \leq N/2 \\ 1 - ran \log\left(\frac{bestF - so(i)}{bestF - worstF} + 1\right); i > N/2 \end{cases} \quad (4)$$

$$x^{pn}(i+1) = \begin{cases} lb + ran(ub - lb) & ; r < z \\ bestS(i) + v_b(W^{pn}.x^a(i) - x^b(i)); r < p \\ v_c.x^{pn}(i) & ; r > p \end{cases} \quad (5)$$

$$v_b \in [-\tanh^{-1}\left(-\frac{t}{T} + 1\right), -\tanh^{-1}\left(-\frac{t}{T} + 1\right)] \quad (6)$$

$$v_c \in \left[-\left(1 - \frac{t}{T}\right), \left(1 - \frac{t}{T}\right)\right] \quad (7)$$

$$p = \tanh|fitness(x^{pn}(i)) - gbestF| \quad (8)$$

$$x^{inh}(i+1) = \begin{cases} I^i(i) & ; ran \leq p_i \\ x^{pn}(i) & ; otherwise \end{cases} \quad (9)$$

$$I^i(i) = x^1(i) + f^i(x^2(i) - x^3(i)) \quad (10)$$



### C. Integrity and variability

The third important collective behavior added here is *Integrity and Variability*. While oscillating, the venous structure of SM retains the “*integrity*” of keeping the track of its previous optimum food location. However, the same sized fragments (obtained from common plasmodium) showcase preferences of food search differently during propagation, due to intra-plasmodium “*variability*”.

Thus, the oscillators’ journey proceeds towards the global optimum independently, on the basis of reporting done by the other fragments. The mathematical modeling of this behavior is expressed in Equation (11). [For *integrity*, the inertial weight (of fragment),  $W^{iv}$ , and its speed,  $V^{iv}$ , are introduced to reach global optimum,  $bestS(i)$ ;  $a_p$  and  $a_g$  are acceleration coefficients;  $popt$  = individual optimum location (representing *variability* in CB of SM)]. The updated and previous position of oscillators due to *integrity and variability* are expressed as  $x^{iv}(i+1)$  and  $x^{iv}(i)$  respectively.

$$\begin{aligned} x^{iv}(i+1) &= x^{iv}(i) + W^{iv}V^{iv} \\ &\quad + a_p \text{ran}(popt(i) - x^{iv}(i)) \\ &\quad + a_g \text{ran}(bestS(i) - x^{iv}(i)) \end{aligned} \quad (11)$$

### D. Redundancy

The epitome of *redundant* behavior is also presented by Physarum plasmodia, due to the presence of the syncytial property. The *redundancy* CB undergoes two phases; “serving at prime cell level” and “sharing information at individual level independently”. The position of fragments is updated using Equation (12) for first phase and using Equation (13) for second phase respectively.

[where,  $x^{r1}(i+1)$  = updated position of oscillator for serving phase;  $x^{r2}(i+1)$  = updated position of oscillator for sharing phase;  $x^r(i)$  = previous location;  $sf$  = serving factor (randomly chosen for a particular fragment undergoing *redundancy*);  $meanS(i)$  = mean of current solution vector;  $x^p(i)$  = randomly selected partner attractant].

Once the “serving phase” updates the position, the “sharing phase” ensures the interconnection in the foraging network, ensuring the linkage to each food source. However, depending upon the fitness of the partner, “sharing phase” updates the position using + or – in Equation (13).

$$\begin{aligned} x^{iv}(i+1) &= x^{iv}(i) + W^{iv}V^{iv} \\ &\quad + a_p \text{ran}(popt(i) - x^{iv}(i)) \\ &\quad + a_g \text{ran}(bestS(i) - x^{iv}(i)) \end{aligned} \quad (11)$$

$$\begin{aligned} x^{r1}(i+1) &= x^r(i) \\ &\quad + \text{ran}(bestS(i) - sf(meanS(i))) \end{aligned} \quad (12)$$

$$\begin{aligned} x^{r2}(i+1) &= x^r(i) \\ &\quad \pm \text{ran}(x^{r1}(i) - x^p(i)) \end{aligned}$$

### E. Complexity analysis

The analysis of computation complexity of SMO and CB-SMO is performed using the mathematical notations. Let the input cells of Physarum be expressed as  $N$  for each instance,  $D$  be the size of dimension and  $T$  be the upper bound of iterations or total number of iterations for the fitness model.

In traditional SMO, the initialization complexity is  $O(D)$ , fitness and sorting complexity is  $O(N + N \log N)$ , weight update complexity is  $O(ND)$  and location update complexity is  $O(ND)$ . Thus, the overall computational complexity of conventional SMO is  $O(D + TN(1 + \log N + 2D))$ .

In CB based SMO with “*synchronization*” behavior, the location update complexity is  $O(N)$  using Equation (1), conditional (Equation (2)) location update complexity is  $O(N^2)$ , sorting complexity is  $O(D)$  and abandoned oscillators’ location update (Equation (3)) has the complexity of  $O(N \log N)$ . Synchronization behavior based SMO has complexity of  $O(T(N + N^2 + D + N \log N)) \approx O(N^2)$ .

In CB based SMO with “*positive-negative feedback including inhibition*” behavior, the location update complexity (Equation (9) - (10)) is  $O(TN(1 + D))$ .

CB based SMO with “*integrity and variability*” behavior has location update complexity of  $O(TN)$  (Equation (11)).

With “*redundancy*” behavior in CB based SMO, the location of each oscillator is updated using Equation (12) and Equation (13) with complexity of  $O(TN)$ .

## DEVELOPMENT OF DESIGN OPTIMIZATION PROBLEM

There are few assumptions considered for modeling of OR-PMSM in this paper; (1) The three phase star connection is identical and symmetrical, (2) The magnetic material used in rotor magnets have uniform properties, and (3) The stator winding has

coincident inductance and resistance in winding.

#### F. UAV-Quadcopter dynamics

The UAV - Quadcopter Energy Conversion system follows four rotating propellers. They have clockwise (exhibited by motor 1 and motor 3) as well as anti-clockwise (motor 2 and motor 4) rotating directions (Fig. 4).

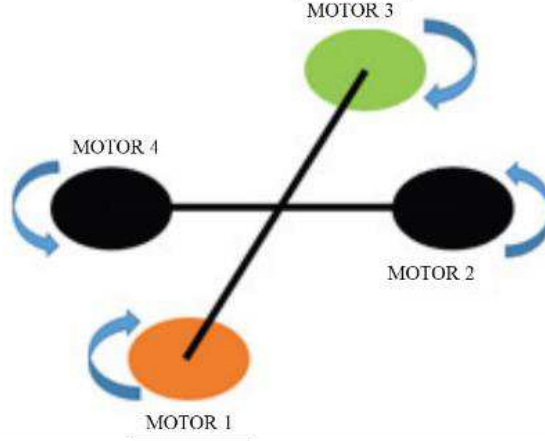


Fig. 4 Representation of rotational direction of UAV motors

The three-degree-of-freedom movement of quadcopter occurs in three axes;  $x$ ,  $y$  and  $z$ . It forms three Euler angles, i.e., roll, pitch and yaw angle. The roll, pitch and yaw angular acceleration are related to the moment of inertia on the corresponding axis and the lifting forces.

The moment of inertia depends upon center of mass ( $M$ ) of UAV, distance ( $R$ ) of middle tip from center of mass, height ( $H$ ) of quadrotor's central part, radius ( $r$ ) of OR-PMSM, mass ( $m$ ) of OR-PMSM and height ( $h$ ) of OR-PMSM etc. The relationships of three angular accelerations of pitch, roll and yaw ( $\ddot{\phi}$ ,  $\ddot{\theta}$  and  $\ddot{\psi}$ ) are presented as follows (Equation (13)):

$$\begin{aligned}\ddot{\phi} &= \frac{l}{\left(\frac{mr^2}{4} + \frac{mh^2}{6} + 2mr^2 + \frac{MR^2}{4} + \frac{MH^2}{12}\right)} b(\omega_4^2 - \omega_2^2) \\ \ddot{\theta} &= \frac{l}{\left(\frac{mr^2}{4} + \frac{mh^2}{6} + 2mr^2 + \frac{MR^2}{4} + \frac{MH^2}{12}\right)} b(\omega_3^2 - \omega_1^2) \\ \ddot{\psi} &= \frac{l}{\left(\frac{MR^2}{2} + 4mr^2\right)} d(\omega_2^2 + \omega_4^2 - \omega_1^2 - \omega_3^2)\end{aligned}\quad (13)$$

Where  $l$  = length of quadcopter arm;  $b$  = thrust factor;  $d$  = drag factor;  $\omega_1, \omega_2, \omega_3, \omega_4$  = angular velocities of motor 1, 2, 3, 4. By keeping constant values of  $M, R, b$  and angular velocities, angular accelerations are re-written as follows (Equation (14)); where  $k_1$  and  $k_3$  are constants).

$$\begin{aligned}\ddot{\phi} = \ddot{\theta} &= f\left(\frac{1}{\frac{mr^2}{4} + \frac{mh^2}{6} + 2mr^2 + k_1} \cdot k_2(\phi/\theta)\right) \\ \ddot{\psi} &= f\left(\frac{1}{4mr^2 + k_3} \cdot k_2(\psi)\right)\end{aligned}\quad (14)$$

#### G. DOP Development procedure - Polynomial regression based surrogate modeling

There appear numerous rotor positions in a single revolution of motor where the rise in cogging torque is observed. For smoother VTOL operation, cogging situation is needed to be suppressed. To save the unnecessary time wasted in re-designing of model, the surrogate model of cogging torque is designed using Polynomial Regression to develop the desired DOP [36].

The predicted or target output (here it is cogging torque),  $P_o$ , with respect to design parameters,  $d_\alpha$  or  $d_l$  of motor based on Response Surface Modeling (RSM) [37] is related as follows (15):

$$P_o = \beta_0 + \sum_{\alpha=1}^3 \beta_\alpha d_\alpha + \sum_{\alpha=1}^3 \sum_{l=\alpha+1}^3 \beta_{\alpha l} d_\alpha d_l + \sum_{\alpha=1}^3 \beta_{\alpha\alpha} d_\alpha^2 + \epsilon \quad (15)$$

Where  $\beta_0$  = offset coefficient;  $\beta_\alpha$  = individual coefficient of  $\alpha^{th}$  variable;  $\beta_{\alpha l}$  = associative coefficient of  $\alpha^{th}$  and  $l^{th}$  variable;  $\beta_{\alpha\alpha}$  = non-linear coefficient;  $\epsilon$  = error between predicted and actual output.

Since, magnet thickness, rotor yoke thickness and motor height are factors which majorly contribute towards the production of unwanted cogging torque. Also, these parameters have influence over the parameters  $m, h$  and  $r$  of OR-PMSM. The angular accelerations of pitch, roll and yaw ( $\ddot{\phi}$ ,  $\ddot{\theta}$  and  $\ddot{\psi}$ ) are related to these design parameters, as can be seen from Equation (14).

For smooth and efficient VTOL operation, the values of angular accelerations of pitch, roll and yaw ( $\ddot{\phi}$ ,  $\ddot{\theta}$  and  $\ddot{\psi}$ ) should be optimum. Hence, these three are chosen as design variables for designing the DOP, (Fig. 5).

The samples obtained by FEM analysis with variation in  $\alpha^{th}$  design variable are listed in Appendix A (Table A.1). The  $\beta$

coefficients of surrogate models are computed using Least Square Method (LSM). The surrogate model obtained for cogging torque is given by Equation (16):

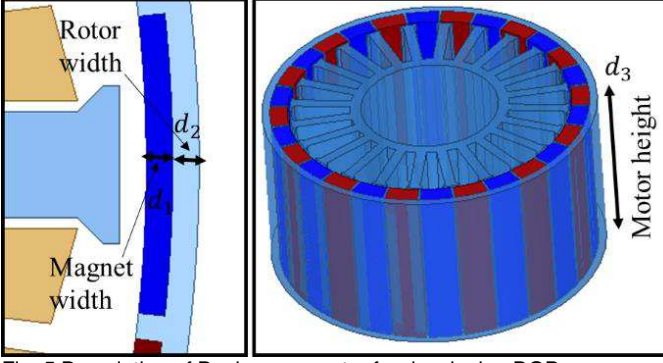


Fig. 5 Description of Design parameter for developing DOP

$$\begin{aligned}
 T_c = & 447.166 - 151.1875d_1 + 16.1289d_2 \\
 & -13.1771d_3 + 4.36d_1d_2 - 0.3875d_1d_3 \\
 & -0.1433d_2d_3 + 24d_1^2 - 5.6556d_2^2 \\
 & + 0.2625d_3^2
 \end{aligned} \quad (16)$$

The root mean square error (RMSE) value obtained for the formulated DOP equation to obtain minimum cogging torque  $T_c$  is 0.1322. This RMSE value shows that the formulated surrogate models are accurate in nature and does not contain any undesired input variable [38]. This surrogate model has been assigned as objective function to the proposed CB-SMO for obtaining the optimal parameters of OR-PMSM for least cogging torque.

## RESULTS AND DISCUSSIONS

This section presents the performance analysis of the proposed modified SMO technique, “CB-SMO” by incorporating CB disturbances one by one. Then, the proposed CB-SMO is utilized in order to solve the formulated DOP.

### H. Performance Analysis of CB-SMO

The proposed CB-SMO is pictorially represented in Fig. 6. Here, CB1 suggests the addition of “Synchronization” behavior to SMO; CB2 suggests the addition of “Positive-Negative Feedback with Inhibition” behavior to SMO, CB3 suggests the addition of “Integrity and Invariability” behavior to SMO and CB4 suggests the addition of “Redundancy” behavior to SMO. The comparative study of performance shown by different algorithms is shown in Table III for various Benchmark Functions.

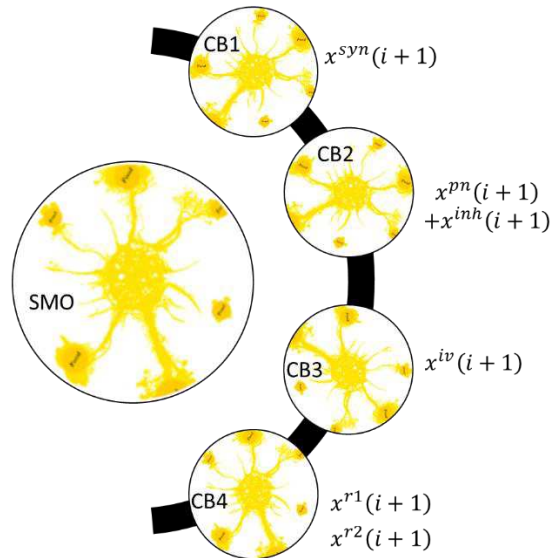


Fig. 6 Descriptive view of SMO along with proposed CBs

From this study, it is observed that:

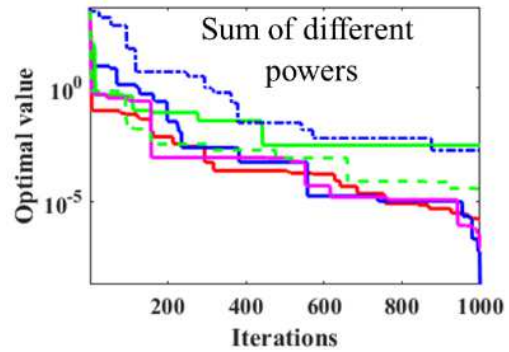
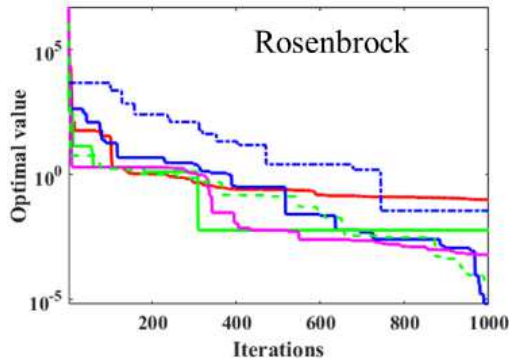
- Adding “only CB1 to SMO” will mostly have negative effect over non-linear problems involving trigonometric functions.
- Addition of “only CB2 to SMO” has increased the efficiency of SMO.

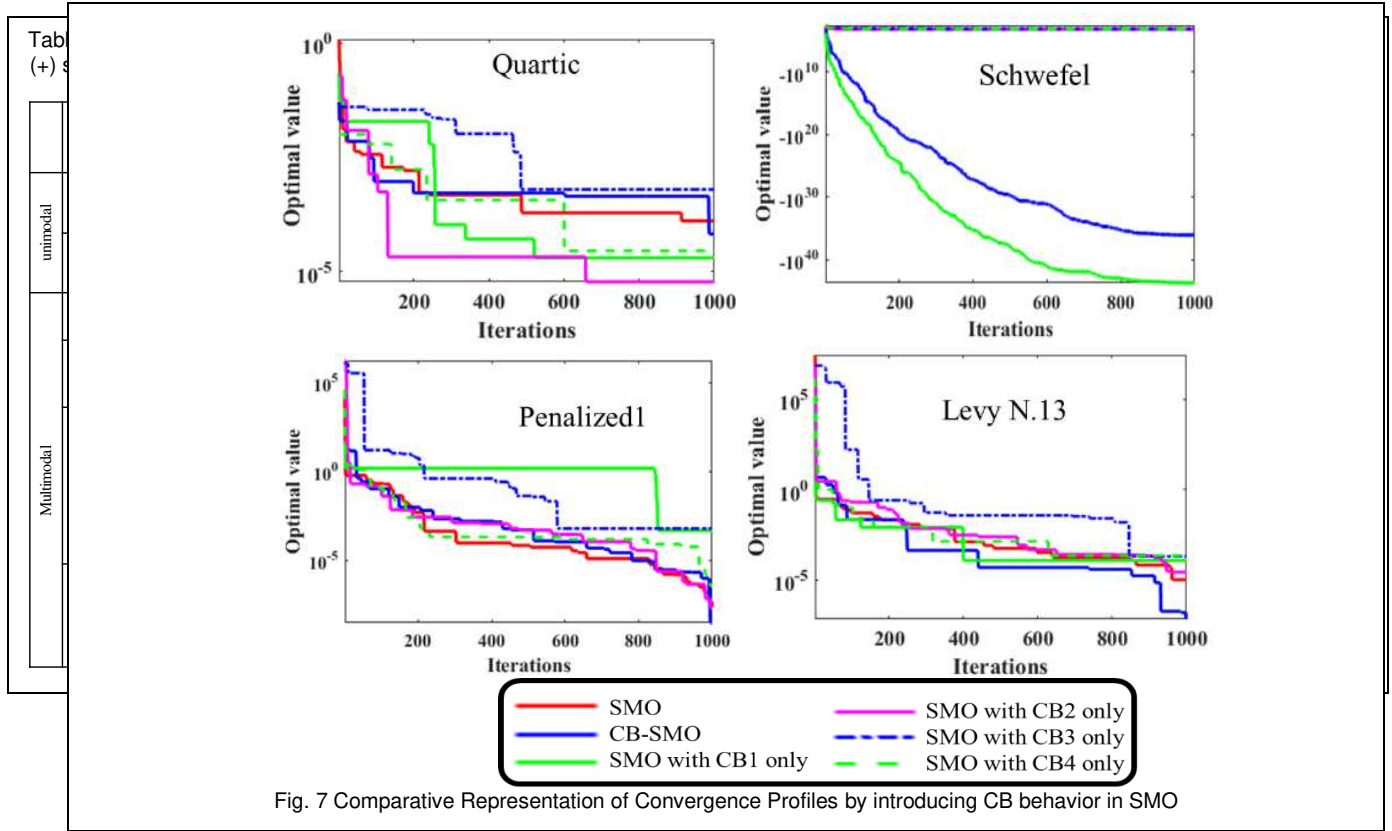
- Addition of CB3 has negative effect while solving problems containing trigonometric functions (exhibits the similar effect as that of adding “CB1 to SMO”).
- By adding CB4, the SMO produces better results (~98% lesser than SMO result) or similar results for Rosenbrock, Sum of different powers, Quartic and Schwefel functions. While for Penalized1 and Levy N.13 functions, SMO produces better result than “SMO with CB4 only”.
- Overall effect of “adding all CBs to the SMO” always shows the positive effect over the performance of SMO.

Table IV represents analysis results for testing the proposed CB-SMO with other traditional and popular MH techniques. The results report that CB-SMO produces precise optimal values for all the Benchmark functions. The convergence profiles, exhibited by different “CB added to SMO” are presented in Fig. 7.

Table IV. Definitions of Benchmark Functions and comparative analysis results with other classical MH techniques; Aquila Grasshopper Optimization (AGO), Aquila Optimizer (AO), Grasshopper Optimization Algorithm (GOA), Artificial Bee Colony (ABC) method, Particle Swarm Optimization (PSO) method, Teaching Learning Based Optimization (TLBO) method and Differential Evolution (DE)

	Names	Description	Range	$f_{min}$	CB-SMO	AGO	AO	GOA	ABC	DE	PSO	TLBO
unimodal	Rosenbrock	$f(x) = \sum_{i=1}^{n-1} [100(x_i^2 - x_{i+1})^2 + (1 - x_i)^2]$	[-30,30]	0	3.2343e-04	8.222e-04	8.306e-04	1.011e+00	0.401e+00	0.000e+00	0.000e+00	0.000e+00
	Sum of different powers	$f(x) = \sum_{i=1}^n ( x_i + 0.5 )^2$	[-100,100]	0	7.7404e-09	2.152e-05	4.556e-06	1.012e-07	0.500e+00	0.000e+00	4.264e-08	1.080e-08
Multimodal	Quartic	$f(x) = \sum_{i=0}^n ix_i^4 + random[0,1]$	[-128,128]	0	4.5931e-05	5.318e-03	2.661e-03	5.812e-03	0.098e+00	3.292e-05	0.004e+00	0.011e+00
	Schwefel	$f(x) = \sum_{i=1}^n (-x_i \sin(\sqrt{ x_i }))$	[-500,500]	-418.982n	-1291.5806	-1213.586	-1212.784	-1217.336	-1202.968	-1208.229	-1200.868	-1362.886
	Penalized1	$f(x) = \frac{\pi}{n} \{10 \sin(\pi y_1) + \sum_{i=1}^{n-1} (y_i - 1)^2 [1 + 10 \sin^2(\pi y_{i+1}) + \sum_{j=1}^n u(x_j, 10, 100, 4)]\}$ ; where $y_i = 1 + \frac{K(x_i - a)^m}{4}$ , $u(x_i, a, k, m) = \begin{cases} 0 & \text{for } -a \leq x_i \leq a \\ K(-x_i - a)^m & \text{for } x_i < -a \end{cases}$	[-50,50]	0	1.4611e-10	1.498e-06	1.505e-05	5.519e-06	1.000e+00	4.711e-03	4.326e-04	3.110e+00
	Levy N.13	$f(x) = 0.1(\sin^2(3\pi x_1) + \sum_{i=1}^n (x_i - 1)^2 [1 + \sin^2(3\pi x_i + 1)] + (x_n - 1)^2 [1 + \sin^2(2\pi x_n)] + \sum_{i=1}^n u(x_i, 5, 100, 4)$	[-50,50]	0	6.6909e-08	1.537e-07	1.353e-06	1.327e-06	1.000e+00	1.349e-02	6.816e-03	0.019e+00





It reveals that although adding individual CB behaviors to SMO, can produce sluggishness in the convergence of SMO. However, the proposed CB-SMO, having incorporated all the CB behaviors, can search the space area thoroughly while going through the computational iterations and can converge to the better optimal value than that produced by SMO.

#### I. Solution to DOP via proposed CB-SMO

In this sub-section, the proposed CB-SMO and other techniques solves DOP to obtain least cogging torque (Eq. (16)), the optimal design parameters are analyzed and compared. The comparative analysis results, obtained by proposed technique and few conventional MH techniques, are presented in Table V. The 'Rank' represents the effectiveness of their performances for getting the solution of the designed DOP.

Table V  
Comparative analysis of Optimal Models obtained by proposed CB-SMO and other classical MH techniques

Algorithm	OR-PMSM-IPIM		Rank
	Parameter values ( $d_1, d_2, d_3$ )	$T_c$ ( $mNm$ )	
CB-SMO	(3.3,1.5,31.9)	41	1
AGO	(3.6,2,31)	42.107	4
AO	(3.7,1.84,31.34)	42.523	6
GOA	(3.4,1.9,31)	41.959	2
ABC	(3.4,1.8,31)	42.297	5
PSO	(3.2,1.2,30.8)	42.637	7
TLBO	(3.2,1.5,32)	42.051	3
DE	(3.2,1.5,32)	42.836	8

The successful implementation of proposed CB-SMO secures rank 1 in producing the least value of cogging torque  $T_c$ , of

41  $mNm$ . While GOA and TLBO secures ranks  $< 4$  and AGO, ABC, AO, PSO, DE secures ranks from 4 to 8 respectively. Table V compares the Initial and Optimal model and their respective torques with respect to the initial and optimized design parameters. Also the dimensions of initial and final models are presented in 2-D and 3-D FEM models of Quadcopters in Fig. 8.

TABLE V  
COMPARATIVE ANALYSIS OF INITIAL AND OPTIMAL MODELS OBTAINED BY PROPOSED CB-SMO

Model	OR-PMSM-IPIM		Remark
	Parameter values (mm) ( $d_1, d_2, d_3$ )	$T_c$ ( $mNm$ )	
INITIAL model	(3,2,31)	43	-
OPTIMAL model	(3.3,1.5,31.9)	41	4.6 % reduction

Further, for verification of influence over the performance of UAV, the “*optimal*” model of OR-PMSM-IPIM is thus designed for Quadcopter application, by assigning the rotational directions as discussed in Fig. 4 previously. Thus, comparison of different essential aspects of Initial and optimal model are shown in the form of Bar graph in Fig. 9.



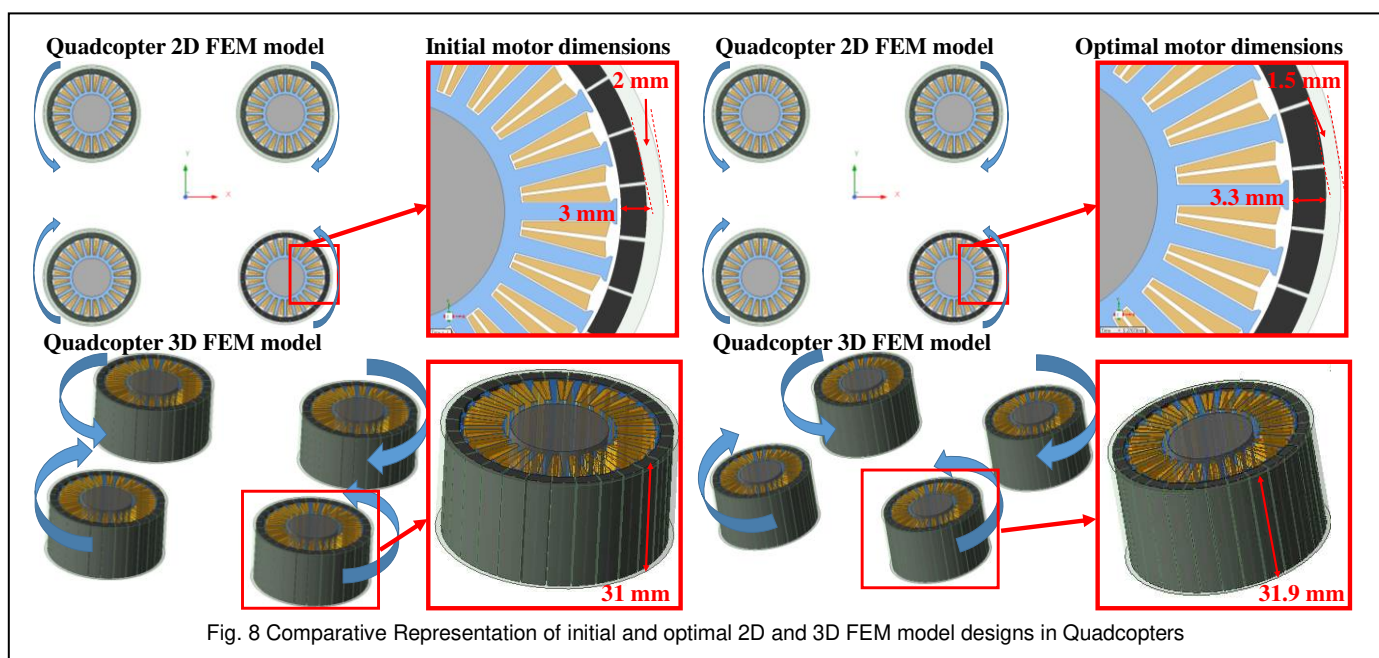


Fig. 8 Comparative Representation of initial and optimal 2D and 3D FEM model designs in Quadcopters

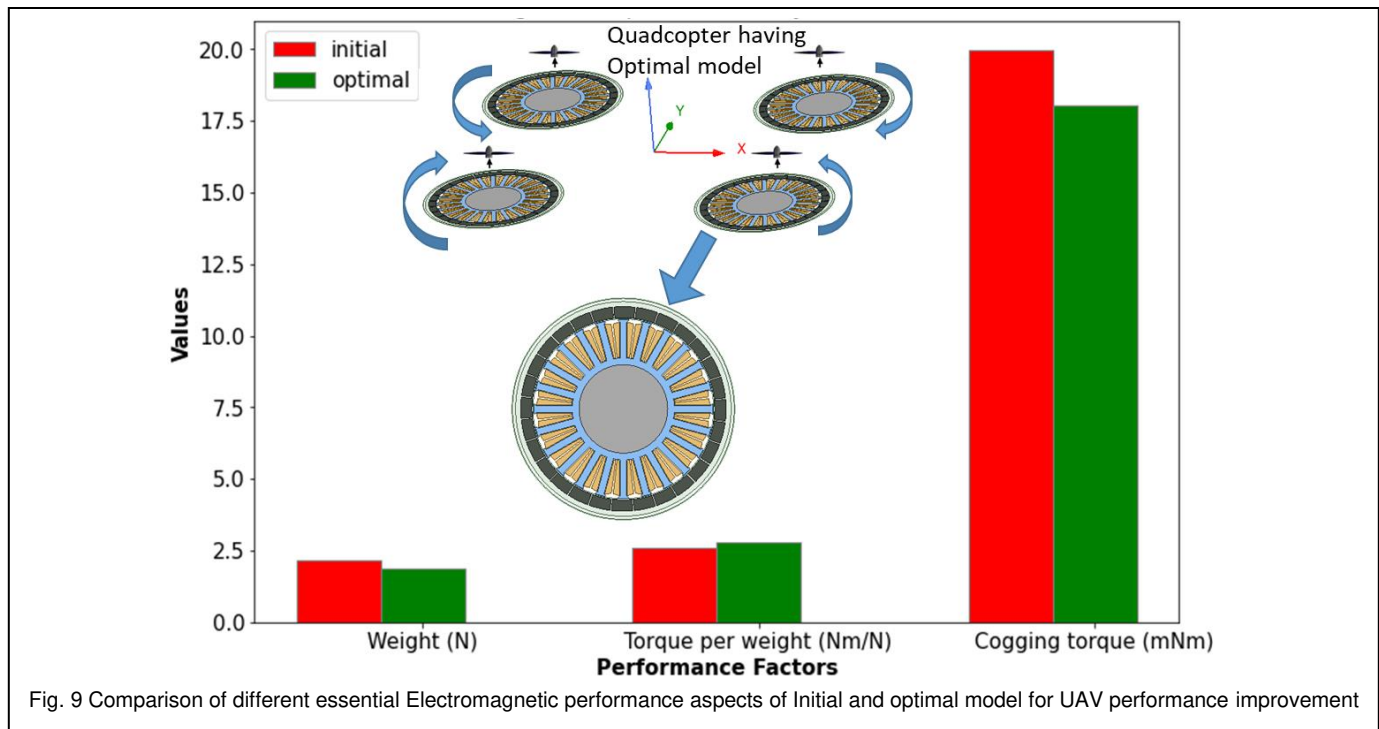


Fig. 9 Comparison of different essential Electromagnetic performance aspects of Initial and optimal model for UAV performance improvement

Initially, the ORPMSM has weight of 2.151 N (0.484 lbs) producing 2.615 Nm/N of torque per weight (t-p-w) ratio. The cogging torque is 19.99 mNm per unit weight. The optimized ORPMSM model, when operated at rated speed, possesses 1.885 N (0.423 lbs) of weight (12.4 % reduction). It produces 18.04 mNm of cogging torque per unit weight (9.8 % reduction). The t-p-w for the optimized model is found to be 2.779 Nm/N (6.3 % increase). This enables UAV device to have better propeller efficiency. The reduced weight of motor ensures reduction in moment of inertia. This helps in obtaining lesser values of radius of gyration. The lesser the value of radius of gyration, lesser is the deflection in pitch, roll and yaw angle. This ensures the ease of overcoming the thrust in lifting load along three dimensional axes. It becomes easier to control such advantageous models in UAV drive system.

### CONCLUSION

Inspired by accuracy results obtained by brainless slime mold optimizer, the OR-PMSM geometry, having unique IPIM structure, has been optimized in this article, using a modified version of SMO.

First, the thorough explanation of all CBs, introduced in conventional SMO, is presented. In the analysis part of the paper, the performance of conventional SMO by adding individual CBs is tested for various Benchmark functions. Their convergence profiles



are compared and complexity analysis is presented.

It is concluded, from various experiments performed, that the addition of CBs to SMO, not only completes the SMO behavior in nature, but also enhances the overall performance of SMO in producing accurate results, for a given objective function. Thus, the performance of CB-SMO is compared with few popular MH techniques. It is found that CB-SMO produces convincingly better results for unimodal as well as multimodal Benchmark functions.

Furthermore, for implementation of the proposed CB-SMO to practical field, the DOP of OR-PMSM-IPIM is designed in order to optimize its performance. For designing DOP, LSM technique is implemented for obtaining cogging torque, as a function of three design parameters, i.e., magnet width, rotor width and motor height.

Then, this DOP is solved by proposed CB-SMO as well as few classical MH techniques, in order to minimize the cogging torque. It is found that CB-SMO outperforms other techniques to supply the least cogging torque by obtaining optimized value of design parameters.

Thereafter, these optimized values of design parameters are utilized to design the overall quadcopter model in order to compare the t-p-w ratios of initial and optimal models.

In UAV application, the energy conversion system should have high propeller efficiency, in order to perform better while undergoing VTOL operation [39-41]. The propeller efficiency is affected by angle of propeller and advance ratio, which is the ratio of freestream air speed to propeller tip speed of quadcopter.

Hence, by comparing initial and optimal models, it is found that following achievements are noticeable out of this work:

- In optimal model, there is nearly 12 % reduction in weight. It leads to almost 10 % reduction in cogging torque per unit weight. This ensures reduction in noise and enhancement in “smoothness” of VTOL operation.
- The overall t-p-w of the optimal model is enhanced by nearly 6 %, as compared to the initial model. This enables UAV device to have better propeller efficiency.
- The reduced weight of motor ensures reduction in moment of inertia. This helps in obtaining lesser values of radius of gyration. (*The lesser the value of radius of gyration, lesser is the deflection in pitch, roll and yaw angle*)
- This ensures the ease of overcoming the thrust in lifting load along three dimensional axes.
- It becomes easier to control such advantageous models in UAV drive system.

## REFERENCES

- [1] Bozkurt, A.; Baba, A.F.; Oner, Y., “Design of Outer-Rotor Permanent-Magnet-Assisted Synchronous Reluctance Motor for Electric Vehicles”, *Energies* 2021, Vol. 14, No. 3739, pp. 1-12, <https://doi.org/10.3390/en14133739>.
- [2] J. M. Ahn, J. C. Son and D. K. Lim, “Optimal design of Outer rotor surface mounted permanent magnet synchronous motor for cogging torque reduction using territory particle swarm optimization”, *Journal of Electrical and Engineering & Technology*, Nov. 2020, DOI: 10.1007/s42835-020-00599-z.
- [3] P. J. Masson, C. A. Luongo, T. Nam, H. D. Kim, D. Mavris, G. V. Brown, D. Hall and M. Waters, “Next generation more electric aircraft: A potential application for HTS superconductors”, *IEEE Trans. On Applied Superconductivity*, vol. 19, no. 3, pp. 1055-1068, June 2009.
- [4] M. Rosu, P. Zhou, D. Lonol, M. Popescu, D. Lin, F. Blaabjerg, V. Rallabandi and D. Staton, “Multi-physics simulation by design for electrical machines, power electronics and drives”, *IEEE Press*, Hoboken, NJ: Wiley 2018.
- [5] Akash, Arumugam, Vijayaraj Stephen Joseph Raj, Ramesh Sushmitha, Boga Prateek, Sankarasubramanian Aditya, and Velloorillom Madhavan Sreehari. “Design and analysis of VTOL operated intercity electrical vehicle for urban air mobility”, *Electronics*, Vol. 11, No. 1, Dec. 2021.
- [6] Finger, D.F., Braun, C. & Bil, C., “Impact of electric propulsion technology and mission requirements on the performance of VTOL UAVs”, *CEAS Aeronaut Journal*, Vol. 10, pp. 827–843, Dec. 2019. <https://doi.org/10.1007/s13272-018-0352-x>.
- [7] W. Cao, B. C. Mecrow, G. J. Atkinson, J. W. Bennett and D. J. Atkinson, “Overview of electric motor technologies used for more electric aircraft (MEA)”, *IEEE Trans. On Industrial Electronics*, vol. 59, no. 9, pp. 3523-3531, Sept 2012.
- [8] M. Mutluer, “Analysis and design optimization of permanent magnet synchronous motor with external rotor for direct driven mixer”, *Journal of Electrical Engineering & Technology*, March 2021, DOI: 10.1007/s42835-021-00706-8.
- [9] E. I. Mbadiwe and E. B. Sulaiman, “Design and optimization of outer rotor permanent magnet flux switching motor using transverse segmental rotor shape for automotive applications”, *Ain Shams Engineering Journal*, vol. 12, pp. 507-516, Sept. 2020.
- [10] C. Guerroudj, Y. L. Karnavas, J. F. Charpentier, I. D. Chasiotis, L. Bekhouche, R. Saou and E. H. Zaim, “Design optimization of outer rotor toothed doubly salient permanent magnet generator using symbiotic organisms search algorithm”, *Energies*, vol. 14, pp. 1-25, April 2021.
- [11] H. M. Taha and I. R. Talnaab, “Designs of PMSMs with inner and outer rotors for electric bicycle applications”, *Kurdistan Journal of Applied Research*, vol. 4, no. 1, pp. 20-25, June 2019.
- [12] Z. Shi, X. Sun, Y. Cai, X. Tian and L. Chen, “Design optimization of an outer rotor permanent magnet synchronous hub motor for a low speed campus patrol EV”, *IET Electric Power Applications*, vol. 14, no. 11, pp. 2111-2118, August 2020.
- [13] K. T. Chau, D. Zhang, J. Z. Jiang, C. Liu and Y. Zhang, “Design of a magnetic geared outer rotor permanent magnet brushless motor for electric vehicles”, *IEEE Trans. on Magnetics*, vol. 43, no. 6, pp. 2504-2506, June 2007.
- [14] M. Z. Ahmed, E. Sulaiman, G. M. Romalan and Z. A. Haron, “Optimal torque investigation of outer rotor hybrid excitation flux Switching machine for in wheel drive EV”, *ARPN Journal of Engineering and Applied Sciences*, vol. 10, no. 19, pp. 8839-8845, Oct. 2015.
- [15] A. -S. Isfanuti, L. N. Tutelea, I. Boldea, T. Staudt and P. E. da Silva, “Outer Ferrite-PM-Rotor BLAC Motor Characterization: FEM-Assisted Optimal Design and Preliminary Experiments,” in *IEEE Transactions on Industry Applications*, vol. 56, no. 3, pp. 2580-2589, May-June 2020, doi: 10.1109/TIA.2020.2979672.
- [16] J. M. Ahn, J. C. Son and D. K. Lim, “Optimal design of Outer rotor surface mounted permanent magnet synchronous motor for cogging torque reduction using territory particle swarm optimization”, *Journal of Electrical and Engineering & Technology*, Nov. 2020, DOI: 10.1007/s42835-020-00599-z.
- [17] S. L. Ho, S. Yang, G. Ni and J. M. Machado, “A modified ant colony optimization algorithm modeled on tabu-search methods”, *IEEE Trans. on Magnetics*, vol. 42, no. 4, pp. 1195-1198, April 2006.
- [18] B. Brandslatter and U. Baumgartner, “Particle swarm optimization- mass spring system analogon”, *IEEE Trans. on Magnetics*, vol. 38, no. 2, pp. 997-1000, Mar. 2002.
- [19] C. H. Im, H. K. Jung and Y. J. Kim, “Hybrid genetic algorithm for electromagnetic topology optimization”, *IEEE Trans. on Magnetics*, vol. 39, no. 5, pp. 2163-2169, Sep. 2003.

- [20] Z. Bayraktar, M. Komurcu, J. A. Bossard and H. D. Warner, "The wind driven optimization technique and its application in electromagnetics", *IEEE Trans. on Antenna Propagation*, vol. 61, no. 5, pp. 2745-2757, May 2013.
- [21] Y. Shi, "Brain storm optimization algorithm", in *Proc. 2nd Int. Conf. Swarm Intell.*, Chongqing, China, pp. 303-309, Jun. 2011.
- [22] Tang, A.D., Tang, S.Q., Han, T., Zhou, H. and Xie, L., "A modified slime mould algorithm for global optimization", *Computational intelligence and neuroscience*, Vol. 21, pp. 1-22, Nov. 2021.
- [23] M. K. Naik, R. Panda and A. Abraham "Adaptive Opposition Slime Mold Algorithm", *Soft Computing*, Vol. 25, pp. 14297-14313, August 2021.
- [24] Y. Xiao, X. Sun, Y. Zhang, Y. Guo, Y. Wnag and J. Li, "An improved slime mould algorithm based on tent chaotic mapping and nonlinear inertia weight", *International Journal of Innovative Computing, Information and Control*, Vol. 17, No. 6, pp. 2151-2176, Dec. 2021.
- [25] A. Hamed, M. Ebeed, A. Refai, M. Sattar, A. Elbaset and T. Ahmed, "Application of SMO for Optimal Allocation of Dstatcom and PV system in Real Egyptian Radial Network", *Sohag International Journal*, Vol. 1, No. 1, pp. 16-24, March, 2021.
- [26] J. Jones, "Applications of multi-agent Slime Mold Computing", *International Journal of Parallel, Emergent and Distributed System*, pp. 1-34, Nov. 2015.
- [27] Y. Liu, A. Heidari, X. Ye, G. Liang, H. Chen and C. He, "Boosting slime mould algorithm for parameter identification of photovoltaic models", *Energy* 234 (2021) 121164.
- [28] H. Jia, W. Zhang, R. Zheng, S. Wang, X. Leng and N. Cao, "Ensemble mutation slime mold algorithm with restart mechanism for feature selection", *International Journal of Intelligent system*, Dec. 2021.
- [29] M. A. Basset, V. Chang and R. Mohamed, "HSMA WOA: A hybrid novel slime mold algorithm with whale optimization for tackling the image segmentation problem of chest X-ray images", *Applied Soft Computing*, Vol. 95, Oct. 2020.
- [30] Z. Cui, H. Hou, H. Zhou, W. Lian and J. Wu, "Modified Slime mold algorithm via Levy Flight", *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, 2020.
- [31] Gao, H.; Liang, G.; Chen, H., "Multi-Population Enhanced Slime Mould Algorithm and with Application to Postgraduate Employment Stability Prediction", *Electronics*, Vol. 11, No. 209, 2022.
- [32] M. K. Naik, R. Panda and A. Abraham, "Normalized square difference based multilevel thresholding technique for multispectral images using leader slime mould algorithm", *Journal of King Saud University- Computer and Information Sciences*, Oct. 2020.
- [33] L. Liu, D. Zhao, F. Yu, A. A. Heidari, J. Ru, H. Chen, M. Mafarja, H. Turabieh and Z. Pan, "Performance optimization of differential evolution with slime mould algorithm for multilevel breast cancer image segmentation", *Computers in Biology and Medicine*, Vol. 138, pp. 1-5, Nov. 2021.
- [34] C. R. Reid and T. Latty, "Collective Behavior and swarm intelligence in Slime Moulds", *FEMS Biology Reviews*, Oxford, Vol. 40, pp. 798-806, August 2016.
- [35] T. Latty and M. Beekman, "Food quality affects search strategy in the acellular slime mould, *Physarum polycephalum*", *Behavioral Ecology*, Vol. 20, pp. 1160-1167, August 2019.
- [36] R. Singh, S. Ranjan, T. Pradhan, and K. Raju Dhenuvakonda, "Calibration and frequency estimation in sensors for electrical parameter measurement using regression and metaheuristic based models", *Expert Systems*, Vol. 40, no. 3, Nov. 2022.
- [37] Khuri, André I., and Siuli Mukhopadhyay. "Response surface methodology", *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol. 2, no. 2, pp. 128-149, Mar. 2010.
- [38] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? –Arguments against avoiding RMSE in the literatureT.", *Geoscientific Model Development*, Vol. 7, pp. 1247-1250, Jan. 2014.
- [39] P. J. Masson, C. A. Luongo, T. Nam, H. D. Kim, D. Mavris, G. V. Brown, D. Hall and M. Waters, "Next generation more electric aircraft: A potential application for HTS superconductors", *IEEE Trans. On Applied Superconductivity*, vol. 19, no. 3, pp. 1055-1068, June 2009.
- [40] M. Rosu, P. Zhou, D. Lonol, M. Popescu, D. Lin, F. Blaabjerg, V. Rallabandi and D. Staton, "Multi-physics simulation by design for electrical machines, power electronics and drives", *IEEE Press*, Hoboken, NJ: Wiley 2018.
- [41] W. Cao, B. C. Mecrow, G. J. Atkinson, J. W. Bennett and D. J. Atkinson, "Overview of electric motor technologies used for more electric aircraft (MEA)", *IEEE Trans. On Industrial Electronics*, vol. 59, no. 9, pp. 3523-3531, Sept 2012.



**MONIKA VERMA** received her B. Tech. Degree in Electrical Engineering from G. B. Pant University of Agriculture & Technology, Uttarakhand, India in 2013, and her M. Tech. Degree in Power Electronics and Drives from Vellore Institute of Technology, India in 2016, and her Ph.D. degree in Electrical Engineering from Delhi Technological University, India in 2022. She is presently working as Software Development Engineer in the Centre of Excellence for Electric Vehicle and Related Technologies, at Delhi Technological University, India. She is Institute of Electrical and Electronics Engineers (IEEE) student member, and Institution of Engineering and Technology (IET) associative member.



**MINI SREEJETH** Born in Kerala, India, she received her B Tech. degree from Mahatma Gandhi University, Kerala, India; her M. Tech. degree from Calicut University, Kerala, India; and her Ph.D. degree from the University of Delhi, New Delhi, India. She joined Delhi College of Engineering (now DTU) as a Senior Lecturer in 2007, where she is presently working as a Professor in the Department of Electrical Engineering. Her research interests include power electronics, modeling, control and the operation of drives. Professor Sreejeth is a Life Member of the Indian Society for Technical Education, New Delhi, India; and a Senior Member of IEEE and WIE.



**MADHUSUDAN SINGH** received his Ph.D. Degree from University of Delhi, New Delhi, India respectively and is presently a Professor in the Department of Electrical Engineering at Delhi Technological University. His research interests are in the area of modeling and analysis of electrical machines, voltage control aspects of self-excited induction generators, power electronics and drives. Dr Singh is a fellow of the Institution of Engineers (IE), India and of the Institution of Electronics and Telecommunication Engineers, New Delhi, India. He is also a member of the IEEE (USA).

#### Appendix A: Samples collected from FEM experiments and impact of selection of prescribed design parameters

For UAV application, electric motor should be reliable, i.e., its critical efficiency should be higher and the motor should be of lighter weight. However, it is contradictory to obtain minimized weight and maximized efficiency of the motor simultaneously. To compensate this adversity, the optimal designing of related parameters of OR-PMSM is essential for designers. This is the reason why parameters  $d_1$ ,  $d_2$  and  $d_3$  (specified in Fig. 5) are selected as design parameters for minimizing cogging torque. Impact of selecting above mentioned design parameters are discussed below:

- The designer should compute the motor design approximately. The electric as well as magnetic loading of motor are related to size and motor's endurance.
- From theoretical point of view, the increase in stack length results in the improvement of motor's efficiency but at the cost of increased mass of the motor.
- The radial length of magnets in OR-PMSM leads to increase in intensity of magnetic induction in rotor magnetic circuit of motor. Thus, increment in magnetic length results in obtaining the condition of magnetic saturation. Therefore, the optimization of radial length of magnets in rotor can rectify magnetic saturation condition.

- However, it may increase the cost and mass of motor. Also, the increased current density in stator winding leads to enhanced copper loss causing reduced efficiency and increment in the temperature of motor.

TABLE A.1  
DESIGN OF EXPERIMENT SAMPLE DATA FOR  $T_c$  (COGGING TORQUE)

No.	$d_1(\text{mm})$	$d_2(\text{mm})$	$d_3(\text{mm})$	$T_c(\text{mNm})$
1	3	2	31	43.0
2	3	0.5	31	23.4
3	3	0.5	30	22.6
4	4	0.5	32	31.9
5	4	0.5	31	30.9
6	4	0.5	30	29.9
7	3.5	0.5	32	31.0
8	3.5	0.5	31	30.0
9	3.5	0.5	30	29.1
10	3	1.5	32	46.9
11	3	1.5	31	45.4
12	3	1.5	30	44.0
13	4	1.5	32	42.1
14	4	1.5	31	40.8
15	4	1.5	30	39.6
16	3.5	1.5	32	40.8
17	3.5	1.5	31	39.5
18	3.5	1.5	30	38.2
19	3	2	32	44.4
20	3	0.5	32	29.9
21	3	2	30	41.6
22	4	2	32	65.9
23	4	2	31	63.9
24	4	2	30	61.7
25	3.5	2	32	45.9
26	3.5	2	31	44.4
27	3.5	2	30	43.0
28	3	1	32	39.0
29	3	1	31	37.7
30	3	1	30	36.5
31	4	1	32	58.8
32	4	1	31	56.9
33	4	1	30	55.1
34	3.5	1	32	35.1
35	3.5	1	31	34.0
36	3.5	1	30	32.9

# Comparative Study on Forecasting of Schedule Generation in Delhi Region for the Resilient Power Grid Using Machine Learning

Lakshmi D, Ravi Shekhar Tiwari, Neelu Nagpal, *Senior Member, IEEE*,  
Neelam Kassarwani, Vishnuvarthanan G, Abhishek Srivastava

**Abstract**—The increasing use of Renewable Energy Resources (RES) in energy generation has led to the transformation of the conventional electrical grid into a more adaptable and interactive system, and this has made electrical load prediction a crucial aspect of smart grid operation. Short-Term Load Forecasting (STLF) is the ultimate requirement for the essentialities, such as planning, scheduling, management, and trading of electricity. In the proposed work, a forecasting engine model is developed to figure out the load of the upcoming twelve months (2020) in the Delhi metropolis, and this is accomplished by integrating real and dynamic meteorological data, calendar data, and load patterns for the successive two years (2017-2018). It is performed using different ensemble models, such as XGBoost, Gradient Boosting, AdaBoost, Random Forest (RF) algorithms, and deep learning models such as Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), Gated Recurrent Unit (GRU) and the Prophet algorithm. The simulation results of the proposed models are obtained on the Python platform using Delhi weather, load, and calendar data. Further, the STLF is analyzed using 14 different models on the basis of 78 scenarios, and 8 data sets are analyzed in conjunction. The train, validation, and test accuracy have been considered as validation metrics, both on hourly and daily load forecasting, to validate the overfitting in terms of the train, validation, and test loss. A comparative study is made to show that the predictions of LSTM and GRU outperform with 100% accuracy.

**Index Terms**—Deep Learning, Electrical Load Forecasting, Feature Extraction, Machine Learning, Ensemble Learning, Resilient Power Grid, Time Series Analysis. Short-Term Load Forecasting

## I. INTRODUCTION

### A. Motivation and Problem Statement

**P**ower system resilience depends on demand forecasting. Accurate load forecasting helps power utilities balance energy production and consumption, operate efficiently, and avoid blackouts and losses [1], [2]. Authorities can buy or sell electricity to other grids or firms in energy markets when

generation is low or high. Bids are submitted at the Day-Ahead Market (DAM) or Real-Time Market at exchanges like the Indian Energy Exchange Limited (IEX). Thus, precise load forecasting helps utilities choose the best bidding approach and maximize economic benefits [3]. Short-Term Load Forecasting (STLF), Medium-Term Load Forecasting (MTLF), and Long-Term Load Forecasting (LTLF) are all parts of Electrical Load Forecasting (ELF). STLF is crucial for operational decisions like maintenance scheduling, energy management, and daily power system operations [4]. STLF predicts load demand for hours and days ahead. This helps utilities optimize infrastructure, allocate resources, and make educated power generation, transmission, and distribution decisions [1], [2]. Accurate short-term load forecasting saves resources and improves power system security [5], [6]. Recently, machine learning algorithms, statistical models, and weather data have improved load forecasting accuracy. Power utilities may now make more trustworthy and informed decisions, benefiting both utilities and consumers [7].

### B. Literature Review

Deep Learning (DL) and Machine Learning (ML) frameworks connected to STLF experimental work were used to study ELF research publications [8], [9], [10], [11], [12]. A variety of DL techniques, such as LSTM, RNN, and GRU, have been implemented to capture the long-term dependencies and trends in electricity demand [13]. These techniques are effective for identifying complex patterns and relationships in data for suitable load forecasting [14]. RNN is a variety of neural networks that are particularly well-suited for modelling sequential data, such as time-series data. The Long Short-Term Memory (LSTM) network for modeling sequential time-series data is a popular RNN variant for short-term load forecasting. It has three gates: an input gate to determine the correct information to be stored in the cell state, a forget gate to discard information, and an output gate to generate information from the cell state. STLF is performed using real-time data and the trained LSTM network. GRU is another type of RNN technique that is designed to remember long-term dependencies in time-series data, making it ideal for forecasting electricity loads. Although GRUs are similar to LSTM networks, they have fewer parameters and gating mechanisms to make them computationally less expensive, faster to train, and easier to understand and interpret. To train a GRU network for short-term load forecasting, archived data on electricity demand is used as the input, and the corresponding electricity load for the next time period is used as the output. Using backpropagation across time, the network is trained to minimize the discrepancy between the

Lakshmi D is with the School of VIT Bhopal University, School of Computing Science and Engineering Sehore-466114, Madhya Pradesh, India (e-mail: lakshmi.lifeofdivine@gmail.com).

R. S. Tiwari is with Mahindra University, Hyderabad, Telangana, India (e-mail: tiwari11.rst@gmail.com).

N. Nagpal is with the Department of Electrical and Electronics Engineering, Maharaja Agrasen Institute of Technology, New Delhi-110086, India (e-mail: nagpalneelu1971@ieee.org)

N. Kassarwani is with the Department of Electrical and Electronics Engineering, Maharaja Agrasen Institute of Technology, New Delhi-110086, India (e-mail: neelam.kassarwani@gmail.com)

Vishnuvarthanan G is with the School of VIT Bhopal University, School of Computing Science and Engineering Sehore-466114, Madhya Pradesh, India (e-mail: gvvarthanan@gmail.com).

A. Srivastava is with the School of VIT Bhopal University, School of Computing Science and Engineering Sehore-466114, Madhya Pradesh, India (e-mail: alnumay@ksu.edu.sa)

projected and real loads. A case study on ELF used more than 20 ML frameworks. Only SVM and ANN have outperformed other ML algorithms. [15]. The case study has involved dynamic variables like weather data, solar irradiation measure, population, electricity price/kWh, and the Gross National Income (GNI) per capita of Cyprus. [17] has successfully experimented with statistical time series models and ensemble ML models for day-load prediction. In the proposed work, the STLF was analyzed using classic ML models, DL models, DL ensemble models, seq-2-seq models, and Dynamic Mode Decomposition (DMD) [16]. The study by [18] has inferred that the applied ANNs and ensemble ANN methods reduce estimation errors as compared to other forecasting methodologies. Further, DL algorithms have proven their effectiveness in big data processing, leading to improved prediction accuracy [19]. Employing forecasting error correction techniques, DL models such as DNN, CNN, RNN, and LSTM have shown high accuracy [20] while the hybrid CNN-LSTM model has been utilized for feature extraction and sequence learning [21]. The performance of DNN was found to be better than the LSTM models to forecast medium- and long-term power consumption patterns [22], [23] has proposed an RNN model using the Input Attention Mechanism (IAM) and Hidden Connection Mechanism (HCM) for the STLF.

Many hybrid models such as LSTM-RNN [24], new ensemble model, SELNet [25], stacked denoising auto-encoders [26], integrated CNN and LSTM [27], multi-layer bidirectional RNN, utilizing both LSTM and GRU [28], and the Bi-LSTM-Auto-Encoder model [29] have been proposed for high-precision prediction of STLF. Recent trends of integration of different models have been reported for feature selection and forecasting respectively by [30] (Auto Correlation Function (ACF) and the Least Square Support Vector Machine (LSSVM)), [31] (ACF and LSTM and GRU based models), [32] (RF and GRU), [33] (Residual Convolutional Neural Network (R-CNN) and multilayered LSTM), and [35] (CNN and BiGRU). Also, the dimensionality reduction has been achieved using Principal Component Analysis (PCA) [34], [35]. The experimental approach with different models including exploratory analysis, statistical time-series models, classical machine learning, and DL models has shown better performance using transfer learning and meta-learning techniques [36].

### C. Contribution of work

This study uses the Delhi load dataset to test 2020 forecast models. The pandemic lockdown has drastically changed workplace operations and work, and in conjunction, this study will also test the load scheduling prediction's efficacy. To have a contemporary insight, papers published from the year 2015 to the first quarter of 2023 have been reviewed. Exploratory analysis, statistical time-series models, classical machine learning, and DL models have been used and tested to determine their efficacy. This research improves power grid resilience and dependability by creating robust load forecasting models. This study improves short-term load forecasting (STLF) with DL and ML algorithms. The Prophet algorithm improves load predictions in this paper. The study now fine-tunes DL and ML hyperparameters and integrates the Prophet approach to improve load prediction. To maximize utility load and generation balance, power generation and demand must be matched. These approaches capture intricate

patterns, seasonality, and other features to properly predict load and optimize power generation and demand.

- This research investigates the effectiveness of the performance of ML and DL methods for short-term load forecasting (STLF) in a metropolitan area that experiences fluctuating weather conditions, with extreme heat in the summer (ranging from 40<sup>0</sup> to 48<sup>0</sup> C) and extreme cold in the winter (ranging from 10<sup>0</sup> to 50<sup>0</sup> C). Adding to the woes, the load pattern in this area is unstable due to its fast-paced urban development.
- To improve precision and dependability, load forecasting systems integrate ML and DL models and assess optimized models. Forecasting models are compared in different settings and datasets.
- The research entails assessing the performance of various forecasting models through a comparative analysis of diverse scenarios and datasets.

### D. Structure of Paper

The paper is organized as follows: Section I provides an introduction, while Section II presents the proposed methodology for forecasting after a literature review. Sections III, IV, and V discuss STLF approaches such as exploratory analysis, time series analysis, ML, and DL models. Section VI discusses the findings, while Section VII concludes.

## II. PROPOSED METHODOLOGY

STLF applying the machine learning technique involving historical data on electricity demand is used to predict the expected electricity load in the near future (typically from a few hours to a few days ahead). This method predicts power generation and distribution needs to help utility firms manage their energy supply cost-effectively. Due to these considerations, the datasets for 2017–2018 and 2020 were pooled. COVID-19 has set a completely different paradigm for lifestyle and working nature. At times, such an unprecedented situation towards electrical load forecasting needs to be considered. The Indian Electrical Grid operates at a frequency of 50 Hz, and if the grid's frequency is above this level, power plants must reduce generation for the next block, which lasts 15 minutes, making each block consist of 15 minutes duration. In a day, 96 blocks ((15 minutes\*96 blocks)/60=24 hours) are used in the Indian scenario. In this way, the power plant can use the Indian Energy Exchange Limited to trade any excess energy produced. The Delhi Electricity Board is the source of the electrical load data used in this research study. Due to Delhi's unpredictable weather, this data source was chosen. Data source opinion is also influenced by Delhi's dense population, mineral-rich and industrial regions, attractive infrastructure, commercial industries, and government services. Thus, electricity consumption varies widely. The dataset includes 2017–2020, including the pandemic year. This dataset fluctuates due to numerous variables.

The adoption of ML algorithms for short-term load forecasting is central to this research. These algorithms, capable of analyzing vast volumes of historical data, can discern patterns and trends far beyond the scope of human capabilities. In this research, the focus is on utilizing Linear Regression (LR), K-nearest neighbors (KNN), and Support Vector Machine (SVM) models. Each of these algorithms has its own strengths and is especially

adept at processing electrical load data and delivering methodical outcomes and transversions. An LR model load-time connection simplifies result interpretation. Neighbor-based KNN captures complex load data patterns. Finally, the SVM's high-dimensional feature space is appropriate for load forecasting generalization.

By exploiting the unique advantages of these three ML models, the aim is to enhance the accuracy and efficiency of short-term load forecasting. Besides ML-based algorithms, STLTF algorithms also rely on input data such as weather data, historical electricity usage, and real-time data from sensors and meters. By combining these different sources of data with machine learning algorithms, utility companies can make accurate predictions about future electricity demand and manage their energy supply cost-effectively. The procedures opted for STLTF are summarized as shown in Fig. 1.

#### A. Data set Description

The electrical load data for the years 2017, 2018, and 2020 is used in this research study. In addition to this, weather data (maximum temperature in Celsius, the minimum temperature in Celsius, and the relative humidity at two different time points) and calendar data (weekends, holidays, and weekdays) were taken into account. The primary rationale behind selecting the data from the Delhi Electricity Board was due to the highly unpredictable weather in the area. The time-series data is transformed into a suitable number of features for creating machine models. A total of eight distinct electrical load/demand data sets are constructed using different features, such as hourly, daily, weather, and calendar data.

Electrical load/demand data consists of two basic components: a timestamp in hours and a distribution of electrical load in megawatts. The maximum temperature in Celsius, the minimum temperature in Celsius, and the relative humidity at two different time points comprise the weather data. Calendar data is represented in the following manner using the label encoding technique: Weekdays (Monday to Friday) are labeled as '0', whereas weekends (Saturday/Sunday) are labeled as '1', and public holidays are labeled as '2'.

These 8 unique combinations of datasets were examined to determine the most effective and suitable features. The years 2017, 2018, and 2020. The dataset was shuffled so that the dependency of the data on the previous day is 100% related to the next day because power consumption is dependent on the weekday (working or non-working), weather conditions, and especially the rush/peak hour, i.e., office and school duration. For that, the dataset has been shuffled, and the DL models are trained for the log of 24, i.e., considering the past 24 hours in the RNN, LSTM, and GRU models. Converting hourly data to daily data requires data conversion. The data sets available for analysis are training and testing. 2017 and 2018 data are used for training. However, 2020 data is utilized during testing. Data testing is difficult because of the insufficiency faced in training the data. Training data are from 2017 and 2018, but test data are from 2020 when the lockdown was proclaimed due to the COVID-19 pandemic. In 2020, electricity use will have changed drastically. Working from home is mandatory as schools, businesses, and corporate offices are closed. Thus, this research employs time series models and ensemble machine learning models.

#### B. Exploratory Data Analysis

The periodicity analysis helps to understand the power demand pattern [36]. To analyze trends, an exploratory analysis was conducted, and five different types of trend analyses were performed, including daily, monthly, holiday, weekday, and weekend load trend analyses, as shown in Fig. 2. The results showed that the highest electricity consumption occurred in July, while January had the lowest consumption. The analysis also revealed that weather conditions had a significant impact on electricity usage, with higher consumption during the summer months and lower consumption during the winter. Moreover, the dataset indicated that weekdays had higher electricity consumption compared to weekends.

1) *Time Series Analysis*: The aim of this study is to predict future electrical load demand in megawatts based on past observations of electrical load usage taken at fixed intervals. For this, four univariate time series models, namely Auto-regressive (AR), Moving Average (MA), Auto-Regressive Integrated Moving Average (ARMA), and SARIMAX (Seasonal Auto-Regressive Integrated Moving Average with Exogenous Factors) have been trained on the hourly and anywise load data sets as shown in Table I. In this table, p denotes the number of auto-regressive terms and q denotes the number of moving average terms. The aforementioned criteria represent the number of past forecast errors that are used to correct the forecast for the current time step; d denotes the number of times the data has been different. Differencing is a technique used to make non-stationary time series data stationary (i.e., data with a constant mean and constant variance). A time series is said to be stationary when the statistical properties (e.g., mean, variance) of the series do not depend on time. An augmented Dickey-

TABLE I  
ARMA AND SARIMAX

Name of the Model	p	d	q	s
Hourly prediction with electrical load data	5	1	1	12
Day-wise prediction with electrical load data	1	0	3	12

Fuller test is also performed to determine the stationarity of the data which functions well under fluctuations in the time series. ARMA has manual ways and inbuilt algorithms to obtain the optimized values of p, d, and q. In this case, the values of p, d, and q are determined using the auto-arma inbuilt function. Stationarity, autocorrelation, and seasonality of time series data determine these parameters. The Auto-Arima function selects d using the KPSS unit root test and p and q using the AIC information criterion. Auto-arma's p, d, and q may not be optimal. Many industries, including energy, employ statistical models because they are easy to execute and require less processing power. Traditional statistical models cannot predict harsh weather, unexpected consumption patterns, or sudden energy demand changes. Univariate statistical models forecast electrical loads using one parameter, like previous energy usage, weather, or population. However, in reality, there are many factors that affect energy consumption, including demographics, economic conditions, consumer behavior, and technological advancements. Thus, one parameter cannot accurately predict the electrical load. Machine learning models may overcome these constraints, and they can use various factors and previous data to predict

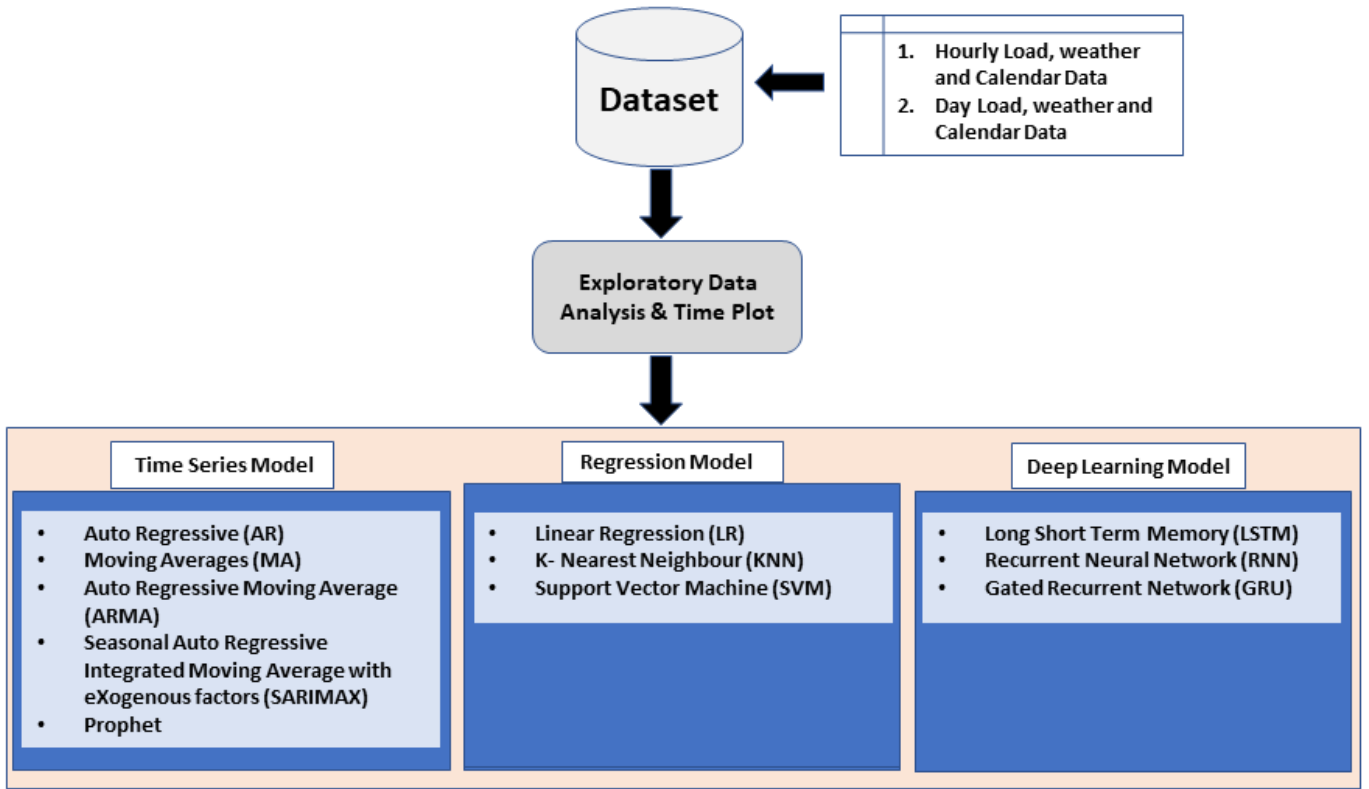


Fig. 1. Proposed Methods used for STLF.

accurately. They can adapt to changing conditions and spot intricate patterns that statistical models. Machine learning models can improve their accuracy with more data and feedback, making them perfect for estimating electricity loads under uncertain and dynamic conditions.

### III. MACHINE LEARNING MODELS

Electrical networks and resources require load-demand prediction research. One can find the optimal machine learning model by training it on diverse datasets. The performance of each model is measured using MAE, MSE, RMSE, and R-Square. These metrics offer distinct perspectives on the model performance: MAE: the mean absolute difference between expected and actual values. It estimates model mistakes. Lower error numbers are desirable. MSE: the mean squared difference between expected and actual values; It is more sensitive to outliers than MAE since it squares differences and considers huge errors more severely. Lower is preferable. RMSE: the square root of MSE measures the residual standard deviation (prediction errors). R-Square (coefficient of determination): the independent variable(s) explain 0–1 of the dependent variable's variation. It measures the model's data fit. The more variance explained, the better. The user has the choice to select the best predictive model by comparing these metrics among models keeping in mind that model selection generally involves balancing complexity and performance. SVM and Random Forest yield superior results but are computationally expensive and more difficult to interpret than Multiple Linear Regression and KNN. The regression model validation metrics are shown in Table II.

TABLE II  
REGRESSION MODEL VALIDATION METRICS

Name of Model	Loss Function			$R^2$ (%)	Data set
	MAE	MSE	RMSE		
LR	665.83	694152.05	833.15	56.78	HL+W+C
KNN	415.24	286639.813	535.38	76.23	DL + W + C
SVM	429.49	303490.63	550.89	74.53	DL and W

HL-Hourly Load;DL- Day Load; W-Weather; C-Calendar

### IV. ENSEMBLE MODELS

The utilization of ensemble techniques with a combination of base learners and a voting approach is a powerful approach for making predictions. It is pertinent to note that the 'Random Forest' algorithm works with homogeneous base learners, specifically as a decision tree algorithm. On the other hand, the remaining ensemble techniques (XGBoost, Gradient Tree Boosting, and AdaBoost) employ heterogeneous base learners allowing for diversity in the models and potentially enhancing the overall performance and robustness of the ensemble. Table III presents the validation metrics for the ensemble models. The validation metrics included in the table provide valuable insights into the accuracy, precision, recall, F1 score, or any other relevant evaluation measures for each ensemble model. These validation metrics serve as important indicators of how well the ensemble models are performing, and they allow for a quantitative assessment of the predictive capabilities and effectiveness of the different ensemble techniques employed. Ensemble models integrate base learners (individual models) to increase prediction accuracy. The various Ensemble methods are



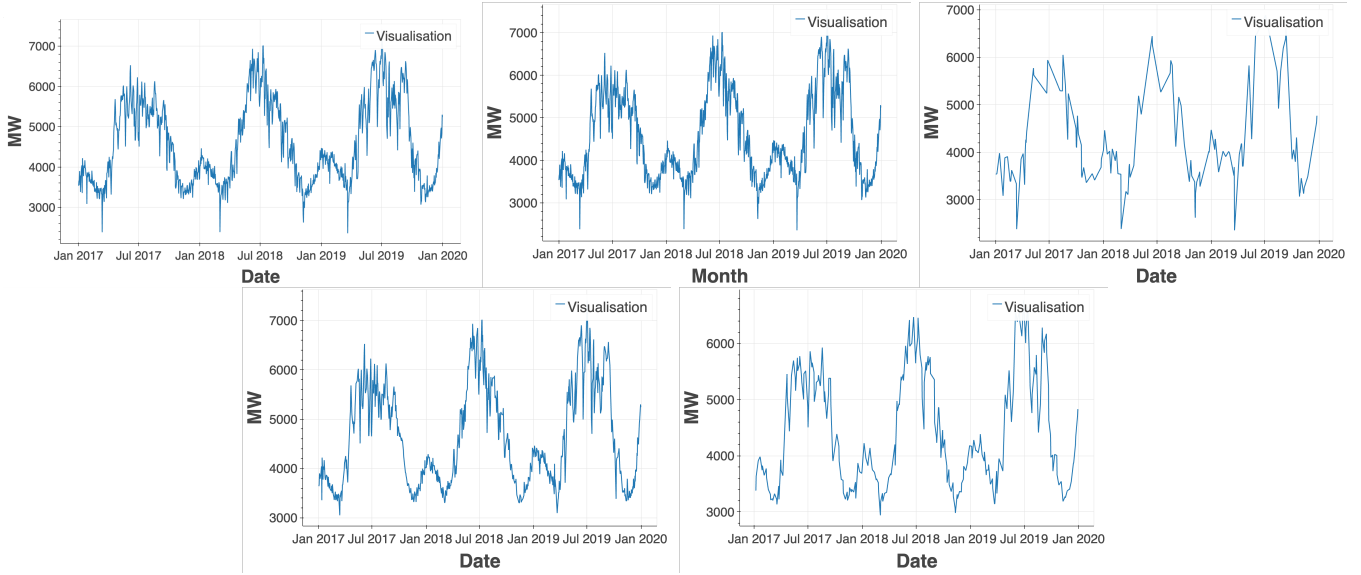


Fig. 2. Exploratory Data Analysis: (i) Daily Load Plot; (ii) Monthly Load Plot; (iii) Holiday Load Plot; (iv) Weekday Load Plot; (v) Weekend Load Plot.

described as follows:

- (i) **Random Forest:** This bagging-based ensemble learning technique uses decision trees as base learners. Bootstrap samples and random feature selection at each node create a "forest" of varied trees.
- (ii) **XGBoost:** Xtreme Gradient Boosting uses boosting-based ensemble learning with decision trees as the most frequent base learners. Boosting corrects faults by fitting successive models.
- (iii) **Gradient Tree Boosting (GBM):** Like XGBoost, GBM uses decision trees as base learners. GBM optimizes a loss function by correcting previous tree errors with each new tree.
- (iv) **AdaBoost (Adaptive Boosting):** Another boosting-based ensemble learning technique prioritizes misclassified training samples from previous models. Base learners with AdaBoost frequently use decision stumps (one-level decision trees).

## V. PROPHET ALGORITHM

Facebook's Prophet algorithm forecasts, and it provides precise time series forecasting without manual adjustment. Time series data patterns are captured and predicted using trend modeling, seasonality modeling, and holiday impacts. **Trend Modeling:** A versatile and customizable function models the time series data's overall trend. It supports linear, non-linear, and abrupt trend change points. Trend modeling uses a data-adaptive piecewise linear or logistic function. Prophet models yearly, weekly, and daily seasons. It models seasonal impacts automatically for reliable forecasts. Fourier series expansion approximates data periodicity to model seasonality.

**Holiday Effects:** The algorithm accounts for holidays and other events that can affect time series. Prophet offers built-in holiday datasets or bespoke holiday lists. Holiday effects allow the algorithm to account for time series behavior during holidays. The Prophet can handle time series outliers and missing values. "Piecewise linear approximation" fits several linear models to various data segments to handle outliers. The data is imputed using available information. The Prophet estimates trends and seasonality to impute missing values.

**Forecasting:** Historical data trains the model to predict the future. Users choose the prediction horizon. The program estimates trends, seasonality, and holiday effects to forecast values. The Prophet decomposes its model into trends, seasonality, holidays, and noise. The Prophet forecasting formula relies on the following: Trend Component: Models time series growth or decline. Prophet uses a piecewise linear trend model. Trend formula:

$$g(t) = (k + k_1 * (t - t_{\text{changept}})) * t + (m + m_1 * (t - t_{\text{changept}})) \quad (1)$$

Here,  $g(t)$  represents the trend value at time  $t$ .  $k$  and  $k_1$  are coefficients that determine the overall trend rate and its rate of change.  $m$  and  $m_1$  are coefficients that control the intercept and its rate of change.  $t_{\text{changept}}$  is the time of the trend changept, where the trend direction changes.

**Seasonality Component:** Seasonality shows time-series cycles. The prophet forecasts yearly and weekly seasonality. The seasonality formula is:

$$s(t) = \sum_{j=1}^J (a_j * \cos((2\pi j * t) / P)) + (b_j * \sin((2\pi j * t) / P)) \quad (2)$$

Here,  $s(t)$  represents the seasonality value at time  $t$ .  $J$  is the number of Fourier terms used to model the seasonality.  $a_j$  and  $b_j$  are the coefficients for the  $j^{\text{th}}$  Fourier term.  $P$  represents the period of the seasonality.

**Holiday Component:** The Prophet considers holidays and other events that may affect the time series. The holiday formula is:

$$h(t) = \sum [i = 1 \text{ to } I] (c_i * \text{is}_{\text{holiday}}(t, \text{holiday}_i)) \quad (3)$$

Here,  $h(t)$  represents the holiday effect at time  $t$ .  $I$  is the number of holidays included in the model.  $c_i$  is the coefficient that determines the impact of the  $i^{\text{th}}$  holiday.  $\text{is}_{\text{holiday}}(t, \text{holiday}_i)$  is an indicator function that returns 1, if time  $t$  corresponds to the  $i^{\text{th}}$  holiday, otherwise 0.

**Noise Component:** Noise is the time series' inexplicable variation. Gaussian is assumed here and the Noise formula is  $e(t) N(0, \sigma)$ . Here,  $e(t)$  represents the noise component at time

t.  $N(0, \sigma)$  denotes a Gaussian distribution with mean 0 and standard deviation  $\sigma$ .

Overall, the formula for Prophet forecasting can be expressed as:

$$y(t) = g(t) + s(t) + h(t) + e(t) \quad (4)$$

Where  $y(t)$  represents the forecast value at time t. Each of the components is explained above. In context, the prophet algorithm is used for the STLTF, which takes account of various factors and gives adequate impetus to balance the electric generation and demand put forth by the utilities. A right proposition is always maintained between the two with the intervention of the prophet algorithm, and Tables II and III are testimonials to the claim made.

## VI. DEEP LEARNING MODELS

The proposed DL techniques and the related block diagram are presented in Fig. 3. The following subsections discuss the procedure to obtain the STLTF using the DL models. Table III shows these models' performance in terms of MAE, MSE, RMSE, and R-squared. XGBoost, GBM, and AdaBoost use decision trees, especially for tabular data, even if one can use heterogeneous base learners. This is required to homogenize their basic learners. However, each model has its own strengths, weaknesses, and algorithmic quirks, making them effective ensemble learners. The proposed DL techniques and the related block diagram are presented in Fig. 3. The following subsections discuss the procedure to obtain the STLTF using the DL models.

### A. Pre-processing

The pre-processing of time-series data involves a sequence of three steps, namely: (i) normalization using the Min-Max technique: (ii) extraction of the weekday, and (iii) obtaining 24-hour lag data. These three steps are carried out to transform the data before it is fed into the deep learning model. Min-max normalization technique is used in data pre-processing to rescale the data values of a feature into a specific range. The purpose of this technique is to normalize the data so that it has a consistent scale and the values are within a particular range, typically between 0 and 1. This prevents data bias and erroneous predictions from large data values. Min-max normalization simplifies feature comparison and scale-sensitive machine learning methods.

Electrical load patterns can vary depending on the day of the week due to factors such as differences in commercial and residential energy usage, work schedules, and other social and economic factors. By extracting the weekday from the time series data, it becomes possible to analyze the data and identify patterns that may be unique to specific weekdays. This can be helpful in making predictions and forecasting energy usage, as well as in developing more accurate energy management strategies (kindly refer to Fig. 6 and Fig. 5 for better understanding and information interpretation). A 24-hour lag refers to the energy usage data from the previous day, which can be used to analyze the variation pattern of energy usage from day to day. This information can be used to develop more accurate energy management strategies, optimize energy usage, and predict energy demand in advance. Additionally, the 24-hour lag data can be used in conjunction with other data features to train machine learning models that can learn to make accurate predictions of energy usage, allowing for more efficient

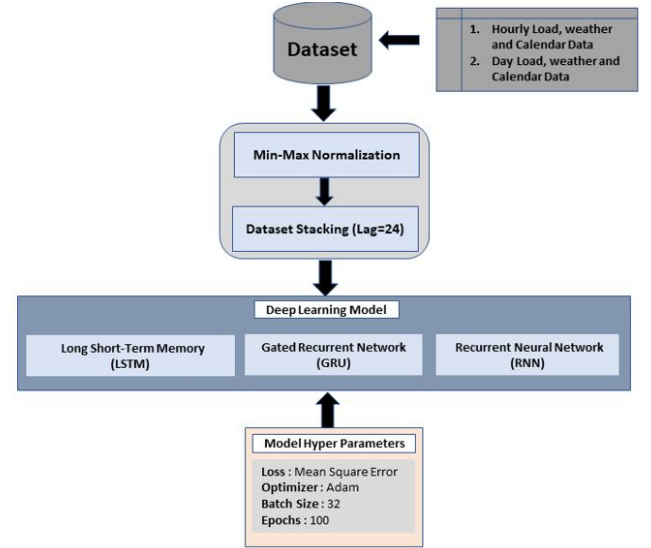


Fig. 3. DL Models used for STLTF.

and effective energy management.

The dataset is pre-processed by performing the following steps:

1) *Normalization*:: The Load and weekday columns are normalized by performing the standard normalization which is shown below.

$$X_{Norm} = \frac{X_i - X_{min}}{X_{max} - X_{min}} \quad (5)$$

As seen in the Pattern of Dataset section, the dataset is in the expected pattern, so the day name has been extracted, and the same is integrated into the dataset so that the model is aware of the seasonality present in the dataset explicitly.

2) *Lag*: Since the dataset is available in hourly format, the dataset is formatted in such a way that the model analyzes the past 24 hours' Load value to predict the expected Load in the Delhi Region.

$$Load_{X_{n+1}} = Model([X_n, X_{n-1}, X_{n-2}, X_{n-3}, \dots, X_{n-19}]) \quad (6)$$

3) *Exploratory Data Analysis*: A visual plot is drawn for each day of the week, i.e. from Sunday to Saturday, to show the load consumption analysis for that day. The usage of electricity rises gradually from 5.00 AM to 3.00 PM across all seven days of the week, starting on Sunday and ending on Saturday. After reaching its peak at 3.00 PM, electricity consumption starts to decline gradually from 7 PM until 12 PM. The plots for only the first and last days of the week are shown in Fig. 6.

### B. Deep Learning Model Configuration

This section discusses the setup and outcomes of experiments involving deep learning models such as LSTM, GRU, and RNN. The dataset distribution for 1-hour load forecasting and day load forecasting is processed with a 24-hour lag. The training, validation, and test sets are divided into an 80:10:10 ratio. The following 'HyperParameter' values are used to control the three deep learning models: (1. 'Optimizer: Adam', 2. 'Loss: Mean Squared Error', 3. 'Batch Size: 32' and 'Epochs: 100'). Two metrics, 'loss and accuracy', have been considered for the validation of the model. To verify the overfitting, the dataset

TABLE III  
ENSEMBLE MODEL AND PROPHET ALGORITHM VALIDATION METRICS

Model	MSE	RMSE	EV Score	$R^2$ Score	Data Set
XG Boost	184579.72	429.63	88.86	84.50	DL+W+C
Gradient T Boosting	183584.47	428.47	89.09	84.59	DL+W+C
Ada Boost	202500.90	450.00	86.88	86.88	DL+W+C
Random Forest	187942.76	433.52	336.54	84.22	DL+W+C
Prophet	4331909.51	2081.32		0.7088	
DL-Day Load; W-Weather; C-Calendar					

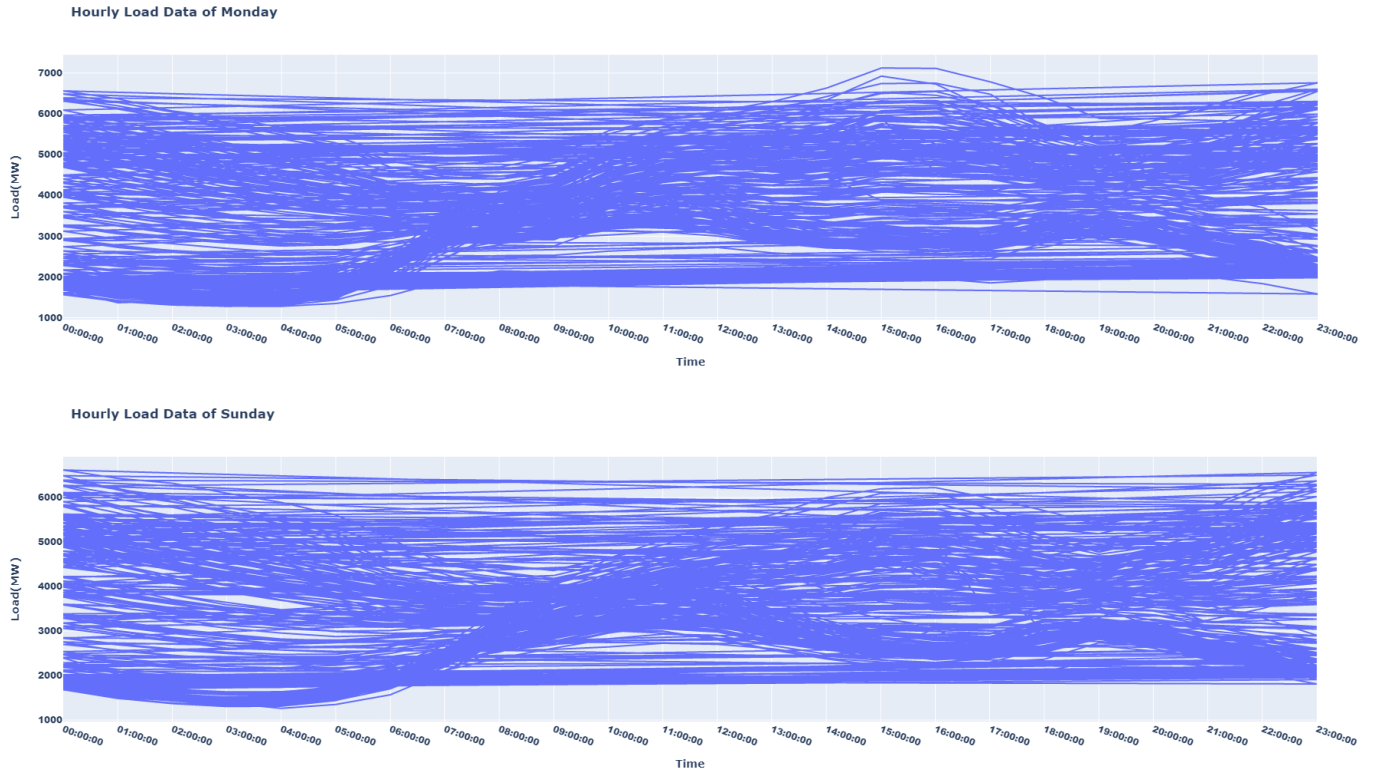


Fig. 4. Day Wise Load Consumption Analysis

is split into train, validation, and test. Similar performance on both the training and test sets is observed signifying a favorable outcome for the appropriate functioning of the model.

1) *Hourly Load Forecasting*: The dataset on hourly load is presented in Fig. 5, and it comes in three different forms: Load (Hourly) alone, Load (Hourly) with weather data, and Load (Hourly) with calendar information. These are utilized for forecasting, and for best performance in terms of training loss and accuracy, and testing loss and accuracy; three models, namely LSTM, GRU, and RNN, are trained for each of the three datasets.

2) *Daily Load Forecasting*: The dataset on daily load is presented in Fig. 4, and it comes in three different forms: Load (Daily) alone, Load (Daily) with weather data, and Load (Daily) with calendar information. They are utilized for forecasting, and using the results acquired, the best performance is plausible in terms of training loss vs. accuracy and testing loss vs. accuracy. To achieve this extent, three models, namely LSTM, GRU, and RNN, are trained for each of the three aforesaid datasets.

## VII. RESULT ANALYSIS

The efficacy of each type of model to forecast hourly load is assessed using performance metrics using the bar graph shown in Fig. 5. The train and test accuracy obtained is 100% for the 'hourly load with calendar data' for the LSTM and GRU models. The LSTM and GRU show the best accuracy of 100% for the 'hourly load data'. For the 'hourly load with weather data', LSTM results in 99% accuracy. Overall, the LSTM model is performing well on all three types of data sets. Only with the load data, it is possible to schedule a 24-hour load. The LSTM model is doing well with all three types of data sets. The findings indicate that scheduling a 24-hour electrical load using only load data is possible. Additionally, the study reveals that calendar data has a greater impact on hourly load forecasting compared to weather data.

Further, the performance metrics for the Daily Data Prediction are shown in Fig. 6. It is inferred from this bar graph that the train and test accuracy obtained is 99% for the 'daily load data' and 'daily load with the calendar data' using LSTM, RNN, and

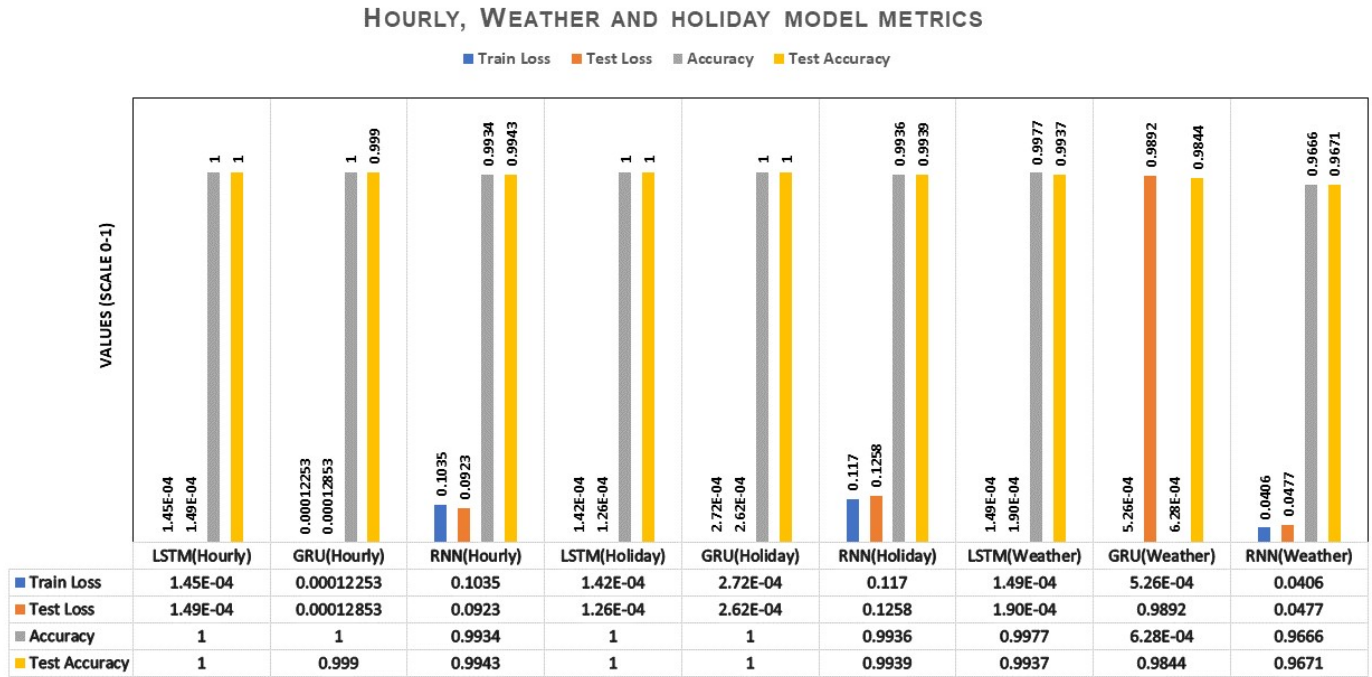


Fig. 5. Model Performance Metrics for the Hourly Data Prediction

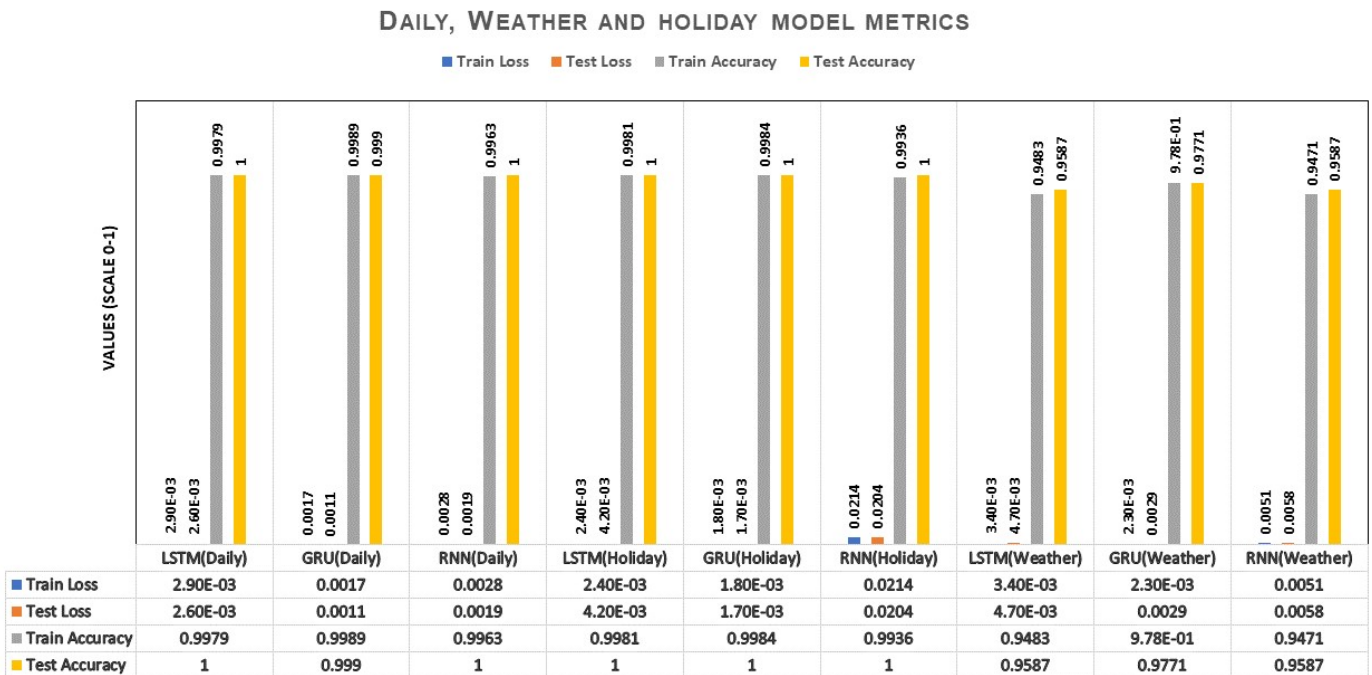


Fig. 6. Model Performance Metrics for the Daily Data Prediction



GRU models. The 'daily load with weather data' GRU model has resulted in 97% accuracy. Only with the load data, it is possible to schedule a 7-day load (from Sunday to Saturday). The findings indicate that scheduling a 7-day electrical load using only the load data is possible. Additionally, the study reveals that calendar data has a greater impact on daily load forecasting compared to weather data.

## VIII. CONCLUSION

The STLF for the Delhi region has been conducted successfully with different prediction models of machine learning, ensemble learning, and deep learning. Considering all these fluctuations, the focus of this study was to validate the model's stability in an unprecedented situation like the COVID-19 pandemic. The Delhi electricity load data for the years 2017, 2018, and 2020 was considered for forecasting. Other influential data, like weather and calendar, that are directly proportional to electrical load consumption, were also considered. We have carried out the experimentation in five different techniques namely: (i) Exploratory Analysis, (i) Univariate time series analysis (with 4 models), 3) Machine learning models (considering 24 scenarios: 8 data sets vs. 3 Models), 4) Ensemble machine learning models (considering 32 scenarios: 8 data sets vs. 4 Models), and 5) Deep learning models (considering 18 scenarios: 6 data sets vs. 3 Models). Among 14 different models, both the LSTM and GRU models have given hourly and daily load predictions with 100% accuracy. Both of these models have exhibited exceptional performance in predicting sequences in time series data considering calendar data as an influencer over weather data. It is concluded from the result analysis that the offered approaches have the potential to anticipate the 24-hour load demand and day-by-day load demand for a region with large fluctuations in weather and load patterns during normal as well as pandemic periods. By applying these forecasting methodologies to MTLF and LTLF, utilities and distribution companies can better anticipate future load requirements and plan their operations accordingly. This proactive approach is effective in managing power distribution efficiently, ensuring timely provision of electricity, and mitigating the risk of peak overload situations. Additionally, it enables utilities to optimize resource allocation, identify potential bottlenecks, and make informed decisions about infrastructure upgrades and investments.

## ACKNOWLEDGMENT

The authors acknowledge the support from the Regional Meteorological Centre, New Delhi, India, for providing the Weather data and Delhi TRANSCO Limited, for providing the electricity data for carrying out this study.

## REFERENCES

- [1] Z. Hamad, I. Abdul rahman, "Deep learning-based load forecasting considering data reshaping using MATLAB Simulink", *Int J Energy Environ Eng*, vol. 13, pp. 853-869, 2022, doi.org/10.1007/s40095-022-00480-x
- [2] N. Jha, D. Prashar, M. Rashid, S. K. Gupta, and R. K. Saket, "Electricity load forecasting and feature extraction in smart grid using neural networks", *Computers & Electrical Engineering*, vol. 96, pp. 107479, 2021.
- [3] R. K. Sethia, S. U. Lalit and A. Siddique, "Power Exchanges in India- Overview and Way Forward", *India Journal of Project Infrastructure & Energy Law*, 2021. <https://ijpiel.com/index.php/2021/12/22/power-exchanges-in-india-overview-and-way-forward/>

- [4] T. Hong, P. Pinson, Y. Wang, R. Weron, D. Yang, H. Zareipour, "Energy forecasting: a review and outlook", *IEEE Open Access J. Power Energy*, vol. 7, pp. 376-388, 2020. doi.org/10.1109/OAJPE.2020.3029979
- [5] N. Vanting, Z. Ma, B. Jorgensen, "A scoping review of deep neural networks for electric load forecasting", *Energy Information*, vol. 4, no. 49, 2021.
- [6] H. Acaroglu, F. Márquez, "Comprehensive review on electricity market price and load forecasting based on wind energy", *Energies*, vol. 14, pp. 7473, 2021. doi.org/10.3390/en14227473
- [7] S. Haben, S. Arora, G. Giasemidis, M. Voss, D. V. Greetham, "Review of low voltage load forecasting: Methods, applications, and recommendations", *Applied Energy*, vol. 304, pp. 117798, ISSN 0306-2619, 2021. doi.org/10.1016/j.apenergy.2021.117798.
- [8] I.K. Nti, M. Teimeh, O. Nyarko-Boateng, "Electricity load forecasting: a systematic review", *Journal of Electrical Systems and Inf Technol*, vol. 7, no. 13, 2020. <https://doi.org/10.1186/s43067-020-00021-8>
- [9] N.S.M. Salleh, A. Suliman, & B. N. Jorgensen, "A Systematic Literature Review of Electricity Load Forecasting using Long Short-Term Memory", *In Proceedings of the 8th International Conference on Computational Science and Technology*, pp. 765-776, Springer, Singapore, 2022.
- [10] A. Azeem, I. Ismail, S.M. Jameel, V.R. Harindran, "Electrical load forecasting models for different generation modalities: a review", *IEEE Access*, vol. 9, pp. 142239-142263, 2021. doi.org/10.1109/ACCESS.2021.3120731
- [11] A.A. Mamun, M. Sohel, N. Mohammad, M.S.H. Sunny, D.R. Dipta, E. Hossain, "A comprehensive review of the load forecasting techniques using single and hybrid predictive models", *IEEE Access*, vol. 8, pp. 134911-134939, 2020. doi.org/10.1109/ACCESS.2020.3010702
- [12] J. Li, et al, "A survey on investment demand assessment models for power grid infrastructure", *IEEE Access*, vol. 9, pp. 9048-9054, 2021.
- [13] I. Yazici, O. F. Beyca, & D. Delen, "Deep-learning-based short-term electricity load forecasting: A real case application", *Engineering Applications of Artificial Intelligence*, vol. 109, pp. 104645, 2022. doi.org/10.1109/ACCESS.2021.3049601
- [14] B. Farsi, M. Amayri, N. Bouguila, and U. Eicker, "On short-term load forecasting using machine learning techniques and a novel parallel deep LSTM-CNN approach", *IEEE Access*, vol. 9, pp. 31191-31212, 2021.
- [15] D. Solyali, "A comparative analysis of machine learning approaches for short-/long-term electricity load forecasting in Cyprus", *Sustainability*, vol.12, no. 9, pp. 3612, 2020.
- [16] M. Lopez-Martin, A. Sanchez-Esguevillas, L. Hernandez-Callejo, J. I. Arribas, & B. Carro, "Novel data-driven models applied to short-term electric load forecasting", *Applied Sciences*, vol. 11, no. 12, pp. 5708, 2021.
- [17] S. Papadopoulos and I. Karakatsani, "Short-term electricity load forecasting using time series and ensemble learning methods", *IEEE Power and Energy Conference at Illinois*, pp. 1-6, 2015.
- [18] A. S. Khwaja, A. Anpalagan, M. Naeem, & B. Venkatesh, "Joint bagged-boosted artificial neural networks: Using ensemble machine learning to improve short-term electricity load forecasting", *Electric Power Systems Research*, 179, 106080, 2020.
- [19] F. M. Butt, L. Hussain, S. H. M. Jafri, H. M. Alshahrani, F. N. Al-Wesabi, K. J. Lone, E. M. T. El Din & M. A. Duhayyim, "Intelligence based Accurate Medium and Long Term Load Forecasting System", *Applied Artificial Intelligence*, vol. 36, no.1, 2022. DOI: 10.1080/08839514.2022.2088452
- [20] N. Son, "Comparison of the deep learning performance for short-term power load forecasting", *Sustainability*, vol. 13, no. 22, pp. 12493, 2021.
- [21] M. Alhussein, K. Aurangzeb, & S. I. Haider, "Hybrid CNN-LSTM model for short-term individual household load forecasting", *IEEE Access*, vol. 8, pp. 180544-180557, 2020.
- [22] N. Son, S. Yang, & J. Na, "Deep neural network and long short-term memory for electric power load forecasting". *Applied Sciences*, vol. 10, no. 18, pp. 6489, 2020
- [23] M. Zhang, Z. Yu, & Z. Xu, Short-term load forecasting using recurrent neural networks with input attention mechanism and hidden connection mechanism, *IEEE Access*, vol. 8, pp. 186514-186529, 2020.
- [24] A. Haque, & S. Rahman, "Short-term electrical load forecasting through the heuristic configuration of regularized deep neural network", *Applied Soft Computing*, vol. 122, pp. 108877, 2022.
- [25] Y. Shen, Y. Ma, S. Deng, C. J. Huang, & P. H. Kuo, "An ensemble model based on deep learning and data pre-processing for short-term electrical load forecasting", *Sustainability*, vol. 13, no. 4, pp. 1694, 2021.
- [26] P. Liu, P. Zheng, & Z. Chen, "Deep learning with stacked denoising auto-encoder for short-term electric load forecasting", *Energies*, vol. 12, no. 12, pp. 2445, 2019.
- [27] S. H. Rafi, R. S. Deeba, & E. Hossain, "A short-term load forecasting method using integrated CNN and LSTM network", *IEEE Access*, vol. 9, pp. 32436-32448, 2021.
- [28] X. Tang, Y. Dai, T. Wang, & Chen, Y. (2019). "Short-term power load forecasting based on a multi-layer bidirectional recurrent neural network", *IET Generation, Transmission & Distribution*, vol. 13, no. 17, pp. 3847-3854, 2019.

- [29] M. Massaoudi, S. S. Refaat, I. Chihi, M. Trabelsi, H. Abu-Rub, & F. S. Oueslati, "Short-term electric load forecasting based on data-driven deep learning techniques", In *IECON 2020 The 46th Annual Conference of the IEEE Industrial Electronics Society*, pp. 2565-2570, 2020.
- [30] A. Yang, W. Li, & X. Yang, "Short-term electricity load forecasting based on feature selection and Least Squares Support Vector Machines, *Knowledge-Based Systems*, vol/ 163, pp. 159-173, 2019.
- [31] S. Bouktif, A. Fiaz, A. Ouni, and M. A. Serhani, "Single and multi-sequence deep learning models for short and medium-term electric load forecasting, *Energies*, vol. 12, no. 1, pp. 149, 2019.
- [32] V. Veeramsetty, K. R. Reddy, M. Santhosh, A. Mohnot, A., & G. Singal, "Short-term electric power load forecasting using random forest and gated recurrent unit", *Electrical Engineering*, vol. 104, no. 1, p. 307-329, 2022.
- [33] M. F. Alsharekh, S. Habib, D. A. Dewi, W. Albatat, M. Islam and S. Albahli, "Improving the Efficiency of Multi-step Short-Term Electricity Load Forecasting via R-CNN with ML-LSTM", *Sensors*, vol. 22, no. 18, pp. 6913, 2022.
- [34] V. Veeramsetty, D. R. Chandra, F. Grimaccia, & M. Mussetta, "Short-term electric power load forecasting using principal component analysis and recurrent neural networks", *Forecasting*, vol. 4, no. 1, pp. 149-164, 2022.
- [35] L. D. Soares, & Franco, E. M. C. (2021), "BiGRU-CNN neural network applied to short-term electric load forecasting", *Production*, vol. 32, 2021.
- [36] E. Lee, & W. Rhee, "Individualized short-term electric load forecasting with deep neural network based transfer learning and meta-learning", *IEEE Access*, vol. 9, pp. 15413-15425, 2021.



**Lakshmi D** is a Senior Associate Professor in the School of Computing Science and Engineering (SCSE) and Assistant Director at the Centre for Innovation in Teaching and Learning (CITL) at VIT Bhopal. She has 26 years of teaching experience and has delivered numerous guest lectures, chaired sessions, and keynote international conferences. She led a 200-person FDP and, in the research frontier, has made 25 international conference presentations, 30 SCOPUS and SCI-indexed journal papers, 8 SCOPUS-indexed chapters, and 24 patents worldwide: two book publications, Best Research Paper awards, and Best Teacher honors.



**Ravi Shekhar Tiwari** is a researcher, innovator, and an engineer. He has 4+ years of industry experience working as an Artificial Intelligence Engineer, Penetration Tester, and MFDI Engineer in Multinational IT companies as well as start-ups. He also holds a position as a reviewer and editor in reputed journals and as an author in technical magazines. His research domain includes Time Series Analysis, Protein Structure Prediction and Generation, Federated Learning, the Internet of Things, Microcontrollers, Gait Analysis, AI and Healthcare, XAI, Cloud Computing, Computer Vision, Parallel and

Distributed Computing in the cloud. Currently, he is pursuing his Master in Technology in Mahindra University with Specialization Artificial Intelligence and Data Science



**NEELU NAGPAL** (Senior Member, IEEE) is currently an Associate Professor in GGSIP University, has been working since last 17 years in EEE Department of Maharaja Agrasen Institute of Technology, Delhi. Her Ph.D in Electrical Engineering was accomplished in Delhi Technological University, Delhi, India. She was the recipient of commendable research award from Delhi Technological University during her PhD course. She completed her Masters with distinction from Delhi University in Control and Instrumentation specialization. Prior to that, she graduated in Electrical Engineering from Delhi College of Engineering. Her research interest includes Stochastic and nonlinear control, state estimation, smart grid technologies, renewable energy integration and artificial intelligence. She has grant of one Australian patent. She is Vice-chair, IEEE Smart Cities Ambassador Program 2023.



**NEELAM KASSARWANI** is currently an Associate Professor in GGSIP University, has been working since last 15 years in EEE Department of Maharaja Agrasen Institute of Technology, Delhi. She received her doctorate in electrical engineering from the National Institute of Technology, Kurukshetra, India. She completed her Masters in Power System from Delhi University in Power System and Apparatus specialization. Prior to that, she graduated in Electrical Engineering in first division from Madan Mohan Malviya Engineering College, Gorakhpur, India. She was a National Scholarship Holder from year 1980-87 for her education. Besides having 20 years of experience in teaching, she has 10 years of industrial experience. She has 8 research publications in International journals, and conferences. Her area of research is investigations on the Power System modelling and control, power quality, dynamic voltage restore, artificial intelligence, and renewable energy systems.



**G. Vishnuvarthanan** is an Associate Professor in VIT-Bhopal's School of Computing Science and Education's Data Science Division. His PhD in Electronics and Communication Engineering was in 2015 for "Tumor Detection and Tissue Segmentation in Multimodal MR Brain Images Using Fuzzy and Optimization Techniques". He spent 14 years teaching at three top Tamil Nadu engineering schools. The greatest impact value of his 100 journal publications in 33 Science Citation Index Database journals is 17.560, and the average impact factor is 4.923. IEEE Transactions on Industrial Informatics (11.648), IEEE Transactions on Fuzzy Systems (12.253), and Elsevier Information Fusion are notable publications. OPTIK, Computers in Biology and Medicine, IEEE Transactions on Cognitive and Developmental Systems, and Applied Soft Computing have nominated him for review. He appreciates medical image processing, AI, ML, and pattern recognition.



**Abhishek Srivastava** is a dynamic Software Engineer at JP Morgan Chase & Co, possessing an exceptional proficiency in Problem Solving, System Architecture, and a comprehensive grasp of Data Structures and Algorithms. Holding esteemed roles as a Microsoft Student Ambassador and a distinguished leader of Google Developer Student Club, he has played an instrumental role in steering global coding endeavours such as GirlScript Summer of Code and Student Code-in, instrumental in equipping emerging talents with proficient programming skills. His professional journey encompasses impactful stints at Songdew, SurveySparrow, and BeofUse, where he has adeptly engineered websites for diverse startups and enterprises. He is a prolific tech blogger and devoted open-source advocate. His insightful compositions are prominently featured on Medium (@abhishek2x).

# Comparative Analysis of ResNet and DenseNet for Differential Cryptanalysis of SPECK 32/64 Lightweight Block Cipher

Ayan Sajwan<sup>1</sup> and Girish Mishra<sup>2</sup>

<sup>1</sup> Delhi Technological University, Delhi - 110 042, India  
ayansajwan2003@gmail.com,

<sup>2</sup> DRDO-Scientific Analysis Group, Delhi - 110 054, India  
gmishratech28@gmail.com

**Abstract.** This research paper explores the vulnerabilities of the lightweight block cipher SPECK 32/64 through the application of differential analysis and deep learning techniques. The primary objectives of the study are to investigate the cipher's weaknesses and to compare the effectiveness of ResNet as used by Aron Gohr at Crypto2019 and DenseNet. The methodology involves conducting an analysis of differential characteristics to identify potential weaknesses in the cipher's structure. Experimental results and analysis demonstrate the efficacy of both approaches in compromising the security of SPECK 32/64.

**Keywords:** Differential Cryptanalysis, Deep Learning, Speck, ResNet, DenseNet

## 1 Introduction

Cryptography is the technique of converting data into an incomprehensible form known as cipher text. It is done by using mathematical principles and algorithms. This crucial sector ensures the security and privacy of modern digital communications and data storage. Throughout history, critical information ranging from military communications<sup>[1]</sup> to commercial transactions<sup>[2]</sup> have been protected via cryptographic processes.

Over the centuries, cryptography has been an art practised by many who have invented techniques to meet some of the information security requirements. The previous two decades have seen the field evolve from an art to a science<sup>[3]</sup>.

Data secrecy, integrity and authenticity are the main goals of cryptography. Confidentiality ensures that only authorised personnel may access the information. Integrity ensures that the data is unchanged throughout transmission or storage. Authentication makes sure that only reliable sources are sharing information

With the aid of encryption algorithms<sup>[4]</sup>, cryptography secures data and communication. This is achieved by converting plaintext into ciphertext, a form



that cannot be deciphered. On the other side, Cryptanalysis is the science of analysing cryptographic systems to find vulnerabilities. These weaknesses can be exploited to obtain the original plaintext or encryption keys. Lightweight block ciphers serve a critical role in cryptography in situations where computational resources are constrained. These ciphers' low computational and memory overhead makes them ideal for secure, effective encryption<sup>[5]</sup>.

The study of decrypting cryptographic methods, or cryptanalysis, is a crucial area for maintaining the security of encryption systems. It involves investigating the mathematical features and design choices to find the flaws in the cipher. Understanding these flaws allows cryptanalysts to create more successful attacks. This helps in increasing the security of the cryptographic systems.

Numerous fields, including cryptanalysis<sup>[6]</sup>, have seen the emergence of machine learning and deep learning as highly effective tools. These methods make use of the models' computational capabilities. They can automatically identify patterns, detect features, and generate predictions. Machine learning techniques can be used in the context of cryptanalysis as well<sup>[7]</sup>. It can be used to analyse and categorise cryptographic data, such as ciphertexts, plaintexts, or encryption keys.

Deep learning, a branch of machine learning, has achieved outstanding results in a number of fields. Ranging from speech recognition<sup>[8]</sup>, computer vision<sup>[9]</sup> to natural language processing<sup>[10]</sup>. Deep neural networks are able to learn complex data representations and identify deep correlations. In this paper, we investigate the use of deep learning methods, more specifically DenseNet and ResNet. They are used for differential cryptanalysis on the lightweight block cipher SPECK 32/64. We intend to compare the effectiveness of ResNet and DenseNet. It ultimately helps us in understanding of the security of the SPECK 32/64 cipher by utilising the expressive potential of deep neural networks.

## 2 SPECK 32/64 Overview

SPECK is a lightweight block cipher developed by the National Security Agency (NSA). It was a part of the Lightweight Cryptography Project of the NSA. Here, the term "lightweight" refers to cryptographic algorithms that are designed for efficient operation and low resource consumption.

This qualifies them for usage in environments with limited resources such as Internet of Things devices<sup>[11]</sup>, wireless sensor networks<sup>[12]</sup>, and embedded systems<sup>[13]</sup>.

The SPECK family consists of a variety of block and key sizes. The block is made up of 2 words, and is of the form  $2n$ . Here,  $n$  is the size of the word which may be 16, 24, 32, 48 or 64 bits. The key size( $k$ ) is  $mn$  bits. The key contains 2, 3 or 4 words depending on the variant. Hence, the SPECK family is of the form

*SPECK*  $2n/mn$  and has ten variants. We use SPECK 32/64 in our work, which denotes 2 words of 16-bits each and 4 keys of 16-bits each as well.

The SPECK family cipher is made up of a Feistel network structure<sup>[14]</sup>. In this network, the input block is split into two equally sized halves. Encryption rounds are then performed on these halves using a subkey. A different subkey is derived for each round. The number of rounds differ in each variant of the family.

The round function of SPECK 32/64 uses a range of bitwise operations for its cryptographic operations. It includes rotation, XOR, and modular addition, to induce confusion and diffusion features, assuring the security of the cipher.

## 2.1 Round Function

Speck's round function is very simple. It is an ARX structure, which means it is made out of the fundamental functions of modular addition ( $\text{mod } 2^k$ ), bitwise rotation, and bitwise addition. They are denoted by  $\boxplus$ ,  $\gg$  and  $\oplus$  respectively. SPECK  $n/m$  represents Speck with  $n$  bit block size and  $m$  bit key size. It produces the next round state  $(\mathbf{L}_{i+1}, \mathbf{R}_{i+1})$  with an input  $k$ -bit subkey  $\mathbf{K}$  and the current cipher state consisting of two  $k$ -bit words  $(\mathbf{L}_i, \mathbf{R}_i)$ . The algorithm is as follows:

$$\begin{aligned}\mathbf{L}_{i+1} &= ((\mathbf{L}_i \gg \alpha) \boxplus \mathbf{R}_i) \oplus \mathbf{K} \\ \mathbf{R}_{i+1} &= (\mathbf{R}_i \ll \beta) \oplus \mathbf{L}_{i+1}\end{aligned}$$

The values of  $\alpha$  and  $\beta$  are constant:  $(\alpha = 7, \beta = 2)$  for Speck32/64 and  $(\alpha = 8, \beta = 3)$  for other members of the Speck family. The cipher text output is generated by applying the round function on the plain text input for 22 rounds in the case of Speck 32/64. However, we refer to round reduced speck in this paper. The key used in each round is generated from a master key by applying a key schedule. The key schedule depends on the member of the Speck family, we refer to Beaulieu et al<sup>[15]</sup> in this paper for the key scheduling.

## 3 ResNet and DenseNet Architectures

### 3.1 ResNet

ResNet or Residual Network is a powerful deep learning architecture first published by Kaiming<sup>[16]</sup>. Resnet is intended to address the issue of disappearing gradients<sup>[17]</sup> in very deep neural networks. It does so, by incorporating residual connections<sup>[18]</sup> or skip connections, which enable the building of deeper and more precise models.

ResNet is composed of many residual blocks or towers that are layered on top of each other and contain a sequence of convolutions and a skip connection. The skip connection is added to a block's output and then passed on to the following block. This helps in reducing the vanishing gradient problem and allows for better model training. Figure 2 depicts the working of a skip connection.

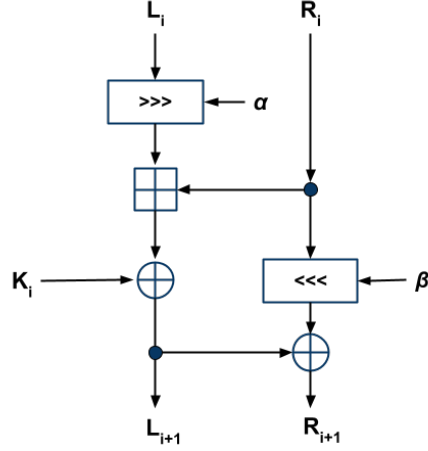


Fig. 1: General round of Speck

**ResNet Structure** The ResNet structure used in this work is the one used by Gohr in 2019<sup>[19]</sup>. It consists of a residual tower of depth ten, having a two layer convolutional network. The convolutional network has 32 filters.

First, Convolution is applied followed by a Batch normalization<sup>[20]</sup> for faster and stable training . It is followed by a Rectified Linear Unit layer<sup>[21]</sup> which introduces Non-linearity to the model. Then a skip/jump connection at the end adds the output of the final rectifier layer of the block to the convolutional block's input and forwards the result to the next block.

The initial layer is a bit-sliced 1 Dimensional Convolution with 32 output channels, which is followed by Batch normalization. Finally, a Rectified Linear Unit is applied to the preceding layer's output. The final result is a  $32 \times 16$  matrix that is fed into the depth-10 Residual Tower.

Finally, the data is flattened and transmitted to the prediction layer. This final layer consists of two densely linked hidden layers of 64 units each, followed by batch normalization, a Rectified Linear unit, and sigmoid activation for a single output head.

### 3.2 DenseNet

DenseNet is made up of Dense blocks and transition layers. DenseNet, which stands for "Densely Connected Networks" is a deep learning architecture designed by Gao Huang et al. originally published in their paper<sup>[22]</sup> in 2017.

DenseNet, like ResNet, aims to solve the vanishing gradient problem by maximising feature reuse. DenseNet introduces dense connections between layers and

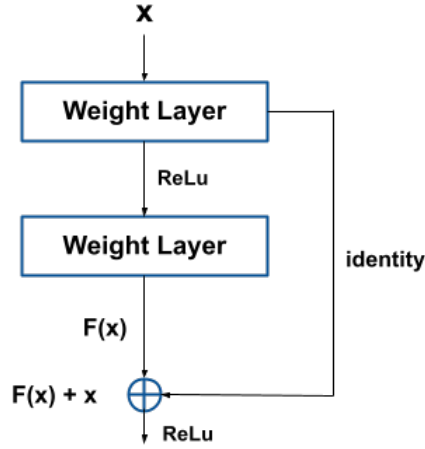


Fig. 2: Skip Connection

blocks as opposed to traditional neural networks, which connect layers sequentially. Unlike ResNet, which utilises an additive approach of adding previous layer output to subsequent layers, DenseNet uses all past outputs as input for future layers. As a result, each layer is directly linked to all the following layers.

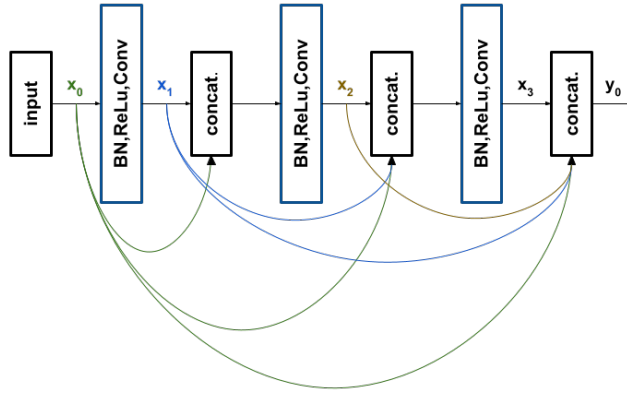


Fig. 3: Dense Connection

**DenseNet Structure** The first layer, like the one in the Resnet model, is a bit-sliced 1 Dimensional Convolution with 32 output channels. It is followed by

Batch normalisation, and lastly a Rectified Linear Unit is applied to the output of the preceding layer. The resulting  $32 \times 16$  matrix is passed into the Dense Network.

DenseNet is made up of Dense blocks and transition blocks. The dense block has a depth-8 and is made up of two layers of 1D-convolution, Batch normalisation, and a Rectified linear unit layer. The convolution layer is made up of 64 filters and a kernel with a size of 3. Finally, the output is concatenated with the layer's input and handed on to the next layer. This occurs eight times since the depth is eight.

To restrict the amount of feature maps and minimise spatial dimensions, transition layers are inserted between dense blocks. The transition layer is made up of 1D convolution with 32 filters, batch normalisation, a ReLu layer, and 1D average pooling. The transition layer's output is subsequently passed on to the following dense block.

The dense block and transition layers are now merged with a depth of 2. This indicates alternating dense block and transition layer, followed by a final dense block. The final result is an overall structure of three dense blocks and two transition layers.

Finally, the data is flattened and transmitted to the prediction layer. This prediction layer consists of two dense hidden layers of 64 units each, followed by batch normalisation, Rectified linear unit, and sigmoid activation, similar to ResNet.

### 3.3 Input Data

Input data: Input consists of a pair of cipher texts ( $C_0, C_1$ ). They are transformed into a  $4 \times 16$  matrix with each row consisting of a word of the ciphertext. This way the data consists of four 16-bit words and therefore the input layer has 64 units. This input data is then passed into the ResNet and DenseNet architecture.

## 4 Experimental Setup and Methodology

### 4.1 Data Generation

The data generation methodology used is similar to the one used by Aron Gohr in 2019. A random number generator is used to create evenly distributed keys  $K_i$  and plain text pairings  $P_i$  with the input difference  $\Delta = 0x0040/0000$ , along with a vector of binary-valued real/random labels  $Y_i$ . If  $Y_i$  is set (=1), the plain text pair  $P_i$  is encrypted for  $k$  rounds to create training or validation data for  $k$ -round Speck, and if not, the second plain text in the pair is changed to a newly created random plain text. This way we have cipher texts belonging to 2 classes: Chosen Input difference (  $\mathbf{Y} = 1$  ) and random input difference (  $\mathbf{Y} = 0$  ). As a result we have  $10^6$  samples for our dataset for training and validation.

#### 4.2 Training and Testing Procedure:

The data set of  $10^6$  samples is used for training in batches of 5000 and run for 20 epochs as opposed to 200 epochs by Gohr . Mean Square Error (MSE)<sup>[23]</sup> loss is used with L2 weights regularization using the Adam algorithm<sup>[24]</sup> for optimization. This is a down scaled version of Gohr’s experiment in which he used  $10^7$  samples for training for 200 epochs.

Testing data also contains a set of  $10^6$  samples with 2 classes. One of the chosen input difference and the other of random input difference. Table 1. provides the list of hyper-paramaters used in training with their values.

Hyper-parameters	values
Sample Size	$10^6$
Batch Size	5000
Epochs	20
Encryption Rounds	5,6,7,8
Optimizer	Adam
Loss function	MSE loss
Cyclic Learning Rate	0.002-0.0001

Table 1: Hyper-parameters for training of model

### 5 Results and Analysis:

In this section, we present the findings of our experiments on the differential cryptanalysis of the round reduced ( rounds 5,6,7,8 ) SPECK 32/64 lightweight block cipher using the ResNet and DenseNet architectures. R5, R6, R7 & R8 refers to the ResNet architecture for rounds 5, 6, 7 and 8 respectively, and similarly D5, D6, D7 & D8 refers to the DenseNet architecture. Table 2 depicts the training and validation accuracy for both the models.

Rounds	R (ResNet)		D (DenseNet)	
	Training	Validation	Training	Validation
5	0.9332	0.6779	0.9309	<b>0.7005</b>
6	0.7952	0.5874	0.7917	<b>0.5923</b>
7	0.6096	0.5267	0.6053	<b>0.5313</b>
8	0.5012	0.4996	0.4998	0.5002

Table 2: Training and Validation accuracy for ResNet and DenseNet models

The DenseNet model achieved slightly better validation accuracy than the ResNet model for rounds 5, 6 and 7. For round 8, both the models failed to give a prominent result since the models could not learn an accurate pattern. Figure 4 below shows the comparison of accuracy for both the models.

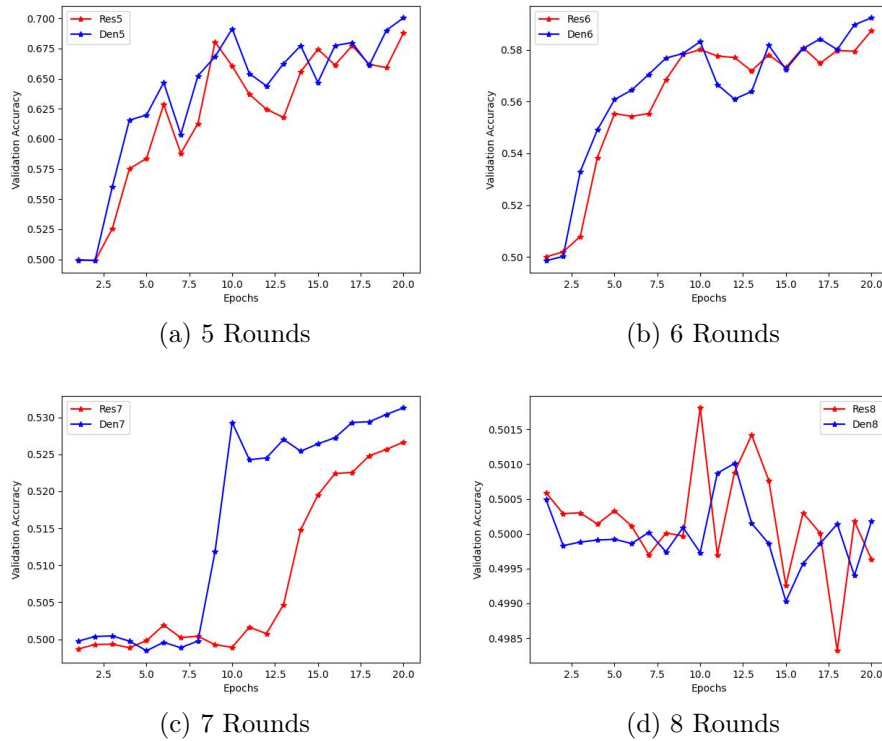


Fig. 4: Validation Accuracy comparison

## 6 Conclusions

In this paper, we compared the ResNet and DenseNet architectures for differential cryptanalysis of the SPECK 32/64 lightweight block cipher. Our analysis attempted to evaluate their accuracy in deciphering the cipher's complicated differential behaviour.

According to our findings, the DenseNet architecture outperforms the ResNet architecture marginally. DenseNet achieved slightly higher predictions of cipher-text differences in the context of differential cryptanalysis. However, neither model produced a satisfactory result for an 8-round (or higher rounds) encryption cipher.

As of now, this work does not include a key retrieval approach. At last, our findings highlight the importance of architecture selection in differential crypt-



analysis. The observed accuracy improvements of DenseNet support its use in scenarios requiring lightweight block cipher analysis.

## Acknowledgements

The Authors would like to show their sincere gratitude to the Scientific Analysis group (SAG), Defense Research and Development Organization (DRDO) for their invaluable support and collaboration throughout the course of this research. The authors would also like to thank Delhi Technological University (DTU), India for providing the opportunity to work in the field

## References

1. Doukas, Nikolaos, and Nikolaos V. Karadimas. "A blind source separation based cryptography scheme for mobile military communication applications." *WSEAS Trans. Commun* 7.12 (2008): 1235-1245.
2. Lamprecht, C., et al. "Investigating the efficiency of cryptographic algorithms in online transactions." *International Journal of Simulation: Systems, Science & Technology* 7.2 (2006): 63-75.
3. Menezes, Alfred J., Paul C. Van Oorschot, and Scott A. Vanstone. *Handbook of applied cryptography*. CRC press, 2018.
4. Mahajan, Perna, and Abhishek Sachdeva. "A study of encryption algorithms AES, DES and RSA for security." *Global Journal of Computer Science and Technology* 13.15 (2013): 15-22.
5. Hatzivasilis, G., Fysarakis, K., Papaefstathiou, I. et al. A review of lightweight block ciphers. *J Cryptogr Eng* 8, 141–184 (2018). <https://doi.org/10.1007/s13389-017-0160-y>
6. C. de Canniere, A. Biryukov and B. Preneel, "An introduction to Block Cipher Cryptanalysis," in *Proceedings of the IEEE*, vol. 94, no. 2, pp. 346-356, Feb. 2006, doi: 10.1109/JPROC.2005.862300.
7. Benamira, A., Gerault, D., Peyrin, T., Tan, Q.Q. (2021). A Deeper Look at Machine Learning-Based Cryptanalysis. In: Canteaut, A., Standaert, FX. (eds) *Advances in Cryptology – EUROCRYPT 2021*. EUROCRYPT 2021. *Lecture Notes in Computer Science()*, vol 12696. Springer, Cham. [https://doi.org/10.1007/978-3-030-77870-5\\_28](https://doi.org/10.1007/978-3-030-77870-5_28)
8. L. Deng, G. Hinton and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: an overview," 2013 *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, 2013, pp. 8599-8603, doi:10.1109/ICASSP.2013.6639344.
9. Q. Wu, Y. Liu, Q. Li, S. Jin and F. Li, "The application of deep learning in computer vision," 2017 *Chinese Automation Congress (CAC)*, Jinan, China, 2017, pp. 6522-6527, doi: 10.1109/CAC.2017.8243952
10. D. W. Otter, J. R. Medina and J. K. Kalita, "A Survey of the Usages of Deep Learning for Natural Language Processing," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 604-624, Feb. 2021, doi: 10.1109/TNNLS.2020.2979670.
11. Dinu, D., Corre, Y.L., Khovratovich, D. et al. Triathlon of lightweight block ciphers for the Internet of things. *J Cryptogr Eng* 9, 283–302 (2019). <https://doi.org/10.1007/s13389-018-0193-x>

12. M. Cazorla, K. Marquet and M. Minier, "Survey and benchmark of lightweight block ciphers for wireless sensor networks," 2013 International Conference on Security and Cryptography (SECRYPT), Reykjavik, Iceland, 2013, pp. 1-6.
13. Manifavas, C., Hatzivasilis, G., Fysarakis, K., Rantos, K. (2014). Lightweight Cryptography for Embedded Systems – A Comparative Analysis. In: Garcia-Alfaro, J., Lioudakis, G., Cuppens-Boulahia, N., Foley, S., Fitzgerald, W. (eds) Data Privacy Management and Autonomous Spontaneous Security. DPM SETOP 2013 2013. Lecture Notes in Computer Science(), vol 8247. Springer, Berlin, Heidelberg. [https://doi.org/10.1007/978-3-642-54568-9\\_21](https://doi.org/10.1007/978-3-642-54568-9_21)
14. Nyberg, K. (1996). Generalized Feistel networks. In: Kim, K., Matsumoto, T. (eds) Advances in Cryptology — ASIACRYPT '96. ASIACRYPT 1996. Lecture Notes in Computer Science, vol 1163. Springer, Berlin, Heidelberg. <https://doi.org/10.1007/BFb0034838>
15. Beaulieu, Ray, et al. "The SIMON and SPECK families of lightweight block ciphers." *cryptology eprint archive* (2013).
16. He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
17. H. Li, J. Li, X. Guan, B. Liang, Y. Lai and X. Luo, "Research on Overfitting of Deep Learning," 2019 15th International Conference on Computational Intelligence and Security (CIS), Macao, China, 2019, pp. 78-81, doi: 10.1109/CIS.2019.00025.
18. Jastrzębski, S., Arpit, D., Ballas, N., Verma, V., Che, T., & Bengio, Y. (2017). Residual connections encourage iterative inference. *arXiv preprint arXiv:1710.04773*.
19. Gohr, Aron. "Improving attacks on round-reduced speck32/64 using deep learning." *Advances in Cryptology—CRYPTO 2019: 39th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 18–22, 2019, Proceedings, Part II* 39. Springer International Publishing, 2019.
20. Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." *International conference on machine learning*. pmlr, 2015.
21. Agarap, Abien Fred. "Deep learning using rectified linear units (relu)." *arXiv preprint arXiv:1803.08375* (2018).
22. Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
23. Toro-Vizcarrondo, Carlos, and T. Dudley Wallace. "A test of the mean square error criterion for restrictions in linear regression." *Journal of the American Statistical Association* 63.322 (1968): 558-572.
24. Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

RESEARCH ARTICLE | SEPTEMBER 05 2023

# Computational fluid dynamics (CFD) simulations of domestic hybrid solar dryer under varying mass flow rates



Mukul Sharma; Deepali Atheaya; Anil Kumar ✉; Pawan Mishra



AIP Conf. Proc. 2863, 020005 (2023)

<https://doi.org/10.1063/5.0155308>



CrossMark

## Articles You May Be Interested In

Performance analysis of a novel solar organic rankine cycle (ORC)

AIP Conf. Proc. (September 2023)

Recent developments and applications of different solar dryers for agricultural crops: A review

AIP Conf. Proc. (September 2023)

Develop pedestrian based TOD index to measure TOD-levels in brownfield areas of Noida

AIP Conference Proceedings (November 2022)

500 kHz or 8.5 GHz?  
And all the ranges in between.

Lock-in Amplifiers for your periodic signal measurements



Find out more



# Computational Fluid Dynamics (CFD) Simulations of Domestic Hybrid Solar Dryer Under Varying Mass Flow Rates

Mukul Sharma<sup>1, b)</sup>, Deepali Atheaya<sup>1, c)</sup>, Anil Kumar<sup>2, 3, a)</sup> and Pawan Mishra<sup>1, d)</sup>

<sup>1</sup>Department of Mechanical Engineering, Bennett University, Greater Noida 201310, India

<sup>2</sup>Department of Mechanical Engineering, Delhi Technological University, Delhi 110 042, India

<sup>3</sup>Centre for Energy and Environment, Delhi Technological University, Delhi 110 042, India

<sup>a)</sup> Corresponding author: [anilkumar76@dtu.ac.in](mailto:anilkumar76@dtu.ac.in)

<sup>b)</sup> [e20soe803@bennett.edu.in](mailto:e20soe803@bennett.edu.in)

<sup>c)</sup> [deepali.atheaya@bennett.edu.in](mailto:deepali.atheaya@bennett.edu.in)

<sup>d)</sup> [pawan.mishra@bennett.edu.in](mailto:pawan.mishra@bennett.edu.in)

**Abstract.** Computational fluid dynamics simulation is a necessary step that saves a huge amount of money and time for researchers. It validates the proposed design by showing the behavior of flow, temperature, wall fluxes, and pressure inside the design. Different solar dryer designs were validated using this technique. A domestic hybrid solar dryer was designed and simulated for Bennett University, Greater Noida, India. The dryer was designed to work sustainably under indirect and mixed-mode operation. In this piece of work, the domestic hybrid solar dryer was simulated under indirect mode operation in unloaded conditions. The domestic hybrid dryer was simulated to track working fluid (air) temperature and absorbed solar flux inside the solar dryer at six different mass flow rates in the range of 0.08125 kg/s to 1.5625 kg/s cases. The mass flow rate of 1.2 kg/s was found to be suitable for the domestic hybrid solar dryer. During all the simulated cases of mass flow rates, the total radiation solar flux at the copper absorber box was 884 W/m<sup>2</sup>. This confirms the optimum mass flow rate of designed parameter of domestic hybrid solar dryer. The temperature inside the drying cabinet was found to be 348 K which was appropriate to dry four different crops.

## INTRODUCTION

Food drying is an essential technique of food storage. While using this technique, food crops with low shelf life can be stored longer. The food is dried using conventional and solar drying [1–3]. The conventional drying technology converts electricity into heat energy, further utilized for the food drying operation [4]. This conversion, however, is neither efficient nor it is economical. The production of conventional electrical power is mainly from fossil fuels that result in the emission of greenhouse gases (GHG) in the environment [5–7]. Therefore, solar drying is preferred over conventional drying. Domestic users majorly use the open sun drying method but, this method is highly inefficient but also, it has several disadvantages such as deterioration of food by birds, insects, animals, rain, fungus, etc. [8, 9]. These issues can be resolved by drying the food in a controlled environment [1, 10, 11].

The controlled environment is provided by dedicated devices named solar dryers. The solar dryers are mainly of three types: direct, indirect, and mixed-mode dryers [12]. Several dryers were designed and fabricated for industrial purposes and domestic purposes [13]. But good quality dried product and high efficient dryer will be a complete package for the domestic users. Numerical design and simulation are essential steps for researchers to validate design.

This saves time and money for the researcher, which can be wasted during fabrication and testing of a non-feasible and inefficient design [1, 14, 15].

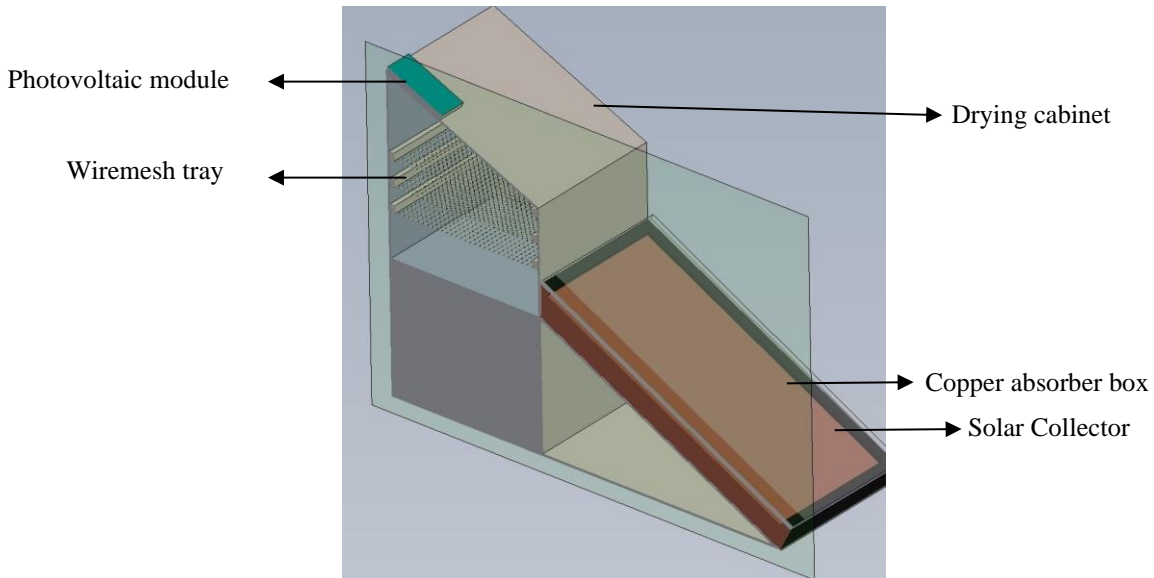
Several researchers prefer the CFD simulation of the dryer to simulate various parameters inside the solar dryers. Singh et al. [16] implemented CFD simulation over an indirect forced convection solar dryer to obtain thermal and dynamic performance at different mass flow rate. The simulated data was validated using experimental results. Mellalou et al. [17] constructed an uneven-span modified greenhouse dryer and interpreted the temperature distribution inside the dryer using CFD simulation. The validation of the simulation was performed using experimental results. Jain et al. [1] simulated a domestic multi-shelf solar dryer using ANSYS FLUENT software. Temperature distribution, the absorbed solar flux, and pressure distribution were reported. Sonthawi et al. [18] designed and simulated a solar biomass hybrid dryer. Temperature and airflow distributions were simulated using ANSYS-FLUENT CFD simulation software. Demissie et al. [19] designed an indirect solar food dryer and simulated the drying chamber's temperature distribution and three-dimensional flow field.

In the present work, the design of the proposed domestic hybrid solar dryer is validated by simulating the static temperature inside the dryer and total radiation heat flux at the solar collector. Furthermore, the static temperature inside the solar dryer design is simulated at different mass flow rates (0.08125 kg/s to 1.5625 kg/s), it helps to fix the system's mass flow rate to achieve appropriate drying temperature for drying food crops.

## METHODOLOGY

### System Information

A domestic hybrid solar dryer is designed for food drying at Bennett University (28.4506° N, 77.5842° E). The proposed domestic hybrid solar dryer comprises of a solar collector, copper absorber box, drying cabinet, three rectangular perforated wire mesh trays, photovoltaic (PV) module, and exhaust fan. The information about dimensions and materials of different components is provided in Table 1. Glass is placed at the top of solar collector and drying cabinet, as shown in Figure 1. The designed domestic hybrid solar dryer can work as an indirect and mixed-mode dryer as per the crop type. During indirect mode working, the glass top of the drying cabinet may be covered by using a suitable insulating material. This insulation will obstruct the approach of direct solar radiation inside the drying chamber.



**FIGURE 1.** Cross-sectional side view of designed domestic hybrid solar dryer.

During indirect forced mode operation, the solar radiation will fall over to the glass, and further, it will be transmitted and absorbed by the copper absorber box (placed inside the solar collector) via conduction, convection, and radiation heat transfer modes. The working fluid, i.e., ambient air, enters inside the solar collector from the inlet, and it will flow over the absorber box surface. The ambient air absorbs the thermal energy from the absorber box due to convection heat transfer and it travels upwards towards the inlet of the drying chamber. When this hot air comes into the contact of the food crop, it takes its moisture away and flows outside through the outlet port due to the force provided by the exhaust fan. This whole process will be continued until the food crop gets dried.

**TABLE 1.** Information about materials used and dimensions of different components of domestic hybrid solar dryer

Components of domestic hybrid solar dryer	Material used	Dimensions (mm)
Solar collector	Acrylic sheet (bottom and boundary), glass top	1100 × 620 × 40
Absorber box	Copper sheet	1000 × 500 × 20
Drying cabinet	Acrylic sheet, glass top	620 × 620 × 701
Wire mesh trays	Stainless steel	610 × 580
Exhaust fan at outlet	Composite plastic resins	80 × 80
Opaque PV module	Aluminum frame, silicon wafers, and glass	25 × 20

## Simulation approach

CFD simulation on designed domestic hybrid solar dryer was performed using ANSYS FLUENT software. For the simulation, various boundary conditions were followed, and the results were calculated by solving the governing equations using the ANSYS FLUENT software.

### *Boundary conditions*

The following boundary conditions were considered for numerical simulation of domestic hybrid solar dryer under indirect mode operation:

- The Initial temperature of the working fluid (air) was 300 K.
- Problem was considered as 3D and steady-state.
- The simulation was performed for different mass flow rates i.e. 0.08125 kg/s, 0.3 kg/s, 0.625 kg/s, 0.9375 kg/s, 1.2 kg/s and 1.5625 kg/s.

- For all circumstances, the Reynolds number stood less than  $5 \times 10^5$ ; therefore, flow exhibited a laminar pattern in the solar dryer.
- Dryer wall was considered motionless furnished with insulation at the outer wall.
- All surfaces in the design were assumed smooth, and flow working fluid was taken frictionless.
- All system parts were taken for the meshing procedure to obtain good results from the CFD analysis.
- Solar load model was chosen to evaluate solar insolation's effects entering into the computational domain.
- Number of iterations for the simulation was set at 3000. Higher iterations were set during steady state solution to achieve higher accuracy.
- Temperature contour and absorbed solar heat flux contours were plotted using ANSYS-FLUENT.

### *Governing Equations*

To perform numerical simulation of the proposed domestic hybrid solar dryer design, suitable equations were selected and were further resolved using ANSYS-FLUENT. CFD simulation is necessary for any standard model to simulate different parameters of the working fluid properties, viz. velocity, temperature, and pressure. It is important to solve various conservation equations (Equations 1-4) governing the flow behavior. These equations can be given as [1, 18]:

- Mass conservation equation:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0 \quad (1)$$

- Momentum conservation equation:

$$\frac{\partial}{\partial t} (\rho \mathbf{v}) + \nabla \cdot (\rho \mathbf{v} \mathbf{v}) = -\nabla p + \rho \mathbf{g} + \mathbf{F} \quad (2)$$

- Energy Conservation equation:

$$\frac{\partial}{\partial t} (\rho E) + \nabla \cdot [\mathbf{v}(\rho E + p)] = 0 \quad (3)$$

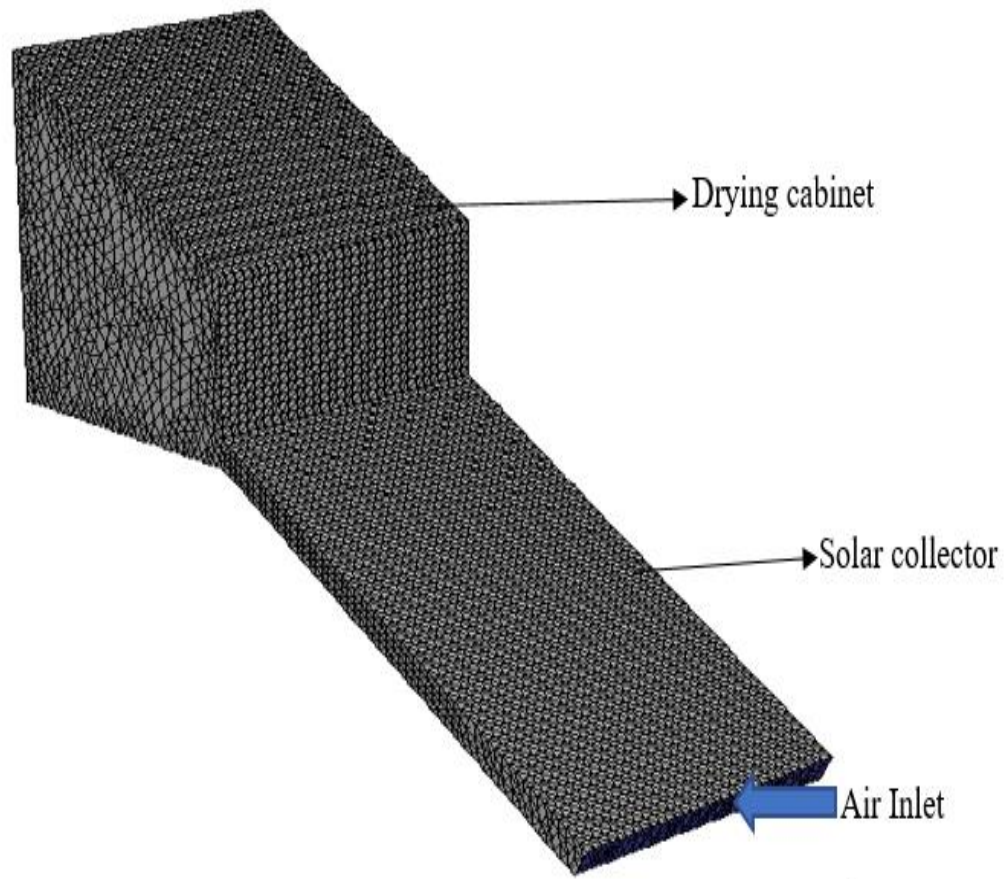
- Heat transfer radiation:

$$\frac{dI(r,s)}{ds} + (\alpha_s + \sigma_s)I(r,s) = \alpha_s n^2 \frac{\sigma T^4}{\pi} + \frac{\sigma_s}{\pi} \int_0^{4\pi} I(r,s) \phi(s,s') d\Omega' \quad (4)$$

## **RESULTS AND DISCUSSION**

The meshed view of the domestic hybrid solar dryer design has been displayed in Figure 2. The meshed elements were generated using quadratic mode. The grid independence test was conducted to get appropriate number of mesh elements that give optimum results. After grid independence test, 1, 46, 322 elements were considered for generation of results. The design parameters for domestic hybrid solar dryer analysis are mentioned in Table 2.





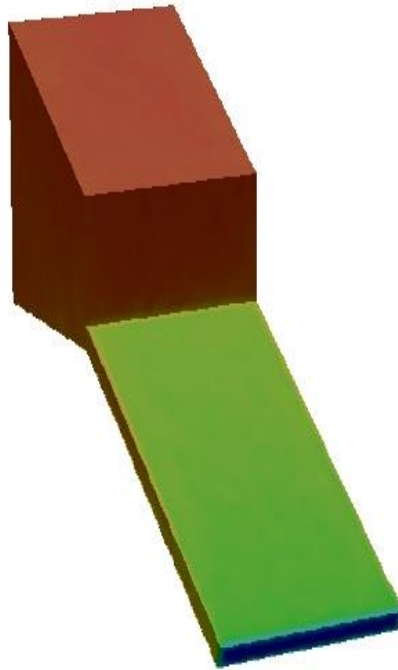
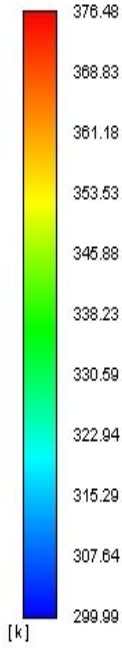
**FIGURE 2.** Meshed view of designed domestic hybrid solar dryer in ANSYS FLUENT 19.1.

**TABLE 2.** Design parameters for numerical simulation of domestic hybrid solar dryer.

Parameters	Value
Solar Insolation	According to solar load model
Latitude and longitude	28.4506° N, 77.5842° E
Date and Time	24 <sup>th</sup> February, 12: 00 PM
Density of working fluid (Air)	1.225 kg/m <sup>3</sup>
Walls	Insulated
Heated wall	Carbon coated copper sheet
Absorptivity of heated wall	0.94
Transmissivity of glass sheet	0.99
Thickness of glass sheet	8 mm
Inlet	Velocity Inlet
Outlet	Pressure outlet
Tilt angle	28°
Density of air	1.164 kg/m <sup>3</sup>
Thermal conductivity of air	0.02588 W/m-K
Density of glass	2500 kg/m <sup>3</sup>
Specific heat capacity of glass	750 J/kg- K
Thermal conductivity of glass	1.05 W/m-K
Convective heat transfer coefficient of copper sheet	13.34 W/m <sup>2</sup> -K
Thickness of copper sheet	1 mm

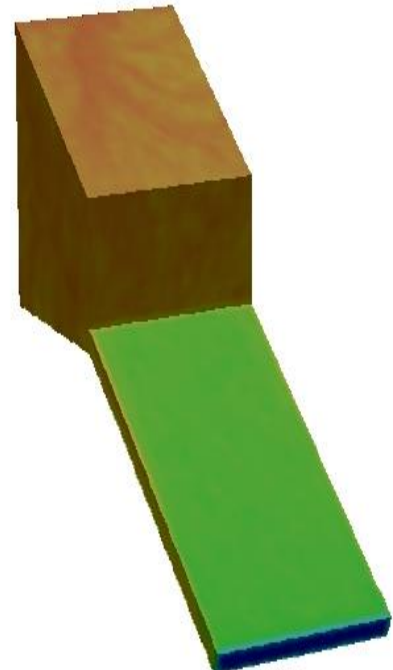
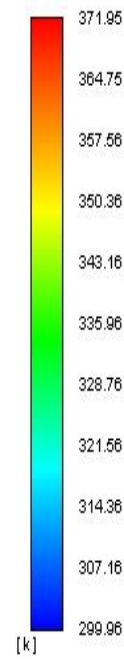
Temperature contours at different mass flow rates have been shown in Figure 3 (a-f). When the mass flow rates are lower (0.08125 kg/sec to 0.625 kg/sec), solar radiation will fall to the glass top of the solar collector system. Further, thermal radiation will be transmitted to the copper absorber metal box. The global and diffused solar radiation readings for the simulation as interpolated by ANSYS FLUENT were 875 W/m<sup>2</sup> and 207 W/m<sup>2</sup>. Under conduction, convection, and radiation, the air flowing from the inlet absorbs the thermal energy from the absorber box and circulates inside the drying cabinet. Due to the lower air mass flow rate, there will be high temperatures inside the drying cabinet. This can be observed clearly in Figure 4(a-c). As the mass flow rate of working fluid increases, the temperature reduction can be noted inside the drying cabinet. From Figure 4(d-f), it is observed that the temperature has been reduced due to an increase in mass flow rate. As per legends of the contours shown in Figure 3 (a-f), it can be observed that as the mass flow rate increases, the temperature in the drying cabinet decreases, which leads to decrease in temperature of outlet air 14. At mass flow rate of 1.2 kg/sec, the temperature inside the drying cabinet is about 348 K which can be observed from Figure 4 (e). There is a slight change (less than 1 K) in temperature after increasing the mass flow rate to 1.5625 kg/sec. Therefore, the mass flow rate of 1.2 kg/sec can be taken as optimum for drying food crops inside the domestic hybrid solar dryer under indirect mode operation.

contour-24  
Static Temperature



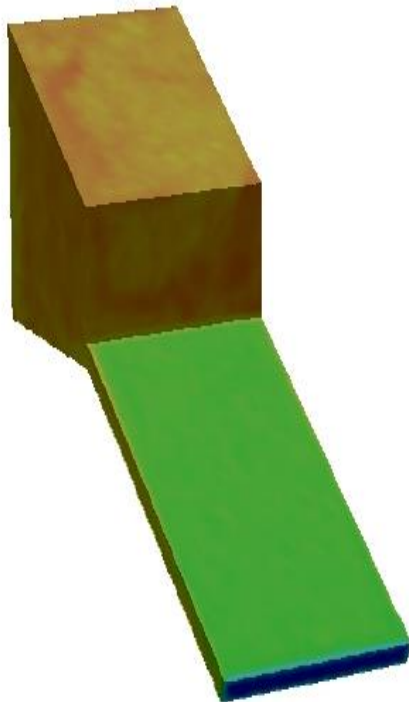
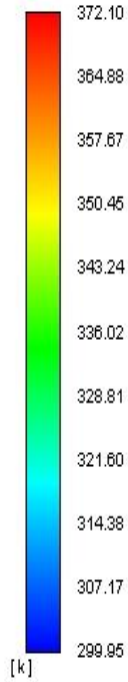
a) Mass flow rate= 0.08125 kg/s

contour-26  
Static Temperature



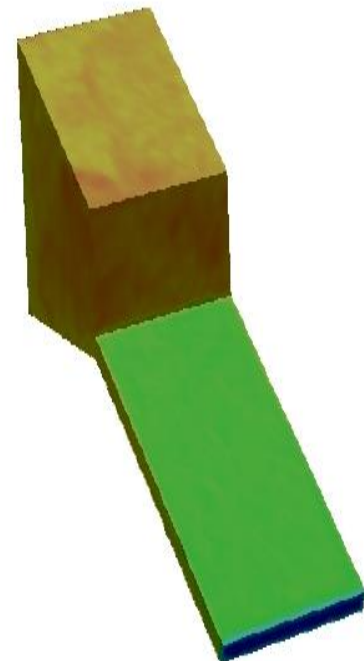
b) Mass flow rate = 0.3 kg/s

contour-23  
Static Temperature

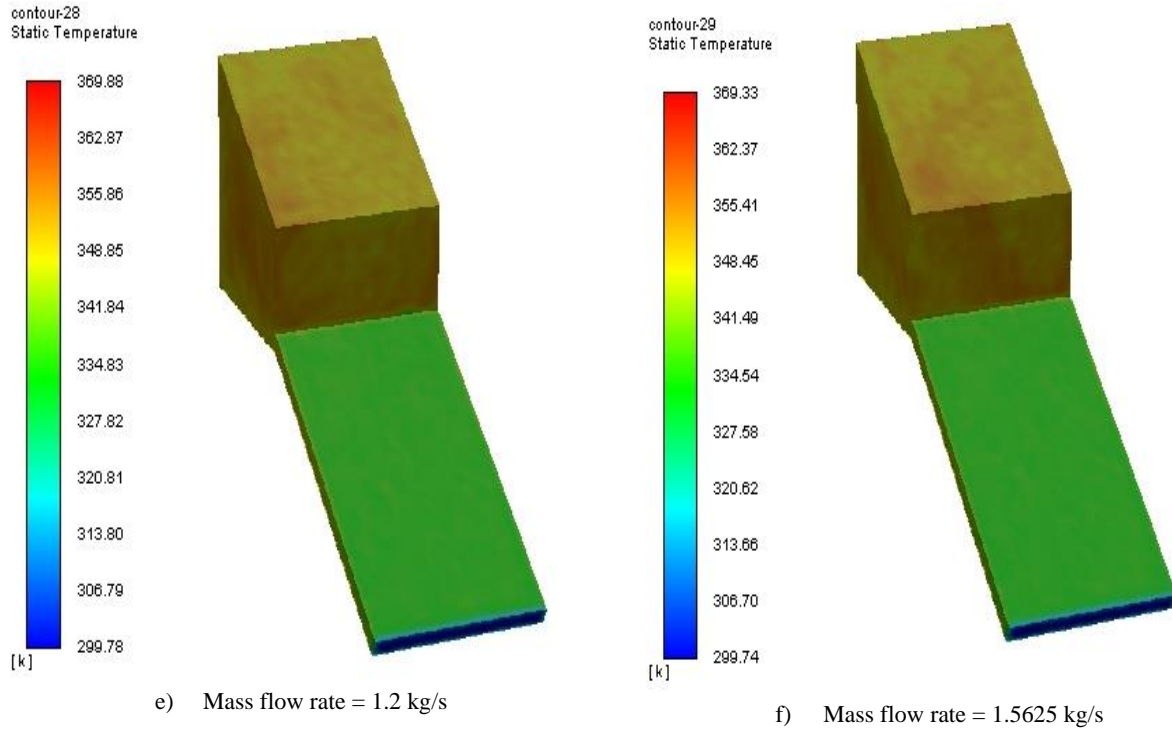


c) Mass flow rate = 0.625 kg/s

contour-27  
Static Temperature

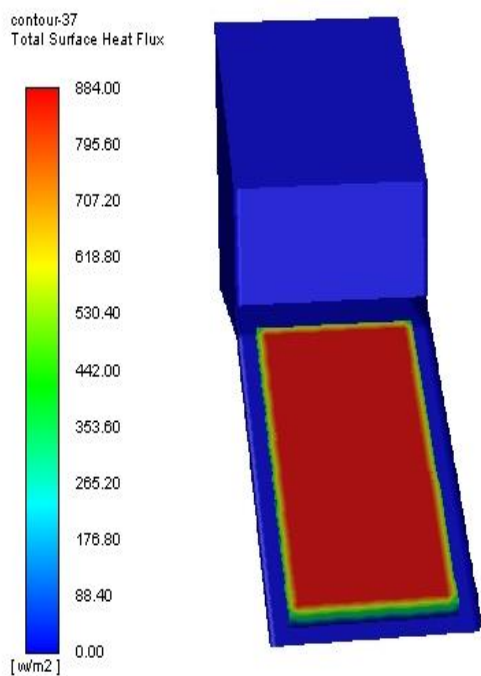


d) Mass flow rate = 0.9375

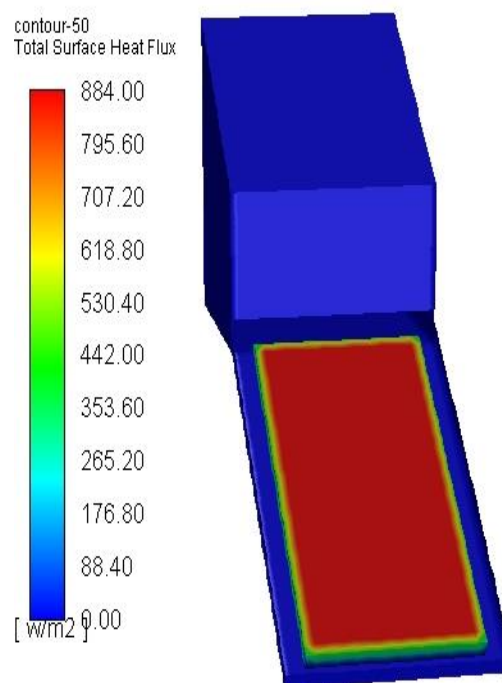


**FIGURE 3.** Static temperature variation inside domestic hybrid solar dryer in indirect mode operation under unload condition.

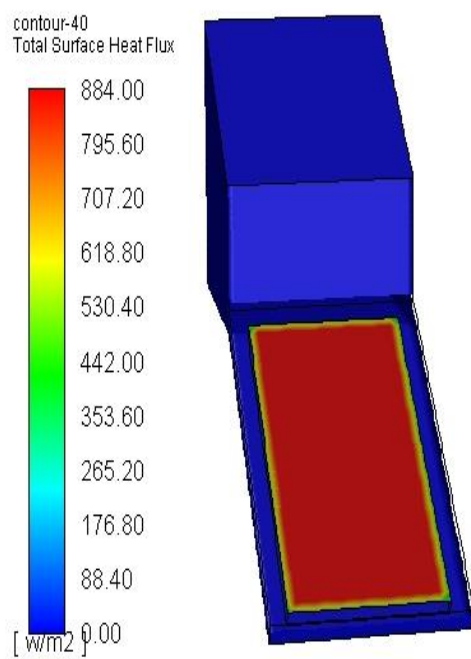
Figure 4 (a-f) shows the total surface heat flux in the domestic hybrid solar dryer under indirect mode operation. The total surface heat flux is slightly similar, as evident from the contours, at different mass flow rates, in the range of 884W/m<sup>2</sup> for all mass flow rates (0.08125 kg/s to 1.5625 kg/s). The reason behind these contours is that the input solar flux, geometry and boundary conditions are the same for all cases of simulated mass flow rates. The absorber box of the collector absorbs the maximum solar flux as compared with other components of the dryer, which validates the system's design [20].



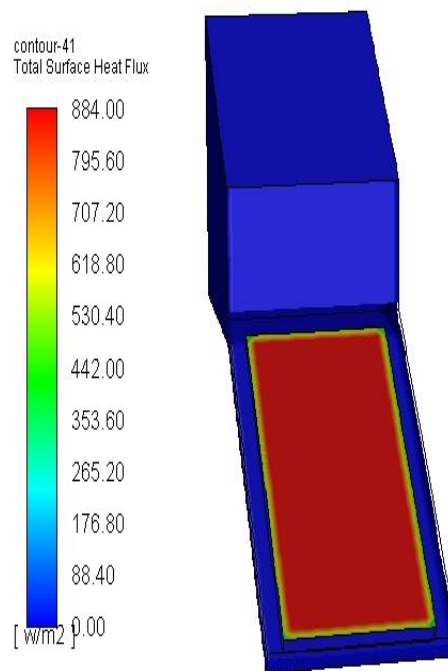
a) Mass flow rate = 0.08125 kg/s



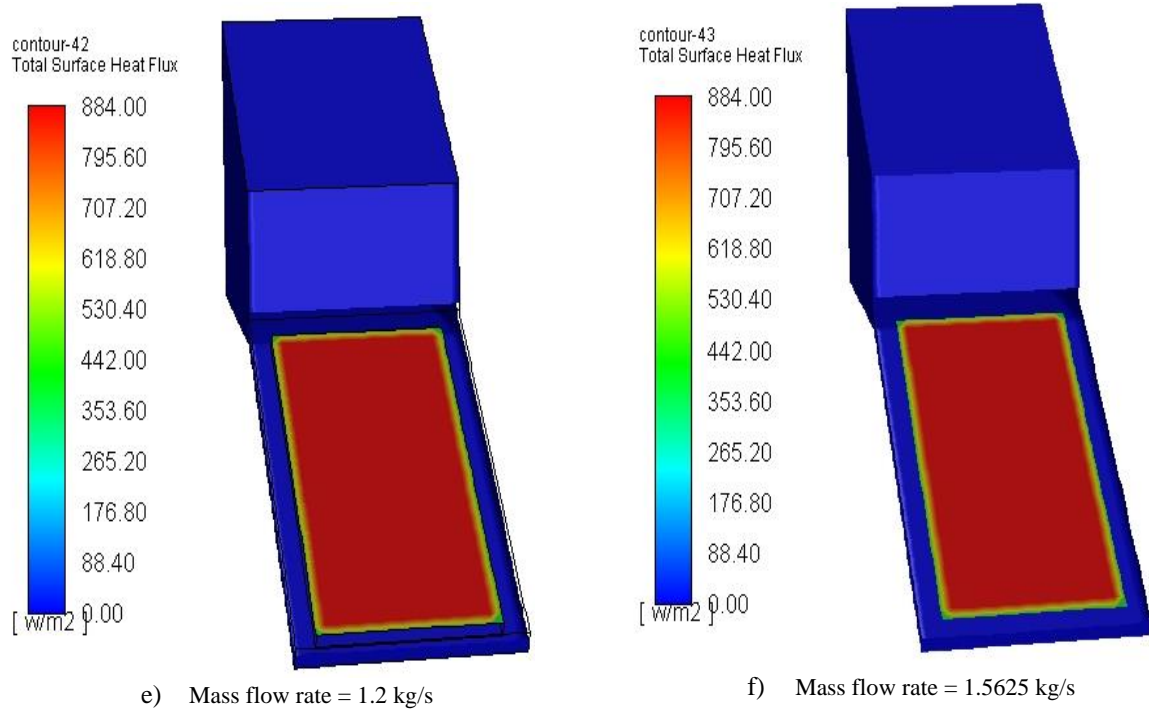
b) Mass flow rate = 0.3 kg/s



c) Mass flow rate = 0.625 kg/s



d) Mass flow rate = 0.9375



**FIGURE 4.** Total surface heat flux inside domestic hybrid solar dryer in indirect mode operation under unload condition

The temperature contour shown in Figure 3 (e) displays the temperature of air inside the drying cabinet as 348 K. Table 3 provides information about crops which can be dried inside the designed solar dryer as per the drying temperature of the crops [21].

**TABLE 3.** Crops suitable for drying in designed solar dryer

Crops	Maximum drying temperature
Carrots	75°C
Green beans	75°C
Potatoes	75°C
Sweet potatoes	75°C

## CONCLUSION

The domestic hybrid solar dryer has been simulated under indirect drying mode in ANSYS FLUENT for clear sky conditions. The following conclusions can be derived from the above discussion:

- The adequate mass flow rate of air inside the proposed simulated dryer is 1.2 kg/s.
- The temperature inside the drying cabinet rises to 348 K under the mass flow rate of 1.2 kg/s.
- The solar collector absorbs the maximum total surface heat flux as compared to other components of the domestic hybrid solar dryer which is in the range of 884 W/m<sup>2</sup> for all the simulated mass flow rate cases.
- Under these climatic conditions, the dryer is suitable to dry carrots, green beans, potatoes, and sweet potatoes.



## NOMENCLATURE

$v$	fluid velocity (in m/s)
$s_n$	the direction vector of the sun
$r$	position vector
$s'$	scattering direction vector
$n$	refractive index

## GREEK LETTERS

$\rho_s$	fluid density (in kg/m <sup>3</sup> )
$\tau$	transmissivity
$\alpha_s$	absorptivity
$\sigma_s$	scattering coefficient
$\sigma$	Stefan-Boltzmann constant (in W/m <sup>2</sup> K <sup>4</sup> )
$\phi$	phase function
$\Omega'$	solid angle (in Steradian)

## ACKNOWLEDGMENT

There is no direct funding for this research work. Authors are highly thankful to Bennett University, Greater Noida, India, and Centre for Energy and Environment, Delhi Technological University, for providing basic infrastructure for compiling this work.

## REFERENCES

1. Jain A, Sharma M, Kumar A, et al (2019) Computational fluid dynamics simulation and energy analysis of domestic direct-type multi-shelf solar dryer. *J Therm Anal Calorim*. <https://doi.org/10.1007/s10973-018-7973-5>
2. Atheaya D (2017) Economics of Solar Drying. In: Prakash O, Kumar A (eds) *Solar Drying Technology: Concept, Design, Testing, Modeling, Economics, and Environment*. Springer Singapore, Singapore, pp 441–462
3. Kumar A, Tiwari GN (2006) Thermal Modeling and Parametric Study of a Forced Convection Greenhouse Drying System for Jaggery: An Experimental Validation. *Int. J. Agric. Res.* 1:265–279
4. Prakash O, Kumar A (2013) Historical review and recent trends in solar drying systems. *Int J Green Energy* 10:690–738. <https://doi.org/10.1080/15435075.2012.727113>
5. Sharma M, Kumar A (2018) Promising biomass materials for biofuels in India's context. *Mater Lett*. <https://doi.org/10.1016/j.matlet.2018.03.034>
6. Saurabh A, Atheaya D, Kumar A (2020) Study of hybrid photovoltaic thermal systems. In: *IOP Conference Series: Materials Science and Engineering*. Institute of Physics Publishing

7. Tiwari D, Atheaya D, Kumar A, Kumar N (2022) Analysis of organic rankine cycle powered by N-number of solar collectors in series. *Energy Sources, Part A Recover Util Environ Eff* 44:6678–6697. <https://doi.org/10.1080/15567036.2022.2096726>
8. Sharma M, Prakash O, Sharma A, Kumar A (2018) Fundamentals and Performance Evaluation Parameters of Solar Dryer. In: Sharma A, Shukla A, Aye L (eds) *Low Carbon Energy Supply: Trends, Technology, Management*. Springer Singapore, Singapore, pp 37–50
9. Sharma M, Atheaya D, Kumar A (2021) Recent advancements of PCM based indirect type solar drying systems: A state of art. *Mater Today Proc*. <https://doi.org/10.1016/j.matpr.2021.04.280>
10. Singh Chauhan P, Kumar A, Tekasakul P (2015) Applications of software in solar drying systems: A review. *Renew Sustain Energy Rev* 51:1326–1337. <https://doi.org/10.1016/j.rser.2015.07.025>
11. Vijayan S, Arjunan T V, Kumar A (2017) Fundamental Concepts of Drying. In: Prakash O, Kumar A (eds) *Solar Drying Technology: Concept, Design, Testing, Modeling, Economics, and Environment*. Springer Singapore, Singapore, pp 3–38
12. Prakash O, Kumar A, Sharaf-Eldeen YI (2016) Review on Indian Solar Drying Status. *Curr Sustain Energy Reports* 3:113–120. <https://doi.org/10.1007/s40518-016-0058-9>
13. Sharma M, Atheaya D, Kumar A (2022) Exergy, drying kinetics and performance assessment of Solanum lycopersicum (tomatoes) drying in an indirect type domestic hybrid solar dryer (ITDHSD) system. *J Food Process Preserv* n/a:e16988. <https://doi.org/https://doi.org/10.1111/jfpp.16988>
14. Prakash O, Ranjan S, Kumar A, Tripathy PP (2017) Applications of Soft Computing in Solar Drying Systems. In: Prakash O, Kumar A (eds) *Solar Drying Technology: Concept, Design, Testing, Modeling, Economics, and Environment*. Springer Singapore, Singapore, pp 419–438
15. Iranmanesh M, Samimi Akhijahani H, Barghi Jahromi MS (2020) CFD modeling and evaluation the performance of a solar cabinet dryer equipped with evacuated tube solar collector and thermal storage system. *Renew Energy* 145:1192–1213. <https://doi.org/10.1016/j.renene.2019.06.038>
16. Singh R, Salhan P, Kumar A (2021) CFD Modelling and Simulation of an Indirect Forced Convection Solar Dryer. *IOP Conf Ser Earth Environ Sci* 795:. <https://doi.org/10.1088/1755-1315/795/1/012008>
17. Mellalou A, Riad W, Hnawi SK, et al (2021) Experimental and CFD Investigation of a Modified Uneven-Span Greenhouse Solar Dryer in No-Load Conditions under Natural Convection Mode. *Int J Photoenergy* 2021:. <https://doi.org/10.1155/2021/9918166>
18. Sonthikun S, Chairat P, Fardsin K, et al (2016) Computational fluid dynamic analysis of innovative design of solar-biomass hybrid dryer: An experimental validation. *Renew Energy* 92:185–191. <https://doi.org/10.1016/j.renene.2016.01.095>
19. Demissie P, Hayelom M, Kassaye A, et al (2019) Design, development and CFD modeling of indirect solar food dryer. *Energy Procedia* 158:1128–1134. <https://doi.org/10.1016/j.egypro.2019.01.278>
20. Xi Q, Long W, Ma Q (2021) Research of flat plate solar air collector in drying. *E3S Web Conf* 248:10–13. <https://doi.org/10.1051/e3sconf/202124802015>
21. Norton B (2017) Characteristics of Different Systems for the Solar Drying of Crops. In: Prakash O, Kumar A (eds) *Solar Drying Technology: Concept, Design, Testing, Modeling, Economics, and Environment*. Springer Singapore, Singapore, pp 69–88

# Constant Current Fault-Tolerant Buck Type Interleaved DC-DC Converter for Battery Charging Applications

Abhishek Chawla  
CoE for Electric Vehicle and Related Technologies  
Department of Electrical Engineering  
Delhi Technological University  
Delhi, India  
abhishekchawla15@gmail.com

Mayank Kumar (SM IEEE)  
CoE for Electric Vehicle and Related Technologies  
Department of Electrical Engineering  
Delhi Technological University  
Delhi, India  
mayankkumar@dtu.ac.in

**Abstract**— DC-DC converters are used in the variety of industrial processes, one of which is the battery charging for electric vehicles (EVs). Since, these converters are more prone to switch faults and to guarantee reliable operation fault tolerant converters are used. The paper presents the design of a 2.2 kW fault-tolerant interleaved buck converter. Additionally, it suggests an algorithm that can identify any switch defect within one switching cycle and adjust the gains to distribute the entire current equally among the healthy legs to provide constant current. In order to have the least amount of ripple in the output voltage and output current, it also updates the phase shift of healthy legs using phase selector. A PI control is implemented along with fault tolerant control which not only avoids sharp overshoots and undershoots but also input disturbances and load variation by adjusting the duty of each leg. The proposed scheme has been analyzed and simulated in MATLAB SIMULINK.

**Keywords**— Constant current charging, DC-DC power converters, electric vehicle, fault-tolerant control, interleaved converters.

## I. INTRODUCTION

With the increasing energy demand of the world, climate change and rapid depletion of the conventional resources and increasing hazardous effects of the pollution in urban areas, there is a global movement going on to bring cleaner and greener resources. Photovoltaic (PV) systems and wind energy systems contribute towards greener tomorrow, but they require energy storage element in order to mitigate the intermittent nature of solar PV and wind [1]- [2]. Because of this, the development of secondary batteries, particularly lead acid and Li-ion batteries are rapidly accelerating at the moment. Li-ion batteries have many benefits, including their high operating voltage, high-power density, lack of memory effect and ease of usage with electric vehicle (EV) power systems. In order to adapt these changes and for user convenience, quick charging of these batteries is necessary. With higher charging current, charging time to fully charge the battery reduces. Therefore, a chain of efficient charging systems is required that charges the battery within minimum time. In addition to the above-mentioned benefits, Li-ion batteries are widely used in EVs due to their high recycling and renewable qualities when compared to other types of batteries [3].

A power source and a DC-DC converter are the two main

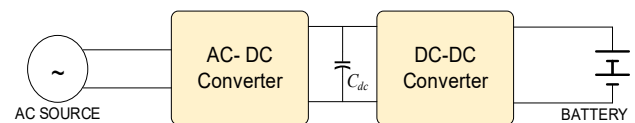


Fig1. Components of a Charger

components of a standard charging system. Fig. 1 represents the components of the charging system. Considering EV charging, it is usually done at the following power levels known as level 1, level 2, and level 3 charging. The level 1 being the slowest one charges upto the power of 1.9 kW. Level 2 charges upto the power level of 19.2 kW while level 3 can charge upto the power of 90 kW [4].

The dc-dc converters play an important role as it is not only just regulate the output voltage but provide high output current in order to charge the battery as soon as possible. One of the converters, that provide high output current with great efficiency is the interleaved buck converter (IBC). With greater efficiency and higher output current it also offers reduction in component size, reduced ripple voltage and reduced output current ripple. Despite having all such advantages, we cannot connect a converter directly to a battery. A proper control is required to charge the battery. Generally, the modes are classified as constant-current (CC) mode, constant-voltage (CV) mode, constant-current constant-voltage (CC-CV) mode and constant trickle charging (CTC) mode [3]. Fig. 2 represents the different regions where battery is charged in constant-current mode and in constant-voltage mode.

The IBC provides a constant output current and charges the battery in CC mode. [5] employs a technique in which the output current values for each phase are averaged using the average current method, and the resulting value is then utilized as the control reference to command each phase separately. Consequently, the phases with lower output currents will increase their duty cycle in order to raise their output current in accordance with the averaged value, and vice versa, to guarantee equal amount of current. The key advantage of the average current approach is sharing precision, but it has a poor fault tolerance capability because even one damaged phase will cause the system to fail. The following paper works towards making the system reliable and fault tolerant by altering the gains of the controller with the aid of fault detection system.

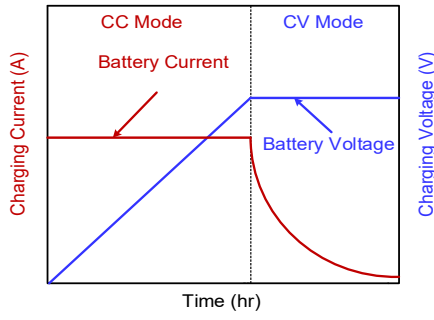


Fig 2. CC-CV modes during battery charging.

A lot of research has been going on in the field of fault diagnosis and fault tolerant control scheme and it is reported in the literature that about 30-35% of faults occurred in the power converters are because of the semiconductor switch faults. The most common switch faults are open circuit faults (OCF) and short circuit faults (SCF). The most critical switch fault is SCF, which results in extreme high current that can turn the complete power converter off. Although the OCF is not as severe as the SCF, but the effect of this fault cannot be ignored. If it is not diagnosed quickly enough, additional switches and circuit components can become over stressed and fail [6]. Therefore, detection of fault is as important as making the converter tolerant from these faults. In [7], gives an extensive overview about different fault diagnostic algorithm. It classifies the algorithm into two categories naming model-based algorithms and signal processing (SP) based algorithms. The SP based algorithms are further classified into time and frequency domain analysis. Also, it compares various SP based algorithm (time and frequency domain wise) and gives an insight about the parameters such as diagnosis criteria, maximum diagnosis time, diagnosis speed and the most important parameter cost. In [8], uses a reconfigurable approach to make the system fault tolerant. Switch voltage is considered as the parameter for the detection of the fault and it is done within one switching cycle. The article in [9] presents a technique of derivative of inductor current based switch fault detection. It compares measured and predicted value of inductor currents and identifies fault in case of change in the behavior of the inductor current. In [10], a stacked interleaved buck converter with a series capacitor and an auxiliary bidirectional boost/buck branch is used to experimentally confirm the converter's ability to tolerate several fault modes. In [11], OCF identified in the converter by monitoring the input voltage, current and power. Also, a 3-phase interleaved converter is made to work as 2- phase interleaved converter but with reduced power ratings in order not to affect the continuity of the system.

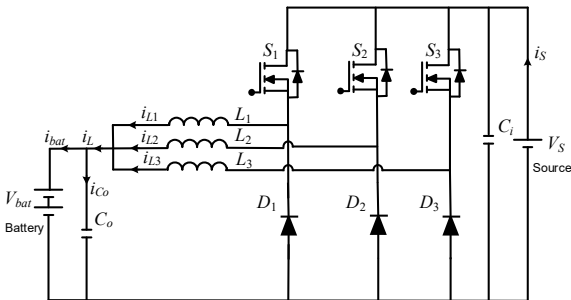


Fig 3. Three phase interleaved buck converter.

## II. STATE SPACE AVERAGE MODELLING OF INTERLEAVED BUCK CONVERTER

The modelling of three phase interleaved buck converter is presented without the consideration of circuit parasitic. The converter is analyzed over a switching cycle and each leg of the converter is operated by providing a phase shift of 120°. Fig 4 shows that the converter operates in 6 different modes which are discussed below:

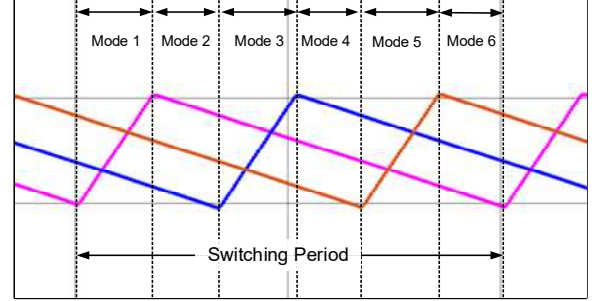


Fig 4. Operating modes of the three-phase interleaved converter

*Mode 1:* Switch  $S_1$  – ON;  $S_2, S_3$  – OFF; and diode  $D_1$  – OFF;  $D_2, D_3$  – ON:

$$\frac{di_{L1}}{dt} = \frac{-V_c}{L_1} + \frac{V_{in}}{L_1} \quad (1)$$

$$\frac{di_{L2}}{dt} = \frac{-V_c}{L_2} \quad (2)$$

$$\frac{di_{L3}}{dt} = \frac{-V_c}{L_3} \quad (3)$$

$$\frac{dv_c}{dt} = \frac{i_{L1}}{C} + \frac{i_{L2}}{C} + \frac{i_{L3}}{C} - \frac{v_c}{RC} \quad (4)$$

*Mode 2, Mode 4 and Mode 6:* Switches  $S_1, S_2, S_3$  – OFF; and diodes  $D_1, D_2, D_3$  – ON:

$$\frac{di_{L1}}{dt} = \frac{-V_c}{L_1} \quad (5)$$

$$\frac{di_{L2}}{dt} = \frac{-V_c}{L_2} \quad (6)$$

$$\frac{di_{L3}}{dt} = \frac{-V_c}{L_3} \quad (7)$$

$$\frac{dv_c}{dt} = \frac{i_{L1}}{C} + \frac{i_{L2}}{C} + \frac{i_{L3}}{C} - \frac{v_c}{RC} \quad (8)$$

*Mode 3:* Switch  $S_2$  – ON;  $S_1, S_3$  – OFF; and diode  $D_2$  – OFF;  $D_1, D_3$  – ON:

$$\frac{di_{L1}}{dt} = \frac{-V_c}{L_1} \quad (9)$$

$$\frac{di_{L2}}{dt} = \frac{-V_c}{L_2} + \frac{V_{in}}{L_2} \quad (10)$$

$$\frac{di_{L3}}{dt} = \frac{-V_c}{L_3} \quad (11)$$

$$\frac{dv_c}{dt} = \frac{i_{L1}}{C} + \frac{i_{L2}}{C} + \frac{i_{L3}}{C} - \frac{v_c}{RC} \quad (12)$$

Mode 5: Switch  $S_3$  – ON;  $S_1, S_2$  – OFF; and diode  $D_3$  – OFF;  $D_1, D_2$  – ON:

$$\frac{di_{L_1}}{dt} = \frac{-V_c}{L_1} \quad (13)$$

$$\frac{di_{L_2}}{dt} = \frac{-V_c}{L_2} \quad (14)$$

$$\frac{di_{L_3}}{dt} = \frac{-V_c}{L_3} + \frac{V_{in}}{L_3} \quad (15)$$

$$\frac{dv_c}{dt} = \frac{i_{L_1}}{C} + \frac{i_{L_2}}{C} + \frac{i_{L_3}}{C} - \frac{v_c}{RC} \quad (16)$$

The state space equation for the above system can be written as follows:

$$\dot{x} = Ax + Bu \quad (17)$$

$$y = Cx + Du \quad (18)$$

where:  $\dot{x}$  = derivative of state vector,  
 $u$  = input vector,  
 $y$  = output vector,  
 $A$  = system matrix,  
 $B$  = input matrix,  
 $C$  = output matrix,  
 $D$  = feed forward matrix.

The above equations, (1) to (16) for different modes can be written in form of matrix and then system matrix and output matrix can be calculated as follows:

$$A = A_1d_1 + A_2d_2 + A_3d_3 + A_4d_4 + A_5d_5 + A_6d_6 \quad (19)$$

$$B = B_1d_1 + B_2d_2 + B_3d_3 + B_4d_4 + B_5d_5 + B_6d_6 \quad (20)$$

$$\text{using } d_1 + d_2 + d_3 + d_4 + d_5 + d_6 = 1 \quad (21)$$

where  $d_1 = d_3 = d_5 = D$ ;  $d_2 = d_4 = d_6 = \frac{1}{3} - D$ ;  $D$  being the duty cycle of the converter. For the above system:

$$A = \begin{bmatrix} 0 & 0 & 0 & \frac{-1}{L_1} \\ 0 & 0 & 0 & \frac{-1}{L_2} \\ 0 & 0 & 0 & \frac{-1}{L_3} \\ \frac{1}{C} & \frac{1}{C} & \frac{1}{C} & \frac{-1}{RC} \end{bmatrix} \quad B = \begin{bmatrix} \frac{d_1}{L_1} \\ \frac{d_3}{L_2} \\ \frac{d_5}{L_3} \\ 0 \end{bmatrix} \quad C = [0 \ 0 \ 0 \ 1] \quad D = [0 \ 0 \ 0 \ 0] \quad (22)$$

The transfer function of the three-phase interleaved buck converter can be calculated as by taking the Laplace inverse transform of (18).

### III. FAULT TOLERANT OPERATION OF INTERLEAVED BUCK CONVERTER

#### A. Operation under Normal Condition

Under normal circumstances, it is anticipated that all of the phases have same parameters. Ideally, all the phases

should share equal amount of current but because of the non-idealities in the system, some phase may have greater current than the other branch. To guarantee equitable distribution of current throughout the phases, the primary converter is managed to operate using a closed-loop system. The closed loop system uses a PI controller followed by a pulse modulation generator to regulate the inductor currents and produce duty cycle. Each phase gets the exact same duty cycle generated by the closed loop current system but are phase-shifted with each other by  $120^\circ$ . The converter's output current is the total of its three phase currents. The output current's ripple frequency is  $n$  times the converter's switching frequency. Also, because of the interleaving the ripples are reduced approximately by  $n$  times. A sliding mode duty ratio control along with current balancing technique is proposed in [12] that provides same current through the phases of the converter. The concept presented in [5] uses average current method for equal distribution of current in each phase. All these techniques ensure constant current through the converter that is required to charge the battery in CC mode. Fig. 5 shows the waveform for  $D < 0.33$  under healthy conditions.

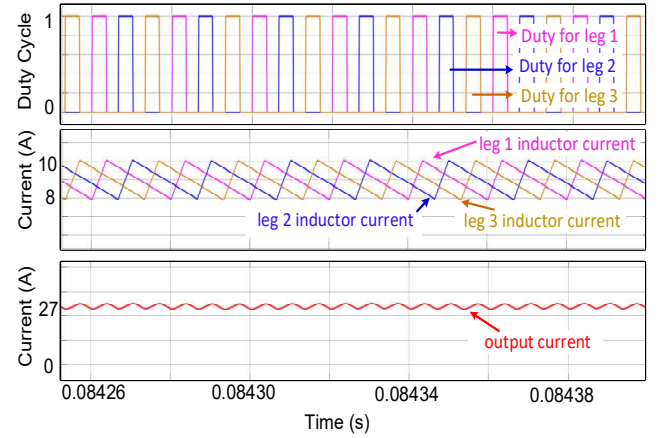


Fig 5. (a) Duty cycle for each leg (b) inductor current for each leg (c) output current of converter.

#### B. Operation under Faulty Condition

The following section deals with the operational process of interleaved buck converter under open circuit fault condition. The aim of the converter is to provide constant current and not to affect the output power of converter even after experiencing a fault. Therefore, to get the desired characteristics, it is assumed that the converter does not operate at the rated power. Fig. 6 shows the overview of the system and how the fault detection system helps in changing gain and adjusting the phase of the legs after the detection of the fault.

Consider the scenario where any of the system's legs experience an open circuit fault, say leg 3. The fault is detected within one switching cycle and the relevant leg's fault signal is raised by the fault-detection algorithm. The gains of the controller are adjusted in such a way that the remaining healthy legs (leg 1 and leg 2) share equal current, and the output current is maintained constant. In addition to that, the phase shifter block adjusts the phases of the duty cycles for the remaining healthy phases. The phases are now phase shifted by  $180^\circ$  ensuring minimum ripple in the output current so that it does not affect the battery's life.



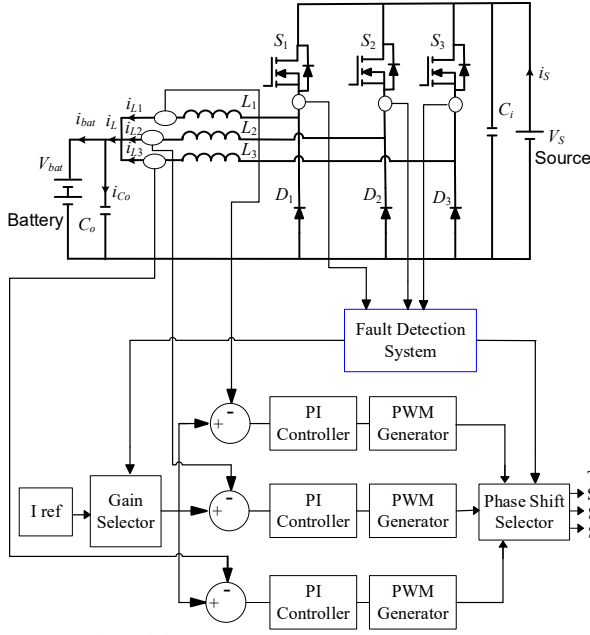


Fig. 6. Overview of the system

#### IV. FAULT TOLERANT CONTROL OF CONVERTER

Over a period of time, many techniques have been developed and reported in literature that deals with a specific fault in the system. The faults in EV chargers ranges from switch level fault, leg level fault, module level fault, measurement level fault, network level fault and system level fault [13]. The proposed is a SP based algorithm that deals with the switch level faults. Generally, there are two prominent switch level faults naming Open Circuit Fault and Short Circuit Fault. Although the performance and efficiency of the power converters are reduced, open-circuit failures are unlikely to result in a catastrophic failure of power converter. Short Circuit failures can lead to shut down of the converter and even it can burn the whole system if not detected and treated. Therefore, it is crucial for these converters to have tolerable functionalities. The process of making a converter fault-tolerant involves several steps. The first step being fault diagnosis. It can be done by finding the defect in the switch and activating the alarm signal. The second step is to isolate the defective phase and then different controlling techniques can be applied to avoid degrading the converter's power ratings.

Picture a situation where leg 1 develops an open circuit fault. Current Sensors are used to sense the leg's current and then these are passed through Analog to Digital Converter (ADC).  $n$  number of current sensors are used. Each sensor contributes towards detecting a fault in the respective leg. These are then added to and transmitted via a magnitude comparator after passing through a phase shifter. For each leg, the same procedure can be carried out again. In the event of an open circuit failure (phase current falling below the threshold current value), each leg's magnitude immediately becomes zero. A fuse is provided to sever the connection in the event of a short circuit fault. The alarm signal for the respective leg gets high. The gains are adjusted as these are controlled using the alarm signals of the legs. The phase shifter block updates the phases for the remaining two healthy legs with  $0^\circ$  and  $180^\circ$  in order not to increase the ripple content in the output current.

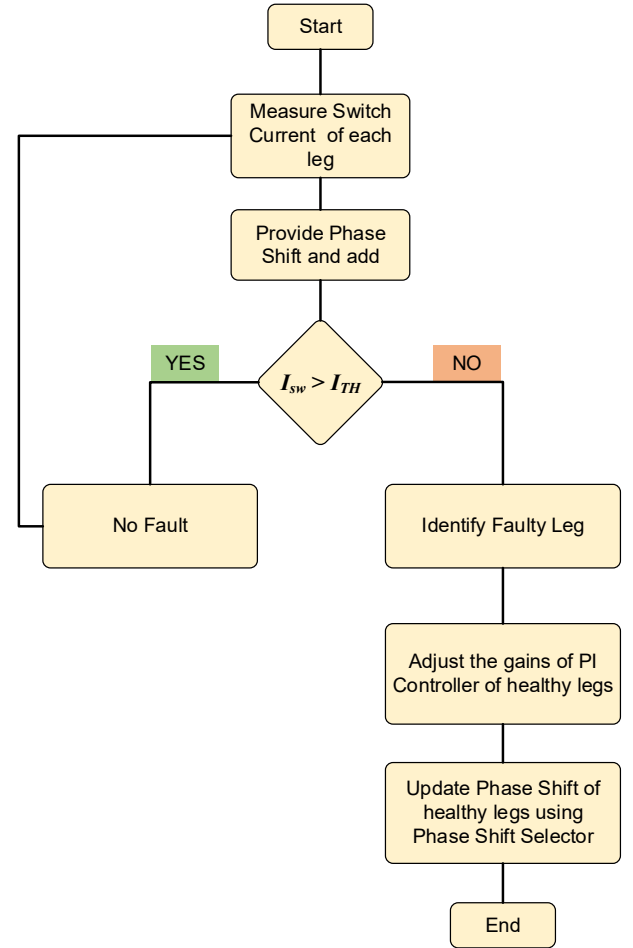


Fig. 7: Flow chart of fault tolerant control scheme.

TABLE I: PHASE SHIFT UNDER FAULTY CONDITIONS

Fault in Leg	Phase of Leg 1	Phase of Leg 2	Phase of Leg 3
No Fault	$0^\circ$	$120^\circ$	$240^\circ$
Fault in Leg 1	-	$180^\circ$	$0^\circ$
Fault in Leg 2	$180^\circ$	-	$0^\circ$
Fault in Leg 3	$0^\circ$	$180^\circ$	-

Table I summarizes the phase shift that needs to be updated in case of fault while Fig 8 gives an insight how the phase shift controller is designed using the table with the help of digital circuits. The PWM generator generates respective duty cycle for each leg. The phase shifter block controls the phase shift of each leg with the help of fault detection signals. For leg 1, fault alarm signal of leg 2 controls the 2:1 Mux, for leg 2, fault alarm signal of leg 1 and leg 3 act as control signal to the mux whereas for leg 3 fault alarm signal of leg 1 and leg 2 act as controlling signal. They provide the respective phase shift in case of fault occurrence and avoid rise of ripple in the output current. Considering the remaining cases, if the second leg or third leg experiences a problem, the overall process might be used, and the results will be similar to that of one considered here. The suggested research uses a signal-based approach to locate the fault and, like any other standard SP technique, is susceptible to erroneous system triggering in highly dynamic environments. Therefore, a tradeoff is to be made between the fault detection time and the reliable operation of the algorithm.

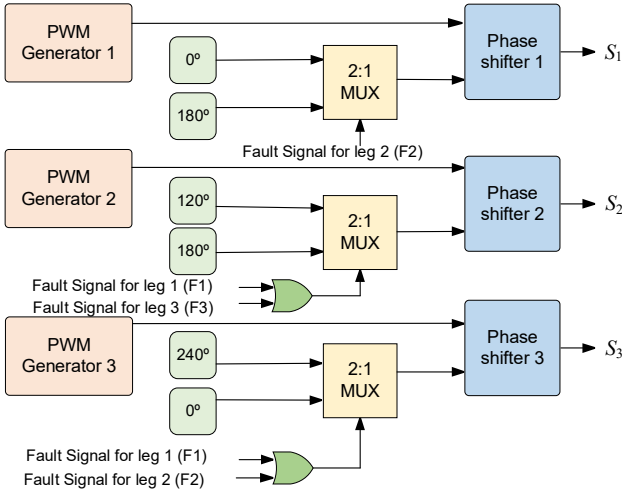


Fig. 8. Block diagram of phase shift selector.

## V. SIMULATION RESULTS

The recommended fault-tolerant control mechanism has been modelled using MATLAB/SIMULINK. Table II lists the parameters that were utilized to simulate the system. The inductor current waveform for each of the legs for  $D < 0.33$  (non-overlapping) is shown in Fig. 9. Moreover, the PI control is used to achieve equitable current distribution across legs. Fig. 10. manifests an open-circuit switch fault, which is simulated in the converter's first phase at time  $t = 0.3$ s. The gains of the PI controller are changed in less than one switching cycle, or less than  $20 \mu s$ , upon the occurrence of the fault, while the current through the faulty leg becomes zero. It also shows the effectiveness of the digitally implemented PI control.

TABLE II: SYSTEM PARAMETERS

Parameter Name	Value
Power Ratings	2.2 kW
Per Phase Inductance	650 $\mu H$
Switching Frequency	50 kHz
Input Capacitance	80 $\mu F$
Output Capacitance	50 $\mu F$
Input Voltage	360-440 V
Nominal Voltage of Battery	72 V
Ah ratings of Battery	40 Ahr
Proportional Gain ( $K_p$ ) of Controller	0.01
Integral Gain ( $K_i$ ) of Controller	70

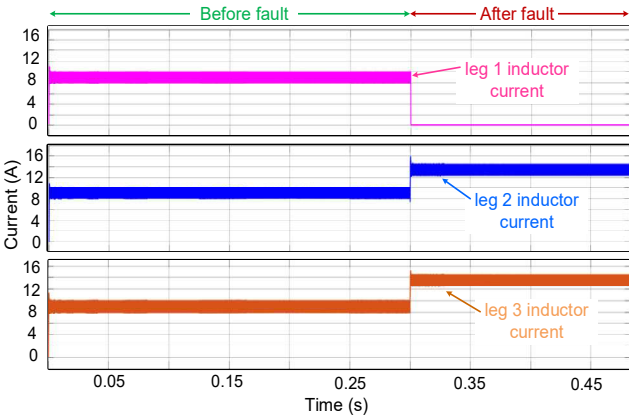


Fig. 9. Inductor current of each leg before and after fault.

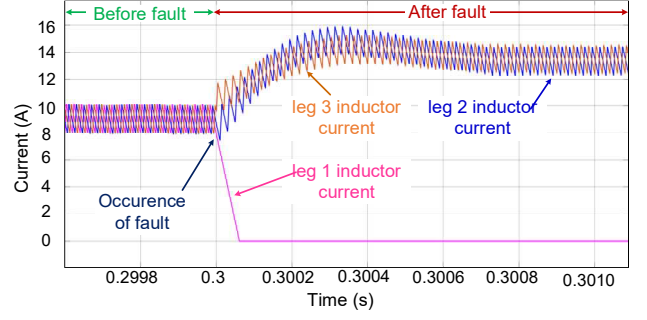


Fig. 10. Effect of fault on inductor current of different legs.

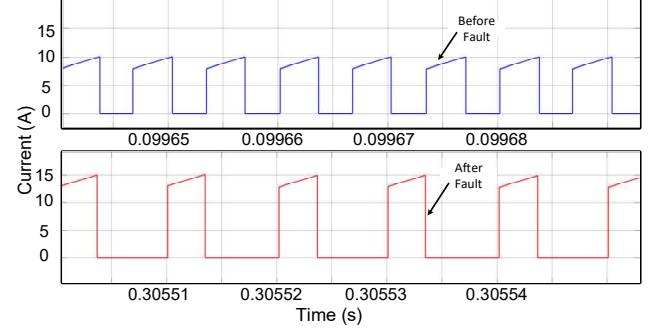


Fig. 11. Input current  $i_s$ , (a) before fault, and (b) after fault.

Fig. 11 depicts the impact of the fault on the input current. Input current appears to be operating as a single channel converter at a frequency of 3 times switching frequency prior to fault but operates at 2 times switching frequency following fault. Moreover, the current is increased to 15A each phase in order to maintain system continuity.

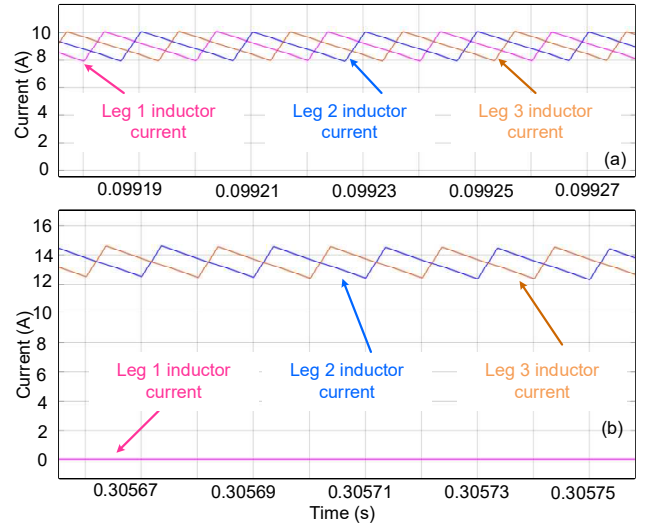


Fig. 12. Inductor currents (a) before fault, and (b) after fault.

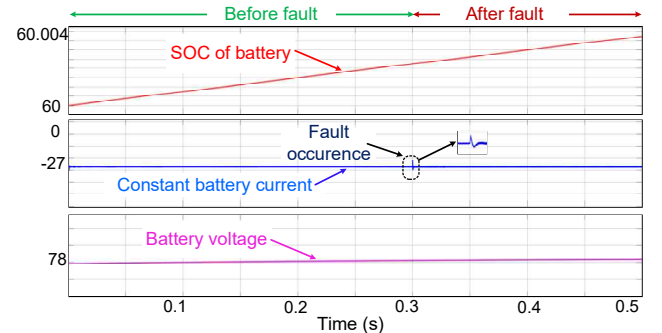


Fig. 13. (a) battery's SoC, (b) battery current, (c) battery voltage.



In order to demonstrate the usefulness of the phase selector block, Fig. 12 makes sure that inductor currents are phase shifted by  $120^\circ$  before the fault and are out of phase i.e.  $180^\circ$  phase shifted with one another after the fault. The current shoots up to almost 15A per phase to maintain same output current. The circuit is digitally built to guarantee the least amount of output current and output voltage ripple for a longer battery life. Fig. 13 depicts the battery being charged at a steady 27A current. The current is kept constant even when one converter leg is turned off, with just a slight decrease in battery current for a few milliseconds. It will take roughly 2 hours to fully charge the battery as it has a capacity of 40 Ah.

## VI. CONCLUSION

One of the most frequent defects in dc-dc converters is a switch fault. Fast detection of these defects enables circuit components to be safeguarded, greatly enhancing converter dependability and reliability. The primary focuses of this study are fault detection, tolerant control, and phase current control of interleaved dc-dc converters. Because the control is implemented digitally, the system's cost is not increased. In order to offer constant current for battery charging applications, including those for EVs and other places where a battery needs to be charged, it integrates closed loop PI control and fault detection capabilities. A phase shift of  $360/n$  is supplied among healthy legs using a phase selector to have the least current ripples. Each leg of the converter is controlled by a PI controller, which shares identical phase shifted current with each leg. The results show that it is resilient to input fluctuations, and it may quickly identify faults so that it may act to fix them. Moreover, the findings are transferable to other non-isolated interleaved converters.

## ACKNOWLEDGMENT

This research supported by the Science and Engineering Research Board (SERB), Department of Science & Technology, Government of India, under the SERB sanction order number SRG/2021/001640.

## REFERENCES

- [1]. M. Haque, S. Das, M. R. Uddin, M. I. Leon and M. A. Razzak, "Performance Evaluation of 1kW Asynchronous and Synchronous Buck Converter-based Solar-powered Battery Charging System for Electric Vehicles," in *IEEE Region 10 Symposium (TENSYP)*, 2020.
- [2]. M. Hossain and M. Islam, "Power stage design of a synchronous buck converter for battery charger application," in *International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, 2018.
- [3]. H. Suryatmojo, "Design Li-Po battery charger with buck converter under partially CC-CV method," in *International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 2020.
- [4]. G. Balen, A. R. Reis, H. Pinheiro and L. Schuch, "Modeling and control of interleaved buck converter for electric vehicle fast chargers," in *Brazilian Power Electronics Conference (COBEP)*, 2017.
- [5]. S. Y. Ou and L. Y. Liu, "Design and implementation of a four-phase converter with digital current sharing control for battery charger," in *IEEE International Telecommunications Energy Conference (INTELEC)*, 2015.
- [6]. S. Siouane, S. Jovanović and P. Poure, "Open-switch fault-tolerant operation of a two-stage buck/buck-boost converter with redundant synchronous switch for PV systems," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 5, pp. 3938-3947, 2018.
- [7]. G. K. Kumar and Elangovan, "Review on fault-diagnosis and fault-tolerance for DC-DC converters," *IET Power Electronics*, vol. 13, no. 1, pp. 1-13, 2020.
- [8]. S. Jagtap and D. More, "Switch Open-circuit Fault Diagnosis and Fault-Tolerant Control for Boost DC-DC Converter," in *Procedia Computer Science*, 2020.
- [9]. E. Pazouki, Y. Sozer and J. A. De Abreu-Garcia, *Fault Diagnosis and Fault Tolerant Operation of Non-Isolated DC-DC Converters..*
- [10]. X. Guo, S. Zhang, Z. Liu, L. Sun, Z. Lu, C. Hua and J. Guerrero, "A new multi-mode fault-tolerant operation control strategy of multiphase stacked interleaved Buck converter for green hydrogen production," *International Journal of Hydrogen Energy*, vol. 47, no. 71, pp. 30359-30370, 2022.
- [11]. E. Ribeiro, A. Cardoso and C. Boccaletti, "Fault-tolerant strategy for a photovoltaic DC-DC converter," *IEEE transactions on power electronics*, vol. 28, no. 6, pp. 3008-3018, 2012.
- [12]. M. Mahmud, Y. Zhao and Y. Zhang, "A sliding mode duty-ratio control with current balancing algorithm for interleaved buck converters," in *IEEE Applied Power Electronics Conference and Exposition (APEC)*, 2018.
- [13]. L. Gaona-Cárdenas, N. Vázquez-Nava, O. Ruiz-Martínez, A. Espinosa-Calderón, A. Barranco-Gutiérrez and M. Rodríguez-Licea, "An Overview on Fault Management for Electric Vehicle Onboard Chargers," *Mdpi*, vol. 11, no. 7, p. 1107, 2022.

# CORRDroid - Android Malware Detection using Association amongst Permissions

\*

Ankita Jain

*Dept. of Applied Mathematics  
Delhi Technological University  
New Delhi, India*

ankitajain\_2k19mc018@dtu.ac.in

Lakshit Rustagi

*Dept. of Applied Mathematics  
Delhi Technological University  
New Delhi, India*

lakshitrustagi\_2k19mc064@dtu.ac.in

Mayank Aggarwal

*Dept. of Applied Mathematics  
Delhi Technological University  
New Delhi, India*

mayankaggarwal\_2k19mc069@dtu.ac.in

Anshul Arora

*Dept. of Applied Mathematics  
Delhi Technological University  
New Delhi, India*

anshularora@dtu.ac.in

**Abstract**—The introduction of Apple’s iPhone in 2007 marked the beginning of a new era for mobile devices and applications. In 2012, more Android smartphones were sold than iPhones, and since then, Android smartphones have become increasingly popular. Android’s ubiquity has attracted the attention of attackers and attacks against the platform are on the rise. Numerous malware applications target mobile devices and compromise sensitive and private information stored on them. Hence, stronger security solutions need to be developed to detect such threats. In the literature, there is a variety of static Android malware detection techniques based on the analysis of manifest file components such as permissions. To the best of our knowledge, none of these attempts have aimed to determine the most distinguishing permission pair based on the association of permissions that could result in greater accuracy. In this study, we propose a risky permission-based malware detection system that uses association between permissions to create a set of important permissions and then uses supervised learning techniques to effectively classify malicious and benign applications. The experimental results demonstrate that the SVM gave the best accuracy of 97.2%.

**Index Terms**—Android Malware, Android Security, Malware Detection, Risky Permissions, Correlation

## I. INTRODUCTION

Between 2021-2022, desktop internet usage dropped from 41.52% to 37.8%, while mobile internet usage increased from 56.05% to 60.66% [1]. Smartphones are becoming increasingly popular over desktops for a variety of reasons. They are smaller and easier to carry which makes them more portable. Moreover, these phones can be used as cameras, music players, GPS devices, etc., in addition to traditional calling and SMS services. Social media also adds to the growing popularity.

When the Apple iPhone came out in 2007, it marked the beginning of a new era for mobile devices. After Google launched Android in 2008, more Android devices were sold than iPhones by 2012 [1]. Today, Android is the most popular mobile OS with more than 2.5 billion active users [2].

Android markets offer a variety of Android apps, but Google’s market is the largest repository [3]. There are currently over 3 million apps available on the PlayStore [4]. More than 3.55 million apps have been downloaded over 10 billion times in the third quarter of 2022 [5]. In recent years, attacks against mobile devices based on Android have increased because Android’s ubiquity has attracted the attention of attackers as well. Unlike Apple’s App Store, Google Play Store does not personally validate released software. A dynamically simulated environment, Bouncer, is used to manage and protect the official market against malicious attacks. It protects against malware attacks but does not inspect supplied software. Also, Android’s open-source model makes it possible to install apps from third-party markets, leading to the proliferation of regional and global app stores.

The developers of malware apps exploit platform weaknesses, collect sensitive user data, extort monetary benefits from telecommunication providers, or set up botnets to control smartphones. Hence, arises the need to develop an efficient Android malware detection model. A malware detection system for Android can be classified into three categories: static detection, dynamic detection, and hybrid detection. Static detection analyzes suspicious codes without actually executing Android applications. Dynamic detection analyzes Android programs by executing their code. The result can be attacks that static analysis cannot detect, but dynamic detection takes a lot of time and processing resources. Hybrid techniques work with a combination of static and dynamic features. This paper

proposes a novel model for detecting Android malware using Risky Permissions and Feature Ranking.

#### A. Motivation

Several works have been proposed in the literature for Android Malware Detection such as [6], [7], etc. However, none of them used association metrics to identify risky permission patterns. Since permissions patterns are very similar in both normal and malware apps, hence, this work aims to identify risky permission patterns using the Cosine method and the Kulczynski measure. Further, our objective is to develop a novel algorithm to effectively detect Android malware using the risky permissions set obtained from the above-mentioned approaches by using supervised learning techniques such as Decision Trees, SVM, Naive Bayes, and Random Forest.

#### B. Contribution

The CORRdroid framework presented in this paper is a new approach to classifying Android malware samples. The main contributions of this work are summarized below:

- We identified a precise set of permissions with association techniques of Cosine and Kulczynski measure to find their correlation.
- We evaluated the set of permissions on seven classifiers to effectively determine Android malware.
- A total of 12,070 benign and an equal number of malware samples were used for experiments and we could achieve the detection accuracy of 97.2% with the proposed model.

#### C. Organisation

The rest of the paper is organized as follows. We review the work related to Android malware detection in Section II. We discuss the proposed methodology in Section III. Section IV summarizes the results obtained from the proposed approach and we conclude the work in Section V.

## II. RELATED WORK

In this section, we have reviewed several studies for Android Malware Detection. We have categorized the related works into three types: Analysis based, Permissions based Android Malware Detection, and Android Malware Detection through different combinations.

#### A. Analysis Based Works

Some of the research works have analyzed permissions of normal apps without detecting the malware samples. In previous studies such as [7] and [8], researchers examined permissions as a means of detecting potentially malicious activity within normal applications. For instance, Grace et al. [7] investigated the potential risks associated with in-app advertisement libraries by scrutinizing permissions and API calls, while Kirin [8] developed a model for establishing security rules to detect risky applications based on permission combinations. Similarly, Holavanalli et al. [9] examined cross-app flow permissions to detect collusion between apps, while Grace et al. [10] identified permission leaks that exposed sensitive user data to other apps.

#### B. Permissions based Android Malware Detection

Alswaina et al.[11] created a reverse engineering framework wherein application permissions are selected, and then fed into machine learning algorithms. Li et al.[12] developed three levels of pruning by mining the permission data and then used SVM to classify different families of malware. The authors in [14] analyzed the permissions to classify malicious and benign apps using machine learning. The gain ratio was employed for feature reduction, and J48, Random Committee, Multilayer Perceptron, Sequential Minimal Optimization, etc. were used to evaluate the selected features. In this paper [18], authors applied two algorithms for permissions selection: information gain and chi-square, and the Bayesian method was used for classification.

The authors in [24] proposed a method in which a Random Forest classifier was utilized and an optimal subset of permissions was identified through the learned model. In the paper [27], a technique called Multilevel Permission Extraction was introduced which detected permission interactions that effectively distinguished between benign and malicious apps. Sokolova et al.[22] proposed a methodology for characterizing the normal behavior of applications. The co-required permissions were modeled as a graph and category patterns were obtained by the performance of an application in its category. Arora et al.[23] implemented an innovative detection model named PermPair which constructed and compared the graphs for normal and malware samples by extracting permission pairs. An efficient edge algorithm was also proposed which helped to eliminate unnecessary edges.

In this paper [15], the authors presented a new technique for automatically identifying permission patterns, which are groups of permissions that developers commonly use together, using SOM and K-means clustering. Kato et al. [16] proposed an Android malware detection technique based on the Composition Ratio (CR) of permission pairs. The CR was defined as the ratio of a permission pair to all pairs in an app. The authors constructed databases about the CR to obtain features without using the frequencies. Finally, eight similarity scores were calculated.

#### C. Android Malware Detection through different combinations

Some of the works aim to detect Android malware with dynamic traffic features such as [30-33], however, since we aim to detect Android malware with permissions, hence, we limit our discussion to the works that have analyzed permissions with other features. Park et al.[13] proposed a method consisting of three levels of analysis to classify applications into three categories namely, benign, suspicious, and malicious based on APIs and Permissions. The authors of [19] implemented an automated malware detection system, MalPat. In this approach, real-world Android app data was used for mining hidden patterns of malware and extracting sensitive APIs used in Android malware. Zhu et al.[21] employed ensemble forest rotation for constructing the model to detect malicious apps using permissions, sensitive APIs, and permission rates as key features.

The authors of [17] presented a framework called Droid-MalwareDetector, which utilized CNN. This framework was designed to automate feature extraction and the selection and utilized intents, API calls, and commonly used permissions to perform comprehensive malware analysis. This paper [25] proposed a new methodology named PIndroid which is a novel Permissions and Intents-based framework with Ensemble methods for identifying Android malware apps. Khariwal et al.[26] proposed a method that found the best set of permissions and intents combined that could give better accuracy using Information Gain. The authors in [28] developed a hybrid Android malware detector based on ranked permissions and network traffic features. The permissions had been ranked based on their frequency in normal and malware datasets along with the additional implementation of support thresholds to further remove the redundant permissions from the rankings. The authors in [20] presented two techniques for malware analysis, static and dynamic analysis. The static approach was based on permissions and the other approach was based on source code.

To the best of our knowledge, no past work has found a correlation between the permissions with the Cosine method and the Kulczynski measure. We have applied these techniques to find the correlation between permissions and further applied machine learning models for malware detection. We explain the proposed model in the next section.

### III. PROPOSED METHODOLOGY

In this section, we discuss the proposed methodology. Figure 1 shows the four phases involved in the methodology. We explain each of the four phases in the following subsections.

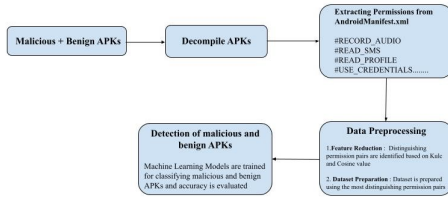


Fig. 1. Proposed Model

#### A. Data Collection

We collected Android applications from the AndroZoo dataset [29] which is a large repository of metadata related to Android applications, with the goal of facilitating Android-related research work. It currently includes over three million unique *apk* files, all of which have been scanned by dozens of Anti-Virus applications to determine which of them is identified as malware. Each application has over 20 categories of metadata, including Virus Total reports. We have compiled 12,000 malicious apps and 12,000 benign APKs.

#### B. Construction of Feature Set

Usually, the features of an application such as the permissions requested by an application are defined in the Android-Manifest.xml file. To extract these features, we used Apktool

which extracts the manifest file from the apk file. By default, an Android application begins with no permissions. When the app has to utilize any of the device's protected functions (sending network requests, accessing the camera, sending SMS messages, etc.), it must obtain the user's permission. Table I shows some examples of permissions requested by Android applications.

TABLE I  
VARIOUS TYPES OF PERMISSIONS REQUESTED BY APPLICATIONS

Permission Name	Description
ACCESS_WIFI_STATE	Allows applications to access information about WiFi networks
ADD_VOICEMAIL	Allows an application to add voicemails into the system
ANSWER_PHONE_CALLS	Allows the app to answer an incoming phone call
BLUETOOTH_CONNECT	Required to be able to connect to paired Bluetooth devices
CAMERA	Required to be able to access the camera device
READ_EXTERNAL_STORAGE	Allows an application to read from external storage
ACCESS_BACKGROUND_LOCATION	Allows an app to access location in the background

#### C. Data Preprocessing

1) *Feature Reduction*: Based on the data obtained from the previous step, there are 231 permissions and, therefore, 53130 permission pairs in total. However, certain permission pairs are not capable of making a better distinction between benign and malicious apps. Hence, in this phase, we have identified the most distinguishing permission pairs based on correlation scores. We aim to find the correlation or association between permissions. The association measures are used to analyze the relationship between different permissions. The correlation score between two permissions shows how much the permissions are associated with each other i.e. if an application contains permission X how much it is likely to contain permission Y and vice versa. To calculate the correlation value, we have used two association measures namely, Kulczynski and Cosine.

- **Kulczynski**- It is the arithmetic mean of the confidence of a permission pair. The confidence of one permission concerning another is defined as the probability of the occurrence of one permission given second permission occurring in the data. The range of kulczynski measure is from 0 to 1, the farther the value is from 0.5 the closer the relationship between two permissions.
- **Cosine** - Cosine is a harmonized version of a simple correlation coefficient as a square root is taken for the product of probabilities in the denominator. The range of cosine is between 0 to 1, the higher the value the closer the positive relationship between two permissions. The farther the value is from 0.5 the closer the relationship.

For every permission pair, their correlation score is calculated, both for malware and benign datasets. The permission pairs are then sorted based on the absolute value of their correlation score. For example, let's consider 3 applications, say A1, A2, A3, and four permissions P1, P2, P3, and P4, as summarized in Table II. In the matrix so obtained, value 1 corresponds to the fact that the permission P<sub>i</sub> is requested by application A<sub>j</sub> and 0 represents its absence.

TABLE II  
SNAPSHOT OF THE DATASET REPRESENTATION

App/Permission	Permission 1	Permission 2	Permission 3	Permission 4
App 1	1	0	1	1
App 2	1	1	0	1
App 3	0	1	0	1

The absolute difference column in Table III shows how much distinguishing the permission pair is. If the permission pair is highly correlated or not correlated at all in both normal and malware apps then it doesn't distinguish between the two types of applications and hence can be discarded. If the permission pair is highly correlated in one of the 2 classes of applications, then it helps in distinguishing between the apps and hence should be considered. The above example shows that pairs P1-P3 and P2-P3 can effectively distinguish between normal and malware applications.

TABLE III  
PERMISSION PAIRS AND THEIR CORRELATION

Permission Pair	Correlation Score		
	Normal Apps	Malware Apps	Abs. Difference
P1 - P2	1	0.8	0.2
P1 - P3	0.2	0.9	0.7
P1 - P4	0.7	0.6	0.1
P2 - P3	0	0.9	0.9
P2 - P4	1	0.4	0.6
P3 - P4	0.9	0.5	0.4

a

2) *Dataset Preparation*: The pairs obtained from the previous step are sorted according to their correlation score difference. The permission pairs with the highest correlation difference are the most distinguishing and hence are selected according to a set threshold. These distinguishing permission pairs are used in this phase to preprocess the data. Two separate datasets have been prepared, corresponding to malware and normal applications respectively. Each column in the dataset corresponds to a permission pair while the rows correspond to the application. For every permission pair corresponding to an application, values are assigned based on the absence and presence of permissions. If both permissions corresponding to the pair are present in the application, then value 1 is assigned to the cell in the matrix. If both the permissions are absent, then value 0 is assigned. If either of the permissions is present, then value 2 is assigned. For example, in reference to the example quoted in the previous step, the threshold value is set as 0.5 and hence all the pairs having differences greater than and equal to 0.5 are selected. Table IV shows the dataset obtained using the selected permission pairs.

#### D. Classification and Evaluation

For the detection of malicious applications, we have used supervised machine learning techniques such as Naïve Bayes, Decision Trees, and SVM. During this phase, two steps are involved, namely training and testing the machine learning models using the preprocessed dataset obtained from the

TABLE IV  
DATASET OBTAINED AFTER PERFORMING PREPROCESSING

Application (Benign)	Permission Pair		
	P1 - P3	P2 - P3	P3 - P4
A1	1	2	2
A2	2	2	1
A3	0	2	1

previous phase. Initially, we performed splitting on the dataset according to the 70-30 rule, and then preprocessing was performed on the training data. The machine learning models are trained using this data and the performance of the models is evaluated using the testing data after preprocessing them.

## IV. RESULTS AND DISCUSSION

In this section, we review the results obtained from our proposed model. We performed all the experiments on a Windows system with 8 GB RAM and an i5 processor. We downloaded the apps for the dataset from Androzoo and extracted permissions using apktool and a Python script to create a CSV of permissions for benign and malware applications respectively.

### A. Analysis of Permission Pairs

In order to perform an analysis we considered a total of 24,140 Android applications, of which 12,070 were malware, and 12,070 were benign applications. A total of 8000 apks from each category are used to train our model and around 4000 apks from both categories are used to test the model. 129 permissions have been extracted from these apks for analysis. The correlation between different permissions in benign and malware applications has been determined using Kulczynski and Cosine coefficients.

Tables V and VI present permission pairs that are highly correlated in malware and normal applications, using kulczynski coefficient, respectively. Tables VII and VIII present permission pairs that are highly correlated in malware and normal applications using cosine coefficients, respectively. The following tables illustrate that some permission pairs are highly correlated in both types of applications while some are correlated in one of the two classes.

TABLE V  
HIGHLY CORRELATED PERMISSION PAIRS IN NORMAL APPLICATIONS (KULCZYNSKI)

First Permission Name	Second Permission Name	Correlation Score
BADGE_COUNT_WRITE	BADGE_COUNT_READ	1.0
WRITE	READ	0.99348
UPDATE_BADGE	UPDATE_COUNT	0.98923
UPDATE_SHORTCUT	UPDATE_COUNT	0.98259
BROADCAST_BADGE	UPDATE_SHORTCUT	0.98101
WRITE	UPDATE_SHORTCUT	0.97950
INTERNET	ACCESS_NETWORK_STATE	0.97716
BROADCAST_PACKAGE_CHANGED	BROADCAST_PACKAGE_REPLACED	0.97560
READ	UPDATE_SHORTCUT	0.97444
WRITE	BROADCAST_BADGE	0.97317
UPDATE_BADGE	UPDATE_SHORTCUT	0.97285

TABLE VI  
HIGHLY CORRELATED PERMISSION PAIRS IN MALWARE APPLICATIONS  
(KULCZYNSKI)

First Permission Name	Second Permission Name	Correlation Score
BADGE_COUNT_WRITE	BADGE_COUNT_READ	1.0
BROADCAST_PACKAGE_CHANGED	BROADCAST_PACKAGE_REPLACED	0.99936
BROADCAST_PACKAGE_ADDED	BROADCAST_PACKAGE_INSTALL	0.99932
INTERNET	ACCESS_NETWORK_STATE	0.99663
WRITE_EXTERNAL_STORAGE	INTERNET	0.99442
WRITE	READ	0.99442
WRITE_EXTERNAL_STORAGE	ACCESS_NETWORK_STATE	0.99293
ACCESS_WIFI_STATE	ACCESS_NETWORK_STATE	0.99084
ACCESS_WIFI_STATE	INTERNET	0.98870
WRITE_EXTERNAL_STORAGE	ACCESS_WIFI_STATE	0.98714
READ_PHONE_STATE	ACCESS_NETWORK_STATE	0.98329

TABLE VII  
HIGHLY CORRELATED PERMISSION PAIRS IN NORMAL APPLICATIONS  
(COSINE)

First Permission Name	Second Permission Name	Correlation Score
BADGE_COUNT_WRITE	BADGE_COUNT_READ	1.0
WRITE	READ	0.99346
UPDATE_BADGE	UPDATE_COUNT	0.98919
UPDATE_SHORTCUT	UPDATE_COUNT	0.98244
BROADCAST_BADGE	UPDATE_SHORTCUT	0.98093
WRITE	UPDATE_SHORTCUT	0.97946
INTERNET	ACCESS_NETWORK_STATE	0.97693
BROADCAST_PACKAGE_CHANGED	BROADCAST_PACKAGE_REPLACED	0.97530
READ	UPDATE_SHORTCUT	0.97435
WRITE	BROADCAST_BADGE	0.97316
UPDATE_BADGE	UPDATE_SHORTCUT	0.97251

TABLE VIII  
HIGHLY CORRELATED PERMISSION PAIRS IN MALWARE APPLICATIONS  
(COSINE)

First Permission Name	Second Permission Name	Correlation Score
BADGE_COUNT_WRITE	BADGE_COUNT_READ	1.0
BROADCAST_PACKAGE_CHANGED	BROADCAST_PACKAGE_REPLACED	0.99936
BROADCAST_PACKAGE_ADDED	BROADCAST_PACKAGE_INSTALL	0.99932
INTERNET	ACCESS_NETWORK_STATE	0.99663
WRITE	READ	0.99442
WRITE_EXTERNAL_STORAGE	INTERNET	0.99442
WRITE_EXTERNAL_STORAGE	ACCESS_NETWORK_STATE	0.99293
ACCESS_WIFI_STATE	ACCESS_NETWORK_STATE	0.99081
ACCESS_WIFI_STATE	INTERNET	0.98864
WRITE_EXTERNAL_STORAGE	ACCESS_WIFI_STATE	0.98712
READ_PHONE_STATE	ACCESS_NETWORK_STATE	0.98319

As can be seen from the above tables, many pairs are present as well as highly correlated in malicious as well as benign applications. As we cannot distinguish between benign and malicious applications using these pairs, they are not useful. In order to identify distinguishing permission pairs, we calculated the difference in correlation scores of benign applications and malware applications for different permission pairs. Table IX and Table X present the permission pairs with the highest correlation difference, these pairs are the most distinguishing and can be used to differentiate between normal and malware applications.

TABLE IX  
MOST DISTINGUISHING PERMISSION PAIRS (KULCZYNSKI)

First Permission Name	Second Permission Name	Correlation Score
USE_CREDENTIALS	SET_DEBUG_APP	0.65473
CALL_PHONE	MOUNT_UNMOUNT_FILESYSTEMS	0.57151
ACCESS_FINE_LOCATION	MOUNT_UNMOUNT_FILESYSTEMS	0.56524
MOUNT_UNMOUNT_FILESYSTEMS	ACCESS_COARSE_LOCATION	0.56040
ACCESS_LOCATION_EXTRA_COMMANDS	CHANGE_WIFI_STATE	0.55005
GET_TASKS	ACCESS_COARSE_LOCATION	0.54712
MOUNT_UNMOUNT_FILESYSTEMS	ACCESS_LOCATION_EXTRA_COMMANDS	0.54481
ACCESS_LOCATION_EXTRA_COMMANDS	CHANGE_NETWORK_STATE	0.53367
ACCESS_FINE_LOCATION	GET_TASKS	0.52613
GET_TASKS	ACCESS_LOCATION_EXTRA_COMMANDS	0.52463
GET_TASKS	WRITE_SETTINGS	0.52067

TABLE X  
MOST DISTINGUISHING PERMISSION PAIRS (COSINE)

First Permission Name	Second Permission Name	Correlation Score
USE_CREDENTIALS	SET_DEBUG_APP	0.68032
ACCESS_FINE_LOCATION	MOUNT_UNMOUNT_FILESYSTEMS	0.67361
MOUNT_UNMOUNT_FILESYSTEMS	ACCESS_COARSE_LOCATION	0.66665
MOUNT_UNMOUNT_FILESYSTEMS	INTERNET	0.63522
MOUNT_UNMOUNT_FILESYSTEMS	ACCESS_NETWORK_STATE	0.63374
WRITE_EXTERNAL_STORAGE	MOUNT_UNMOUNT_FILESYSTEMS	0.62147
VIBRATE	MOUNT_UNMOUNT_FILESYSTEMS	0.61912
MOUNT_UNMOUNT_FILESYSTEMS	READ_PHONE_STATE	0.61708
MOUNT_UNMOUNT_FILESYSTEMS	WAKE_LOCK	0.61263
MOUNT_UNMOUNT_FILESYSTEMS	CAMERA	0.60960
READ_EXTERNAL_STORAGE	MOUNT_UNMOUNT_FILESYSTEMS	0.60924

### B. Filtration of Permission Pairs

We have considered the permission pairs with the highest difference in correlation scores in benign and malware apps. We used different thresholds to filter out the permission pairs and then use them as features to train the model. We first find the most distinguishing permission pairs by setting a threshold and only selecting the permission pairs whose correlation score difference is greater than the threshold. As we increase the threshold, the number of filtered permission pairs decreases. After preprocessing the filtered permission pairs, we train our machine learning models with the data and find out the accuracy of our proposed algorithm.

### C. Detection Results

Tables XI and XII summarize the results obtained using different machine-learning models and different thresholds. In the case of Kulczynski coefficient, we get the best results from SVM polynomial classifier which gives 97.2% accuracy with 677 permission pairs. Decision Trees give the maximum accuracy with 677 permission pairs. Naïve Bayes gives the best results (91.1% accuracy) with 850 permission pairs. Similarly, using the cosine coefficient, we get the best results from SVM polynomial classifier which gives 96.8 % accuracy with 688 permission pairs. Decision Trees give the maximum accuracy with 576 permission pairs. Naïve Bayes gives the best results (91.9%) with 872 permission pairs. On adding or removing more pairs the accuracy of our algorithm starts decreasing and hence our algorithm gives the best result accuracy of 97.2% with 677 significant permission pairs.

TABLE XI  
RESULTS (KULCZYNSKI)

Models/Thresholds	(0.30)	(0.32)	(0.34)	(0.36)	(0.38)	(0.40)
No of pairs	850	677	543	445	349	257
Naive Bayes (bernoulli)	90.1	89.9	89.6	89.3	89.6	89.5
Naive Bayes (Gaussian)	91.1	90.8	90.5	89.9	90.1	89.8
Decision Tree (entropy)	95.8	95.7	94.9	95.1	94.5	94.3
Decision Tree (gini)	96.0	96.1	95.5	95.3	95.6	95.1
Random Forest	95.9	96.0	95.6	96.0	95.6	95.4
SVC (linear)	96.3	96.5	96.4	96.2	96.2	96.1
SVC (poly)	97.1	97.2	96.8	97.0	96.8	96.6

TABLE XII  
RESULTS (COSINE)

Models/Thresholds	(0.26)	(0.28)	(0.30)	(0.32)	(0.34)	(0.36)
No of pairs	872	688	576	475	395	304
Naive Bayes (bernoulli)	91.3	90.8	90.7	90.4	90.5	90.1
Naive Bayes (Gaussian)	91.9	91.8	91.5	91.7	91.2	90.9
Decision Tree (entropy)	95.4	95.5	95.6	95.3	95.5	95.4
Decision Tree (gini)	95.4	95.6	95.9	95.5	95.5	95.4
Random Forest	95.8	95.6	95.8	95.7	95.7	95.4
SVC (linear)	96.3	96.3	96.3	96.0	96.1	96.0
SVC (poly)	96.5	96.8	96.7	96.2	96.5	96.4

## V. CONCLUSION AND FUTURE WORK

In this work, we proposed a novel Android malware detection solution named CORRDroid that extracts permissions from the applications and ranks the permission pairs based on their correlation score difference in benign and malware applications using Kulczynski and Cosine coefficients. The proposed system classifies the applications using top permission pairs which are highly distinguishing. The results proved that the proposed methodology gives better results in comparison to the approach when all the permissions are used for the classification of applications. In our future work, we will incorporate other attributes as well like intents, hardware features, and other manifest file components.

## REFERENCES

- [1] Christo Petrov, "51 Mobile vs. Desktop Usage Statistics For 2023" <https://techjury.net/blog/mobile-vs-desktop-usage/>, 2023
- [2] J.Callahan, "The history of Android: The evolution of the biggest mobile OS in the world" <https://www.androidauthority.com/history-android-os-name-789433/>, 2022
- [3] B.Curry,"Report: Android Statistics 2021" <https://www.businessofapps.com/data/android-statistics/>, 2022
- [4] B.Carlton, "Report: Android Statistics 2021" <https://www.businessofapps.com/data/android-statistics/>, 2021
- [5] A.Sharma, "Report: Top Google Play Store Statistics 2022 You Must Know" <https://appinventiv.com/blog/google-play-store-statistics/>, 2022
- [6] L. Ceci, "Report: Number of apps available in leading app stores Q3 2022" <https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/>, 2022
- [7] M.C. Grace, W.Zhou, X.Jiang, and A.R. Sadeghi, "Unsafe exposure analysis of mobile in-app advertisements", in Proceedings of the fifth ACM conference on Security and Privacy in Wireless and Mobile Networks, ACM, New York, NY, USA, 101–112.
- [8] W. Enck, M. Ongtang, and P. McDaniel, On Lightweight Mobile Phone, on lightweight mobile phone application certification, in Proceedings of the 16th ACM conference on Computer and communications security, ACM, New York, NY, USA, 235–245.
- [9] S. Holavanalli et al., "Flow Permissions for Android," 2013 28th IEEE/ACM International Conference on Automated Software Engineering (ASE), Silicon Valley, CA, USA, 2013, pp. 652–657.
- [10] M. Grace, Y. Zhou, Z. Wang, and X. Jiang, Systematic Detection of Capability Leaks in Stock Android Smartphones, NDSS, 2012.
- [11] F. Alswaina and K. Elleithy, "Android Malware Permission-Based Multi-Class Classification Using Extremely Randomized Trees," in IEEE Access, vol. 6, pp. 76217–76227, 2018.
- [12] J. Li. et al., "Significant Permission Identification for Machine-Learning-Based Android Malware Detection," in IEEE Transactions on Industrial Informatics, vol. 14, no. 7, pp. 3216–3225, July 2018.
- [13] J. Park, H. Chun and S. Jung, "API and permission-based classification system for Android malware analysis," 2018 International Conference on Information Networking (ICOIN), 2018, pp. 930–935.
- [14] S. J. K., S. Chakravarty and R. K. Varma P., "Feature Selection and Evaluation of Permission-based Android Malware Detection," 4th International Conference on Trends in Electronics and Informatics, 2020, pp. 795–799.
- [15] Z. Namrud, S. Kpodjedo, A. Bali and C. Talhi, "Deep-Layer Clustering to Identify Permission Usage Patterns of Android App Categories," in IEEE Access, vol. 10, pp. 24240–24254, 2022.
- [16] H. Kato, T. Sasaki and I. Sasase, "Android Malware Detection Based on Composition Ratio of Permission Pairs," in IEEE Access, vol. 9, pp. 130006–130019, 2021.
- [17] A. T. Kabakus, "DroidMalwareDetector: A novel Android malware detection framework based on convolutional neural network", Expert Systems with Applications, Volume 206, 2022, 117833,ISSN 0957-4174.
- [18] S. R. T. Mat et al., "A Bayesian probability model for Android malware detection", ICT Express, Volume 8, Issue 3, 2022, Pages 424–431.
- [19] G. Tao, Z. Zheng, Z. Guo and M. R. Lyu, "MalPat: Mining Patterns of Malicious and Benign Android Apps via Permission-Related APIs," in IEEE Transactions on Reliability, vol. 67, no. 1, pp. 355–369, 2018.
- [20] N. Milosevic, A. Dehghantanha and K.-K. R. Choo, "Machine learning aided Android malware classification", Comput. Elect. Eng., vol. 61, pp. 266–274, Jul. 2017.
- [21] H.-J. Zhu et al., "DroidDet: Effective and robust detection of Android malware using static analysis along with rotation forest model", Neuro-computing, vol. 272, pp. 638–646, Jan. 2018.
- [22] K. Sokolova, C. Perez and M. Lemercier, "Android application classification and anomaly detection with graph-based permission patterns", Decis. Support Syst., vol. 93, pp. 62–76, Jan. 2016.
- [23] A. Arora, S. K. Peddoju and M. Conti, "PermPair: Android Malware Detection Using Permission Pairs," in IEEE Transactions on Information Forensics and Security, vol. 15, pp. 1968–1982, 2020.
- [24] J. Park et al., "Analysis of Permission Selection Techniques in Machine Learning-based Malicious App Detection," 2020 Third International Conference on Artificial Intelligence and Knowledge Engineering, pp. 92–99.
- [25] F. Idrees et al., "PIndroid: A novel Android malware detection system using ensemble learning methods", Computers & Security, Volume 68, 2017, Pages 36–46.
- [26] K. Khariwal, J. Singh and A. Arora, "IPDroid: Android Malware Detection using Intents and Permissions," Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), 2020, pp. 197–202.
- [27] Z. Wang et al., "Multilevel Permission Extraction in Android Applications for Malware Detection," 2019 International Conference on Computer, Information and Telecommunication Systems, pp. 1–5.
- [28] M. Upadhayay, A. Sharma, G. Garg and A. Arora, "RPNDroid: Android Malware Detection using Ranked Permissions and Network Traffic," 2021 Fifth World Conference on Smart Trends in Systems Security and Sustainability (WorldS4), 2021, pp. 19–24.
- [29] K. Allix et al., "AndroZoo: Collecting Millions of Android Apps for the Research Community," IEEE/ACM 13th Working Conference on Mining Software Repositories, 2016, pp. 468–471.
- [30] A. Arora, S. K. Peddoju and S. K. Peddoju, "Malware Detection Using Network Traffic Analysis in Android Based Mobile Devices," 2014 Eighth International Conference on Next Generation Mobile Apps, Services and Technologies, Oxford, UK, pp. 66–71, 2014.
- [31] A. Arora and Sateesh K. Peddoju, "Minimizing Network Traffic Features for Android Mobile Malware Detection", In Proceedings of the 18th International Conference on Distributed Computing and Networking, ACM, New York, NY, USA, Article 32, 1–10, 2017.
- [32] A. Arora and S. K. Peddoju, "NTPDroid: A Hybrid Android Malware Detector Using Network Traffic and System Permissions," 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering, New York, USA, pp. 808–813, 2018.
- [33] A. Arora, S. K. Peddoju, V. Chouhan, and A. Chaudhary, "Hybrid Android Malware Detection by Combining Supervised and Unsupervised Learning", In Proceedings of the 24th Annual International Conference on Mobile Computing and Networking, ACM, New York, USA, 798–800, 2018.



# Deep Convolutional Neural Network With Attention Module for Seismic Impedance Inversion

Vineela Chandra Dodda<sup>✉</sup>, *Member, IEEE*, Lakshmi Kuruguntla<sup>✉</sup>, *Member, IEEE*,  
Anup Kumar Mandpura<sup>✉</sup>, *Member, IEEE*, Karthikeyan Elumalai<sup>✉</sup>, *Member, IEEE*,  
and Mrinal K. Sen<sup>✉</sup>, *Member, IEEE*

**Abstract**—Seismic inversion is an approach to obtain the physical properties of the Earth layers from the seismic data, which aids in reservoir characterization. In seismic inversion, spatially variable physical parameters, such as impedance ( $Z$ ), wave velocities ( $V_p$ ,  $V_s$ ), and density, can be determined from the seismic data. Among these, impedance is an important parameter used for lithology interpretation. However, the inversion problem is nonlinear and ill-posed due to unknown seismic wavelet, observed data band limitation, and noise. This requires complex wave equation analysis, prior assumptions, human expert effort, and time to analyze the seismic data. To address these issues, deep learning methods were deployed to solve the seismic inversion problem. In this article, we develop a deep learning framework with an attention module for seismic impedance inversion. The relevant features from the seismic data are emphasized with the integration of the attention module into the network. First, we train the attention-based deep convolutional neural network (ADCNN) by supervised learning with predefined acoustic impedance (AI) labels. Next, we train the ADCNN in an unsupervised way with the physics of the forward problem. In the proposed method, the predicted AI is used to calculate the seismic data (calculated seismic), and error is minimized between the input seismic data and calculated seismic data. Unsupervised learning has an advantage when the labeled data are inadequate. The proposed network is trained with Marmousi 2 dataset, and the predicted experimental results show that the proposed method outperforms in comparison to the existing state-of-the-art method.

**Index Terms**—Attention module, impedance inversion, neural networks, seismic data.

## I. INTRODUCTION

THE seismic reflection method is an effective method used to get the Earth subsurface layers information. In seismic

reflection method, a pulse with short duration (seismic wavelet) is sent from the Earth's surface. The pulse penetrates inside the subsurface layers of the Earth to a certain depth. However, due to the impedance contrast between the adjacent layers, the waves are reflected from layer boundaries, and the reflected waves are recorded at the Earth's surface. The recorded seismic data contains information about the seismic source, reflection coefficients and noise. The reflection coefficients have information about the Earth layers, because it is derived from the layer impedance. To obtain the reflection coefficients from the seismic data, the data undergoes several processing steps, such as deconvolution, denoising, and NMO correction [1]. After these processing steps, the seismic data retains only reflection coefficients.

The processed seismic data have amplitude and time information which gives only structural interpretation. Therefore, to obtain stratigraphic interpretation and reservoir characterization, we need to inverse the seismic data/reflection coefficient that gives the physical parameters of the layers. It is known as inverse modeling (seismic inversion). The seismic inversion retrieves the physical properties of the Earth layers from the seismic reflection data. In seismic inversion process, spatially variable physical parameters, such as layer impedance ( $Z$ ), P-wave ( $V_p$ ), S-Wave ( $V_s$ ) velocity and density, porosity, sand/shale formation, and gas saturation, are estimated from the seismic data. These parameters have physical and geological meaning about the Earth subsurface layers, which helps in reservoir characterization [2], [3]. Seismic impedance inversion, AVO inversion, and full waveform inversion are the commonly used seismic inversion methods, which helps to obtain the Earth subsurface properties. Among all these methods, seismic impedance inversion method is extensively used in the seismic industry and is an important goal in reflection seismology. Seismic impedance inversion is a powerful method for the Earth subsurface layers analysis, reservoir characterization, and fluid prediction. The impedance is a rock property, which gives information about lithology, porosity, and other factors [4]. However, the impedance inversion problem is usually ill-posed, nonlinear, and nonunique because of unknown seismic wavelet, noises, and band limited nature of observed seismic data [5]. All these issues are to be taken into consideration while solving an inverse problem.

Since 1960s, researchers have put forward many seismic impedance inversion methods, which are categorized based on poststack and prestack seismic data [6]. In the poststack

Manuscript received 18 May 2023; revised 12 August 2023; accepted 19 August 2023. Date of publication 29 August 2023; date of current version 12 September 2023. This work was supported by the Department of Science and Technology, Science and Engineering Research Board (DST-SERB), India, through Core Research under Grant CRG/2019/001234. (Corresponding author: Karthikeyan Elumalai.)

Vineela Chandra Dodda and Karthikeyan Elumalai are with the Department of Electronics and Communication Engineering, SRM University, Amaravathi 522502, India (e-mail: vineelachandra\_dodda@srmap.edu.in; imkarthi@gmail.com).

Lakshmi Kuruguntla is with the Department of Electronics and Communication Engineering, Koneru Lakshmaiah Educational Foundation (KLEF), Vaddeswaram 5223002, India (e-mail: lakshmi\_kuruguntla@srmap.edu.in).

Anup Kumar Mandpura is with the Department of Electrical Engineering, Delhi Technological University, New Delhi 110042, India (e-mail: amandpura@gmail.com).

Mrinal K. Sen is with the Department of Geological Sciences, Jackson School of Geosciences, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: mrinal@utexas.edu).

Digital Object Identifier 10.1109/JSTARS.2023.3308751

inversion method, the acoustic impedance is estimated from the seismic data by integrating well data and the basic stratigraphic interpretation. However the prestack inversion methods transform the seismic angle/offset into P-impedance, S-impedance, and layer's density by integration of well data and horizon information. Further, the poststack inversion methods are divided into two types: deterministic and stochastic inversion methods [7]. The deterministic inversion methods are based on optimization methods, which can provide a good fitting model. These optimization methods aim to minimize the error between synthetic and observed seismic data. These methods produce smooth models but the uncertainties about the predicted values are not assessed [8]. Most commonly used deterministic inversion methods are band limited recursive inversion [9], colored inversion [10], and sparse spike inversion methods [11]. However, the stochastic seismic inversion methods retrieve the best-fit inverse model from the seismic data based on the probability density function of the data, which helps to assess the uncertainties [12]. Most commonly used prestack inversion methods are simultaneous inversion [13], elastic impedance inversion [14], and AVO inversion methods [15], [16], [17].

Nevertheless, the conventional methods used in seismic inversion have some limitations, such as complex wave equation analysis, longer simulation times, and more human expert effort to analyze the seismic data. Moreover, the conventional methods incur convergence issues and high computational cost. Hence, to solve those limitations, researchers have come up with ideas to use artificial intelligence techniques in various geophysics problems, such as fault interpretation [18], seismic data denoising [19], seismic horizon estimation [20], seismic inversion [21], and so on [22], [23]. Deep learning (DL) is the subpart of machine learning that has prominence and applicability in wide areas of science and engineering [22], [24]. The DL has lots of scope to explore the seismic data for various applications, such as denoising, seismic inversion, and interpretation. In contrast to conventional seismic inversion methods, DL methods does not necessarily require the forward operator or wavelet matrix explicitly [25]. In [22], the Earth surface elastic model is estimated (seismic inversion) from seismic data using convolutional neural network. The robustness of the network to predict P-impedance of seismograms is tested and has shown good accuracy for seismic data generated with source wavelet phase outside the training data. However, the CNN was unable to predict P-impedance for the seismic data generated with various wavelet frequencies. Further in [26], temporal convolutional network (TCN) was proposed to estimate seismic impedance from the seismic data. TCN network is a combination of both RNN and CNN, which overcomes the limitation of overfitting in CNN and gradient vanishing in RNN [27], [28]. Moreover, long-term and short-term dependencies are captured by the network without the need of large number of learnable parameters. Later in [29], fully convolutional residual network (FCRN) with transfer learning approach was used for seismic impedance inversion. Although, FCRN has shown good accuracy and robustness against noise and phase difference of the seismic data, but the results were not accurate when tested with seismic data of different geological features. Hence, the authors proposed transfer

learning approach, i.e., the parameters of FCRN trained on Marmousi 2 data were used as initialization for a new FCRN. In the next step, FCRN is trained with traces from overthrust model and tested the performance of FCRN. However, the major obstacle is in the availability of labeled data. Hence, researchers worked on alternative approaches to predict impedance with the limited usage of labeled seismic data. Therefore in [30], physics constrained seismic impedance inversion method was proposed based on DL where 2-D bilateral filtering constraint was proposed to improve the spatial continuity of the inversion results. In addition, it also reduces the nonuniqueness of the inversion problem. Later in [31], cycle-consistent generative adversarial network (CCGAN) was used for seismic impedance inversion. The CCGAN extracts information contained in the unlabeled data and in addition adversarial learning helps in better prediction rate. Moreover, a neural network visualization method was adopted to visualize the features learned from the trained model and compared with conventional open-loop CNN model. However, CC-GAN suffers from training instability like most of the GAN models. Hence in [32], Wasserstein cycle-consistent GAN-based network was proposed. Here, the authors improved the CCGAN with integration of Wasserstein loss with gradient penalty as the loss function. The network was tested on the 3-D seismic advanced modeling data.

However, in the field of geophysics, geological information is in nature multiscale in seismic data. The extracted feature by the CNN kernel plays different roles for different tasks. Different feature maps (FPs) obtained from various kernels acquire a variety of different features, which together contribute to accurate results. Attention focuses on processing these informations to achieve better accuracy under limited resources. Hence, we propose to integrate the attention module with the CNN and improve the accuracy of network model for the estimation of acoustic impedance. A block attention module was integrated into the network to extract features from the two dimensions; channel and spatial axes. The channel attention module emphasizes “what” features need to be extracted from the input data whereas spatial attention module says from “where” the feature has to be extracted in the input data. Therefore, the attention module helps for efficient information flow in the network thus leading to better representation power. We applied the attention module for two cases: supervised and unsupervised. In supervised case, we used true labels of AI and estimated the optimum parameters to predict impedance. In unsupervised case, the input to the network is seismic data, wavelet and low-frequency model of seismic data. Here, we do not use true AI labels. The unsupervised learning (UL) method finds applications where there is no labeled data. We demonstrate the effectiveness of the proposed method in each of these cases. In our work, we considered the P-impedance inversion in all the cases. The contributions of this article are as follows.

- 1) We introduced an attention mechanism to improve the CNN performance, i.e., convolutional block attention module (CBAM) and allow the neural network to focus on certain regions of an image that are most relevant to the task. The combination of both the channel and spatial attention blocks allows CBAM to selectively focus on the

most important channels and spatial locations within an FP, allowing the CNN to effectively capture both local and global contextual information.

- 2) Our study involved an analysis of two different approaches for learning network parameters: supervised and UL methods. In cases, where labeled data are available, supervised learning can be employed. This approach can also be utilized in transfer learning, where a pretrained model (obtained through supervised learning) is initialized before training the network in an unsupervised way. On the other hand, UL is employed when labels are not present. Consequently, this article encompasses both supervised and UL methodologies.
- 3) We used a novel activation function scaled exponential linear unit (SELU) during the training process. The advantages of the SELU activation function are improved performance, self-normalization, stability, and efficiency. Moreover, we used Bayesian optimization (BO) tuner to optimize the hyperparameters, which resulted in time saving when compared to manual tuning.

The rest of this article is organized as follows. In Section II, we discuss the methodology, which contains mathematical model formulation in Section II-A and we present the proposed method in Section II-B. In Section III, we illustrate and analyze the results of existing and proposed method. Finally, Section IV concludes this article.

## II. METHODOLOGY

### A. Mathematical Model for Impedance Inversion

In this section, we formulate the mathematical model of the impedance inversion problem [1]. According to convolutional model, the seismic trace is modeled as

$$S = W * r + \text{noise} \quad (1)$$

where  $S \in R^n$  is the seismic trace,  $W \in R^{n \times n}$  is the seismic wavelet, and  $r \in R^n$  is the normal incidence reflection coefficient, which can be represented in terms of impedance  $z$  as

$$r = \frac{z(t+1) - z(t)}{z(t+1) + z(t)} \quad (2)$$

where  $z = v\rho$  in which  $v$  is the velocity,  $\rho$  is the density, and  $t$  is the layer number. The extraction of reflection coefficients from the seismic trace is viewed as an inverse problem, given the seismic trace and wavelet information. In general, the inverse problems are nonunique and ill posed whereas in the seismic data, this is due to the band limited nature of the wavelet. Therefore, the recorded seismic traces  $S$  is band limited (low and high frequencies are filtered by the wavelet). Hence, we add constraints, such as sparse reflectivity series, known wavelet, and a low-frequency model, to obtain a unique solution for inverse problem. Let  $s_i$  denote the  $i$ th seismic trace of length  $M$ . The group of  $N$  seismic traces  $\{s_1, s_2, s_3, \dots, s_N\}$  is expressed as

$$S = [s_1 \ s_2 \ \dots \ s_N].$$

Let  $z_i$  be the corresponding acoustic impedance traces

$$Z = [z_1 \ z_2 \ \dots \ z_N].$$

Here, we considered the system to be noise-free. The (1) is reduced to

$$S = W * r. \quad (3)$$

### B. Proposed Method

In this section, we first describe the attention module. Second, we discuss the proposed method with supervised learning approach for seismic impedance inversion and third, we describe the UL approach for seismic impedance inversion.

1) *Attention Module*: In this module, visual system of humans was taken as inspiration where we use an attention mechanism, i.e., series of glimpses to focus on main scenes rather than processing the whole scene for better visualization. Similarly, we added the attention block in the CNN architecture to better capture the features from the input data. The extracted FP is given as an input to the attention module, which further helps to obtain features from both channelwise and spatial attention based on CNN architecture [33]. The FP obtained from hidden layer say  $L \in \mathbb{R}^{C \times H \times W}$  is given as input to the attention module, then it outputs channel attention map  $A_c \in \mathbb{R}^{C \times 1 \times 1}$  and spatial attention map  $A_s \in \mathbb{R}^{1 \times H \times W}$ . The attention process is given as follows:

$$\begin{aligned} L' &= A_c(L) \otimes L \\ L'' &= A_s(L') \otimes L'. \end{aligned} \quad (4)$$

The channel attention values are copied to the spatial dimension and vice-versa in (4).  $L''$  is the final output from the spatial attention module. In channel attention module, FPs are created based on interchannel relationship between the features. The spatial dimension of the input FP is squeezed using both average pooled and max pooling to improve the representation power of networks. The obtained average pooled and max pooled features are  $L_c^{\text{avg}}$  and  $L_c^{\text{max}}$ , respectively. These features are passed to a network which has one hidden layer of size set to  $\mathbb{R}^{c/k \times 1 \times 1}$ , where  $k$  is the reduction ratio. The output feature vectors are merged using elementwise summation given as

$$A_c(L) = \sigma(Z_1(Z_0(L_c^{\text{avg}})) + Z_1(Z_0(L_c^{\text{max}}))) \quad (5)$$

where  $\sigma$  is the sigmoid function,  $Z_0$  and  $Z_1$  are the weights and SELU activation function is used after  $Z_0$ . In spatial attention module, FPs are obtained using interspatial relationship between the features which focuses on “where” is the informative part. Here, the max pooling and average pooling are applied to the channel axis and the obtained feature descriptors are concatenated. The spatial attention map  $A_s(L) \in \mathbb{R}^{H \times W}$  is created after passing the concatenated descriptors through a convolution layer

$$A_s(L) = \sigma(f^{3 \times 3}([L_{\text{avg}}^s; L_{\text{max}}^s])) \quad (6)$$

where  $f^{3 \times 3}$  is a convolution operation with size  $3 \times 3$  and  $\sigma$  is an activation function “sigmoid.” These channel and spatial attention modules can be arranged in series or parallel. In our case, sequential arrangement has shown better accuracy than the

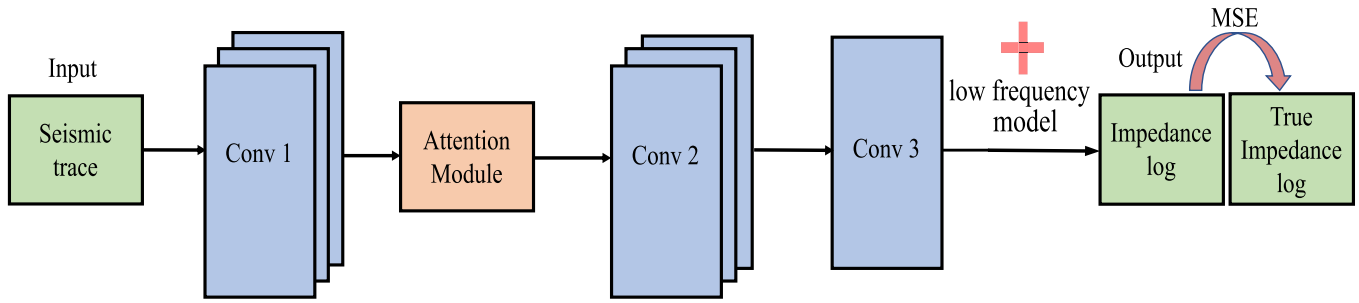


Fig. 1. Network architecture of ADCNN for supervised learning.

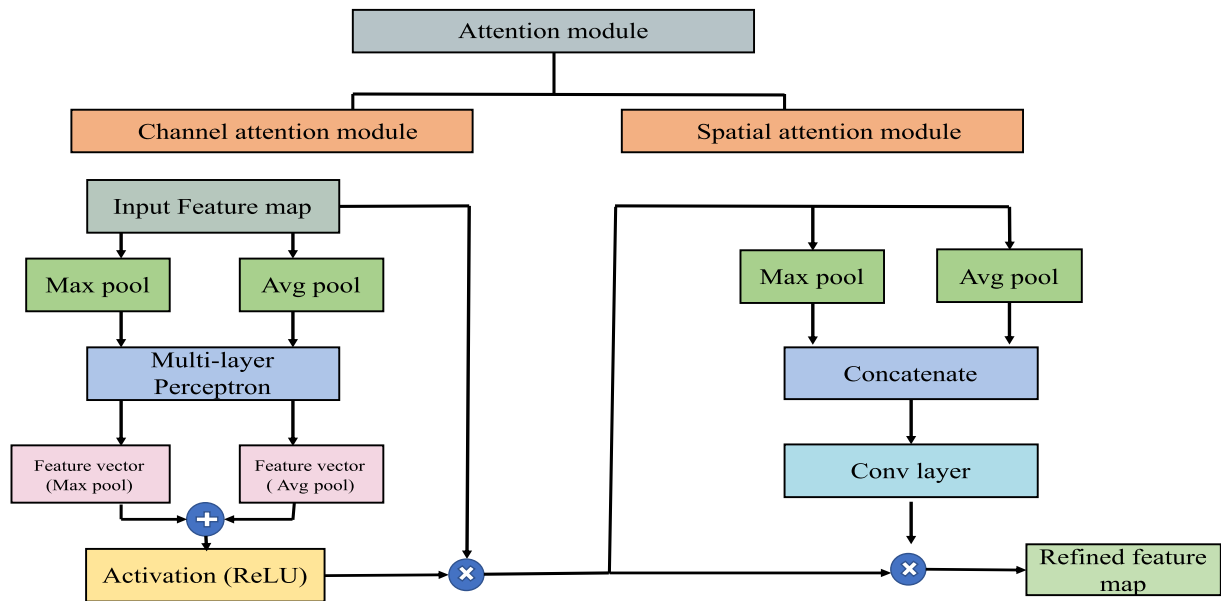


Fig. 2. Architecture of attention module.

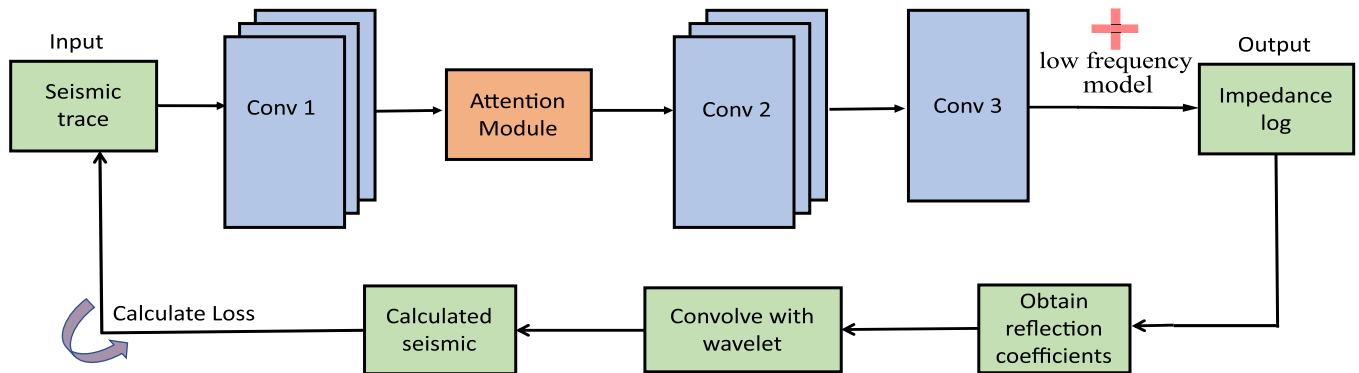


Fig. 3. Network architecture of ADCNN for UL.

parallel arrangement. The output of the complete attention module is given as input to the next layer. The detailed architecture of the attention module is as shown in Fig. 2.

2) *Supervised Learning*: CNNs are widely used in various research fields and achieved good results due to their feature

extraction capability [34]. We used poststack seismic data to estimate AI. The proposed network architecture for supervised case is as shown in Fig. 1. We used three convolution layers (layer1, layer2, and layer3), hence the network was named as deep convolutional neural network. These layers convolve the



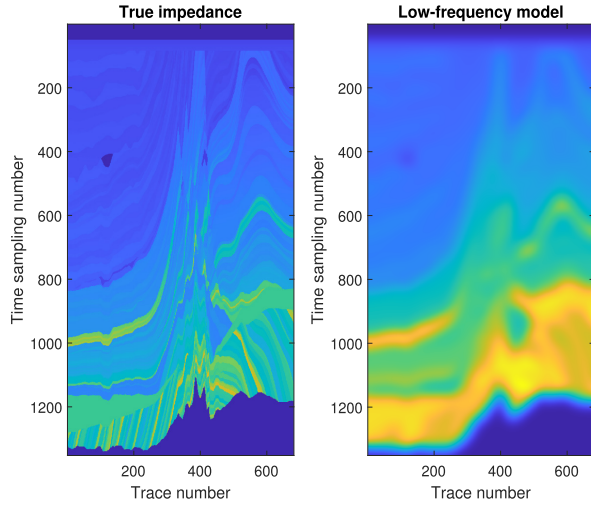


Fig. 4. True impedance and its low-frequency model.

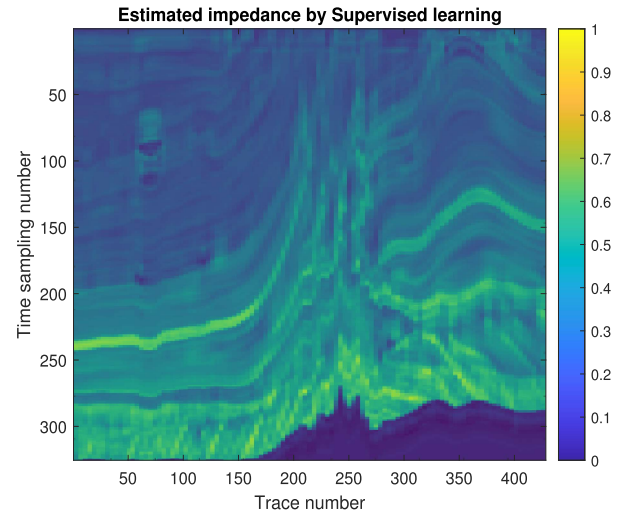


Fig. 6. Estimated impedance by supervised learning.

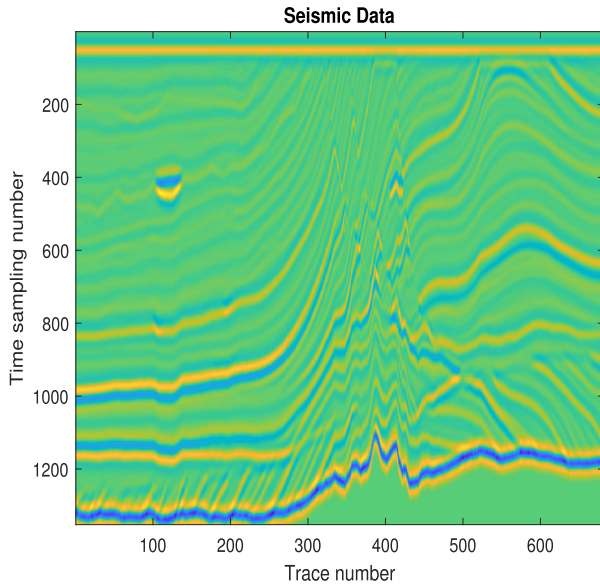


Fig. 5. Seismic data.

input vector with a defined kernel given in layer configuration. The size of layer1 is  $60 \times 1$  in which 60 is the number of output FPs with a stride of 1. The stride is a parameter in network filter that determines the movement of filter. The size of kernel is chosen in accordance with the central frequency of the source wavelet to capture maximum features.

The attention module is placed in a sequential order after layer1 in the network, as shown in Fig. 1. The output of attention module is passed to next convolution layer (layer2). The size of layer2 is  $30 \times 1$  in which 30 is the number of output FPs. Layer3 is the same size as layer1 and layer2 and has one output channel with stride 1. After convolution layer, we add nonlinearity to the network with activation function. Various activation functions, such as tanh, sigmoid, ReLU, ELU, and SeLU, exist in the literature. We used SeLU because it helps the network to converge faster with a good fit compared to existing activation

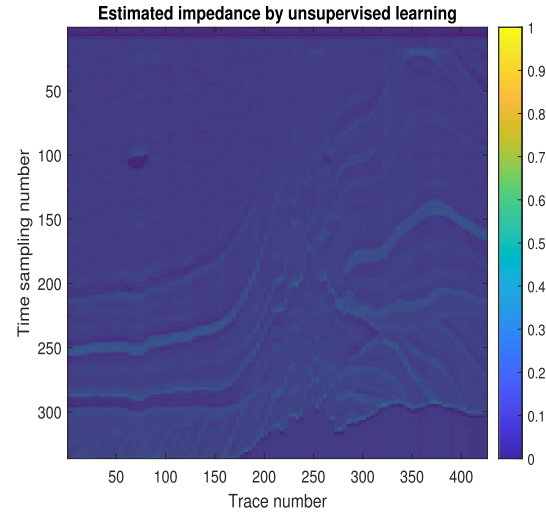


Fig. 7. Estimated impedance by UL.

functions. In addition, it helps to prevent vanishing gradient problem, which usually occurs with sigmoid function. During the training process, the output of the network is compared with the true impedance log and the loss is calculated using mean squared error (MSE) as the cost function

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N \|(z^i - \hat{z}^i)\|_2^2 \quad (7)$$

where  $z^i$  is the true AI data and  $\hat{z}^i$  is the predicted AI data.

3) *Unsupervised Learning*: In case of UL, where the true AI data are not available (such as field data), we use UL approach to estimate AI from input seismic data. The network architecture of the UL approach is shown in Fig. 3. The difference lies in the terms of cost function when compared to supervised learning. Here, we minimize the error between input seismic data and calculated seismic data. The output generated from the network (predicted impedance) is used to generate seismic trace (forward

**Algorithm 1:** An Algorithm for Attention Module.**Require:** input data

- Step 1: Input the feature map. Compute max pooling and avg pooling on spatial dimension to obtain the descriptors  $L_{avg}^c$  and  $L_{max}^c$ .
- Step 2:  $L_{avg}^c$  and  $L_{max}^c$  are given as input to the shared multi layer perceptron to obtain  $A_c(L)$ .
- Step 3: Element wise summation needs to be done as shown in (5). Then Multiply input feature with the obtained  $A_c(L)$  and initialize as an input to spatial attention module.
- Step 4: To obtain spatial attention map, do avg-pooling and max pooling along the channel axis.
- Step 5: Concatenate the feature descriptors  $L_{avg}^s$ ,  $L_{max}^s$  generated and apply convolution operation to obtain spatial attention map  $A_c(L)$ .
- Step 6: multiply input with  $A_s(L)$  to obtain final refined feature map.

modeling) given in (1). The low-frequency model is added to the network output, calculate the reflectivity and convolve with source wavelet to obtain calculated seismic trace ( $s_{cal}$ )

$$r = \frac{AI[i+1] - AI[i]}{AI[i+1] + AI[i]} \quad (8)$$

$$s_{cal} = r * w. \quad (9)$$

The calculated seismic trace is compared with input seismic trace (S) to minimize the loss using the MSE as a cost function

$$MSE = \frac{1}{N} \sum_{i=1}^N ||(s^i - s_{cal}^i)||_2^2 \quad (10)$$

where  $s^i$  is the input seismic trace and  $s_{cal}^i$  is the calculated seismic trace. The optimum weights and biases are obtained by minimizing the cost function mentioned in (7) and (8) using backpropagation algorithms. Various optimization algorithms, such as stochastic gradient descent, adaptive gradient algorithm, root-mean-square propagation, adaptive moment estimation (ADAM) [35], have been studied in literature. In our work, ADAM is used as an optimization algorithm for back propagation. Let  $\theta = \{W^k, b^k\}$ , the Adam optimizer update equation for  $\theta_t$  is given by

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{k}(t)} + \epsilon} \hat{l}(t) \quad (11)$$

where  $\hat{k}(t)$  and  $\hat{l}(t)$  are first and second moments evaluated from  $k_t/1 - \beta_2$  and  $l_t/1 - \beta_1$  after bias corrections. The terms of exponentially moving averages ( $l_t$  and  $k_t$ ) are obtained by using the formula  $l_t = \beta_1 l_{t-1} + (1 - \beta_1) g_t$  and  $k_t = \beta_2 k_{t-1} + (1 - \beta_2) g_t^2$ , respectively. The exponential decay rates are  $\beta_1$  and  $\beta_2$  for the first and second moments with values 0.9 and 0.999, respectively [19]. Here,  $g_t$  is the gradient calculated with regard to time and learning rate ( $\eta$ ) is chosen as 0.001.

**Algorithm 2:** An Algorithm for Seismic Impedance Inversion Using ACNN by Supervised and Unsupervised Learning.**Require:** seismic data(S), True AI ( $z^i$ ), Initialized weights and biases(W, b), number of epochs ( $N_{epochs}$ ), batch size ( $N_{batch}$ )**Supervised learning**

- Step 1: Initialize the parameters such as W, b, batch size in the network. Randomly sample the data for training.
- Step 2: **for**  $N_{epochs}$  steps **do**
- Step 3: Input the seismic data  $S$  to the network in Fig. 1 and predict the AI ( $\hat{z}^i$ )
- Step 4: update the weights(W) and biases (b) using (7)
- Step 5: **end for**

**Unsupervised learning**

- Step 1: Initialize the parameters such as W, b, batch size in the network. Randomly sample the data for training.
- Step 2: **for**  $N_{epochs}$  steps **do**
- Step 3: Input the seismic data  $S$  to the network in Fig. 3 and predict the AI ( $\hat{z}^i$ )
- Step 4: Calculate reflection coefficients from pre-dicted AI and convolve with wavelet to obtain seismic trace(calculated\_seismic)
- Step 5: update the weights and biases by minimizing the cost function in (10) using ADAM.
- Step 6: **end for**
- Output:** Optimized parameters.

## III. NUMERICAL RESULTS

The results of seismic impedance inversion are demonstrated in this section, and the proposed method is compared with the existing methods. The results are validated on Marmousi 2 model, which is briefly explained in the following. The efficiency of the proposed method is analyzed and compared with the state-of-the-art existing method [25]. To measure the accuracy of the proposed method, MSE, Pearson's correlation coefficient (PCC), and coefficient of determination are computed between the estimated and true impedance traces.

## A. Marmousi 2

The Marmousi 2 dataset is an extension of classical Marmousi model created by Allied Geophysical Laboratories [36]. The classical Marmousi model consists of single reservoir, which was widely used for AVO analysis and to validate the imaging algorithms. The classical Marmousi model was extended to Marmousi 2, it is based on the Northern Quenguela Trough in the Quanza Basin of Angola. The Marmousi 2 model covers upto 3.5 km in depth and 17 km across. The model consists of 199 horizons and in addition water layer was extended to 450 m thus leading to complex stratigraphic details.

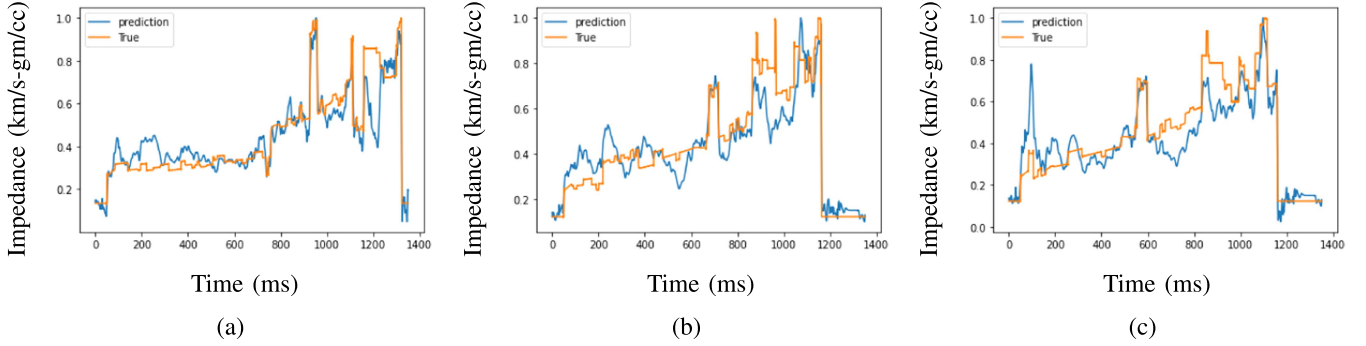


Fig. 8. Comparison of predicted and true impedance for the existed method (supervised): (a) at 200, (b) at 500, (c) at 562 m.

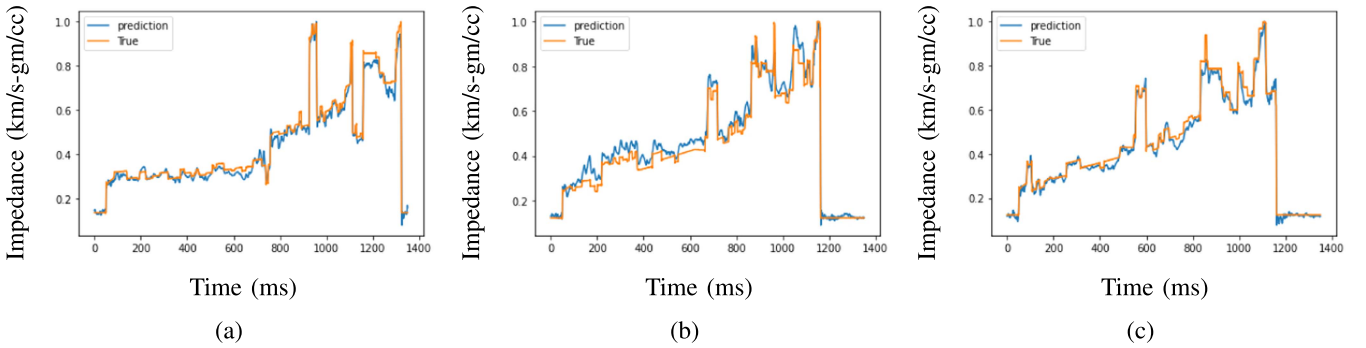


Fig. 9. Comparison of predicted and true impedance for the proposed method (supervised): (a) at 200, (b) at 500, (c) at 562 m.

### B. Training the Network

The acoustic impedance logs for Marmousi 2 are obtained by multiplying their p-velocity and density logs shown in Fig. 4. For each impedance log, we calculate the corresponding seismic trace using (1), as shown in Fig. 5. Both the impedance logs and seismic traces are normalized before training the network. We selected 60% of data for training the network and 40% of data are used for testing. In our work, first we train the network in a supervised way with true AI labels. We trained the network with 2000 epochs and a batch size of 32. An epoch means training the network with the complete training data once. After training (supervised), the network is tested with test data to estimate the acoustic impedance, which is shown in Fig. 6.

We randomly extracted the true and predicted impedance traces and a comparison plot is made, as shown in Figs. 8 and 9, for the existing and proposed method, respectively. For instance, consider the trace at depth 200 m, we can see a good correlation in Fig. 9(a) compared to Fig. 8(a). The importance of attention module is visualized through FPs, which are plotted in Fig. 10. We selected every 8th FP among the configured 60 FPs where Fig. 10(a) shows input FP to the attention module, i.e., channel attention module. Fig. 10(b) and (c) denotes the output of channel attention module and spatial attention module, respectively. From Fig. 10(c), we observe that the prominent features are extracted from the output of attention module. This clearly indicates the prominence of attention module. In

TABLE I  
COMPARISON OF VARIOUS METRICS WITH THE EXISTING CNN METHOD FOR MARMOUSI 2 DATA (SUPERVISED)

Metrics	Proposed method	CNN
MSE	0.00457	0.00651
PCC	0.9832	0.9751
$r^2$	0.977	0.807

particular, we notice that output features from the attention block show the layer boundaries very clearly.

Further, we trained the network in an unsupervised way, i.e., without the need for true AI, as shown in Fig. 3. The hyperparameters are chosen as in the case for supervised learning. The trained network is tested with test data and the estimated AI is as shown in Fig. 7. For comparison, we have taken random impedance traces at various depths which are shown in Figs. 11 and 12. The training loss curve is shown in Fig. 13. From these plots, we can observe the superiority of the proposed method compared to the existing method. When compared to the supervised learning, UL (correlation of 0.9764) has less correlation since in supervised we use true AI labels where as in unsupervised, we do not have an idea of true AI labels.

Table I shows various performance metrics used to evaluate the proposed method when compared to existing methods on Marmousi 2 dataset. The brief description about these metrics



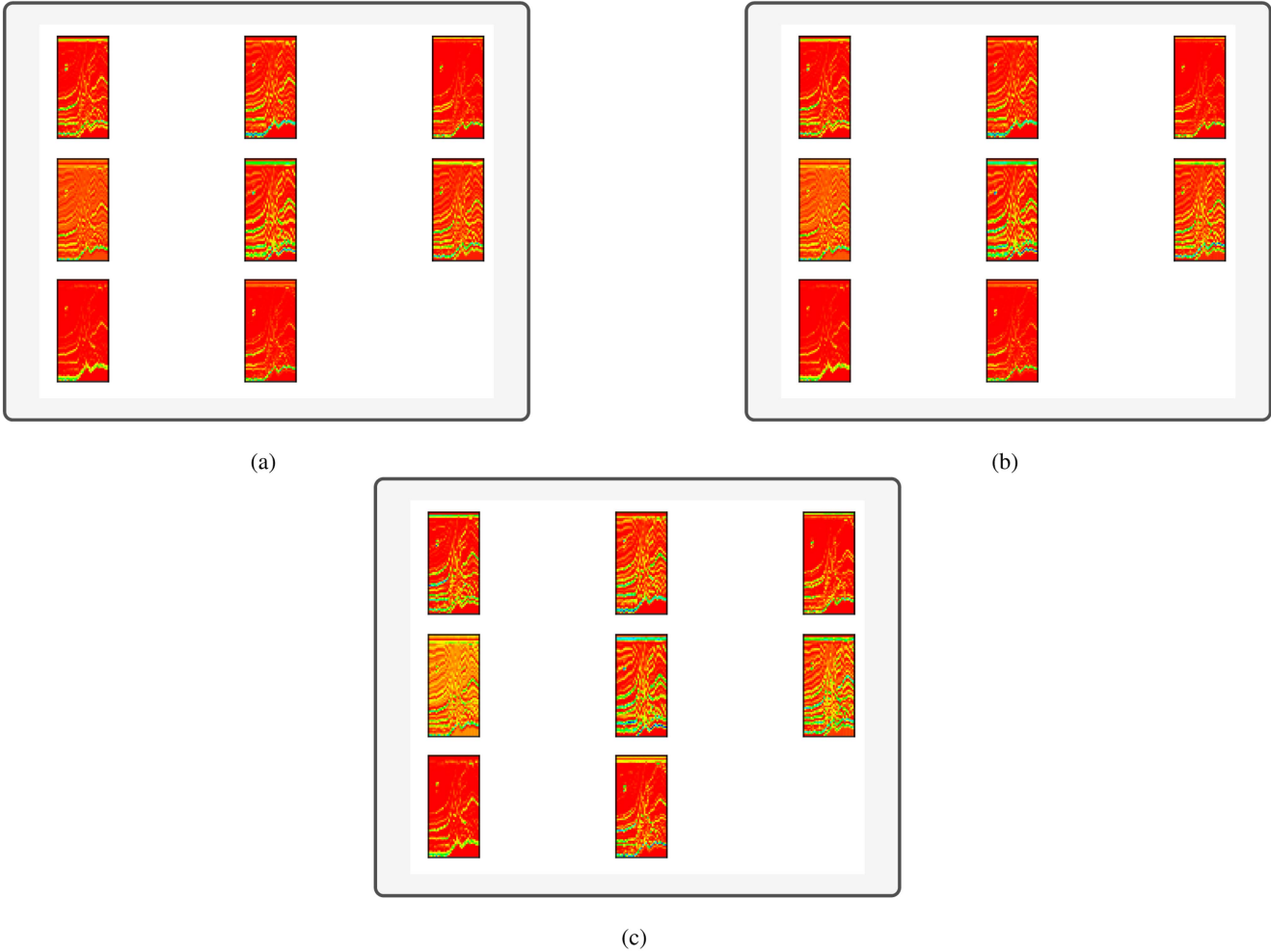


Fig. 10. FPs (randomly taken eight FPs from 60 FPs) for the attention module. (a) Input FPs to the channel attention module. (b) Output of the channel attention module. (c) Output of the spatial attention module.

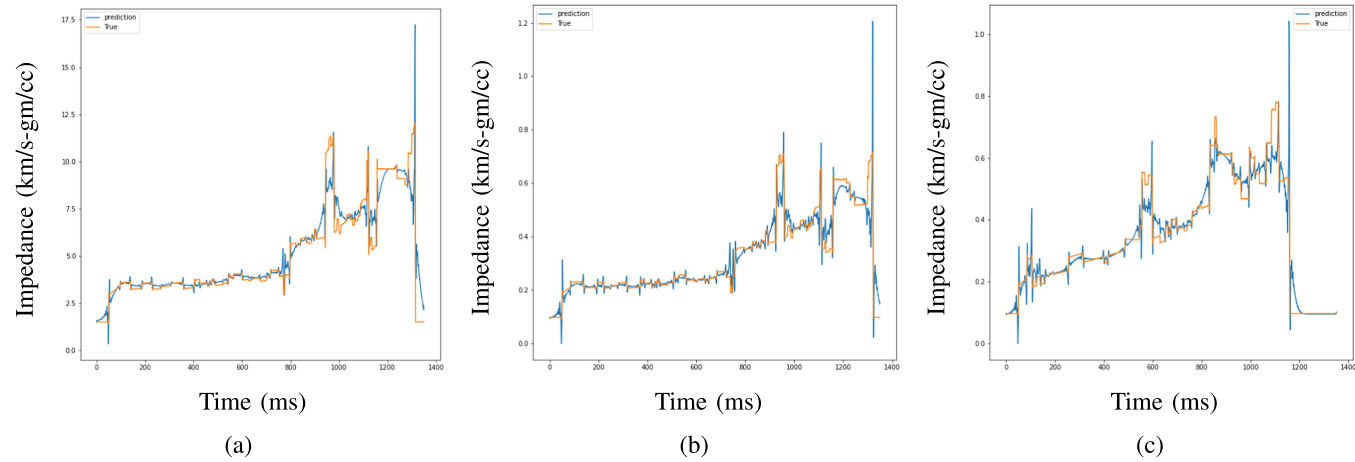


Fig. 11. Comparison of predicted and true impedance for the existed method (unsupervised): (a) at 151, (b) at 200, (c) at 562 m.

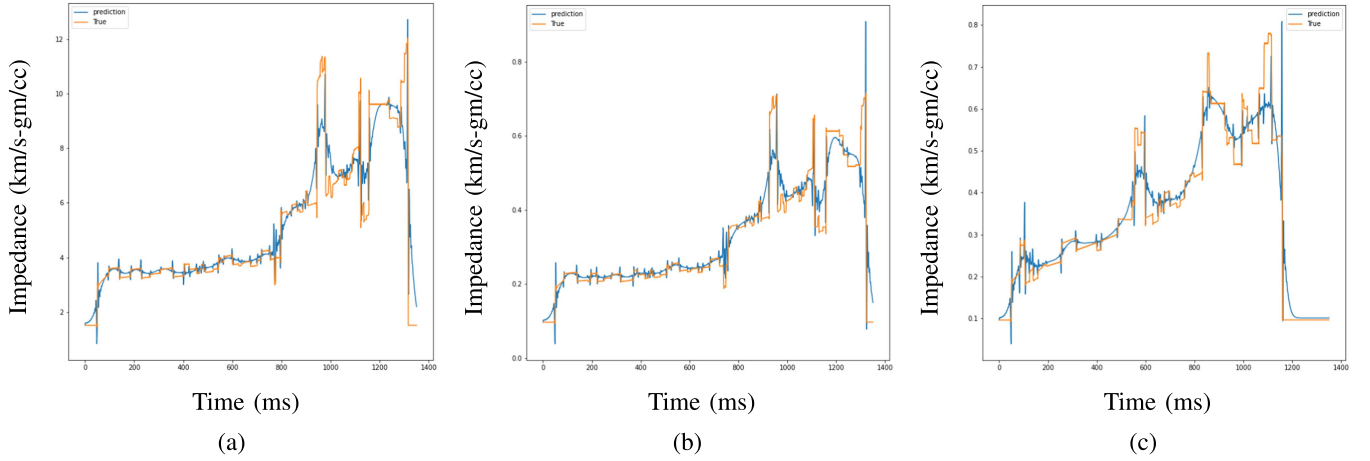


Fig. 12. Comparison of predicted and true impedance for the proposed method (unsupervised): (a) at 151, (b) at 200, (c) at 562 m.

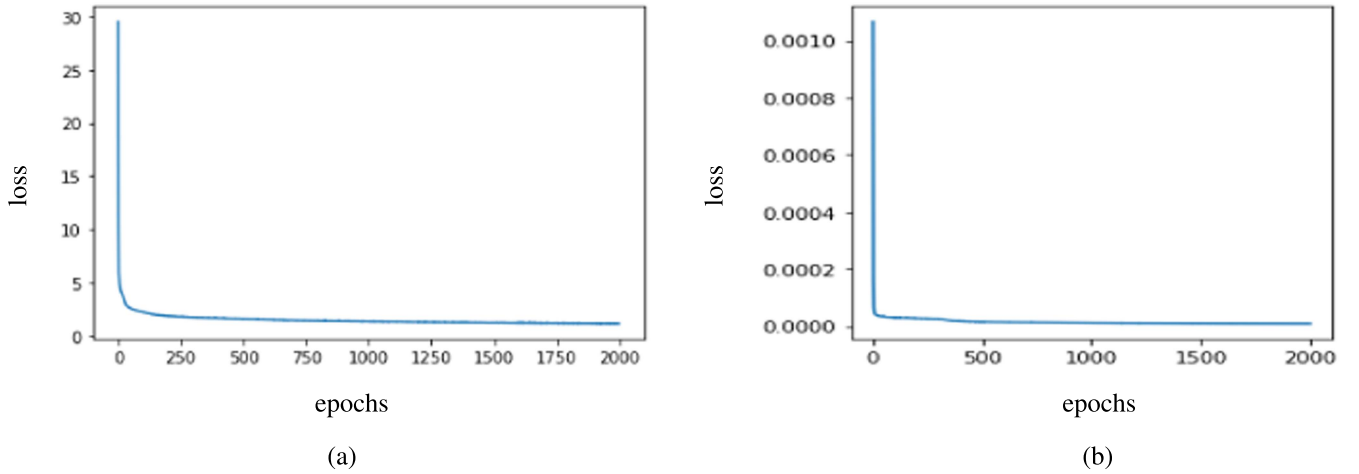


Fig. 13. Training loss. (a) Supervised learning. (b) UL.

are given in the following. Assume that the estimated measurement  $\{\hat{z}^i\}_{i=1}^N$  and its corresponding ground-truth  $\{z^i\}_{i=1}^N$  is available for supervised learning.

**Mean squared error:** The average squared difference between the estimated values and actual values gives the MSE of the data points

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N \|(\hat{z}^i - z^i)\|_2^2. \quad (12)$$

**Coefficient of determination ( $r^2$ ):**  $r^2$  provides measure of how well observed outcomes are obtained from the model based on the proportion of total variation of outcomes given by the model

$$r^2 = 1 - \frac{\sum_{i=1}^N \|(\hat{z}^i - \bar{z})\|_2^2}{\sum_{i=1}^N \|(\bar{z} - z^i)\|_2^2} \quad (13)$$

where  $\bar{z}$  is the average of  $\{z^i\}_{i=1}^N$

**Pearson correlation coefficient:** PCC is a statistic used to measure the correlation between two variables (data). It gives

information about the magnitude and direction of correlation

$$r^2 = \frac{\sum (z^i - \bar{z})(\hat{z}^i - \text{mean}(\hat{z}^i))}{\sqrt{\sum (z^i - \bar{z})^2 \sum (\hat{z}^i - \text{mean}(\hat{z}^i))^2}}. \quad (14)$$

The results are produced by performing simulations on an Intel Xeon Silver 4216 CPU @2.10 GHz (two processors) with 256 GB RAM, 64-bit operating system. The software used is Spyder environment from Anaconda Navigator. The training loss is calculated for both supervised and unsupervised approaches shown in Fig. 13. It took around 2 min. to run the python code and obtain the results for supervised case where as for unsupervised case it took 5 min to get the results. The reason behind this is as UL has to perform forward modeling to generate calculated seismic data. As a result, computation time is increased compared to the supervised case. The hyperparameter tuning is done through BO tuner. The obtained parameters (shown in Table II) are used in the training process where number of layers is chosen as 3 and ADAM as optimizer with a learning rate of 0.001 to minimize the cost function. In addition, we performed noise resistance tests for the proposed method. We added

TABLE II  
OPTIMAL HYPERPARAMETERS

Hyperparameter	Optimized value from BO tuner
Number of layers	3
Filters	60,30,1
Activation function	SeLU
Learning rate	1e-03
Optimizer	Adam
No. of epochs	2000
Batch size	32

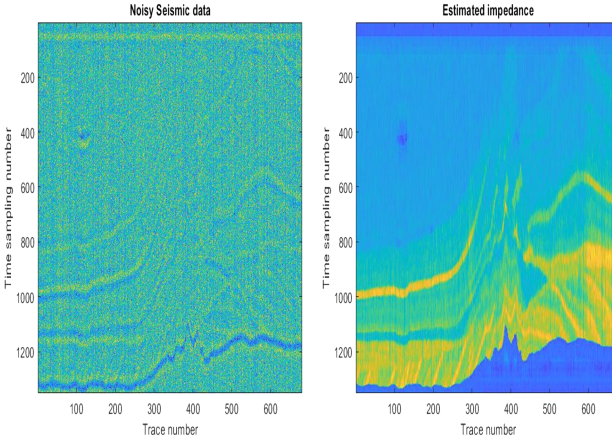


Fig. 14. Noisy seismic data and estimated P-impedance.

gaussian noise to the seismic data and analyzed the accuracy of inversion (supervised learning). The proposed method works efficiently even if the input seismic data are noisy with PCC of 0.963. The noisy seismic data and estimated impedance graphs are shown in Fig. 14.

However, the proposed method has some limitations, which is common with DL methods. The training data are very important in data-driven methods. In general, in these methods, the training and testing data with similar characteristics shows better performance. With the use of different distributed and large amount of data in the training process, we can obtain the generalized model, which works well on any test data but at the cost of computational resources. Hence, in the future we would like to use the concept of federated learning to better optimize the computational resources.

#### IV. CONCLUSION

In this work, we presented a novel approach to address the impedance inversion problem. Our method involves incorporating the attention module CBAM into the neural network architecture, enabling the retrieval of salient features from the input data. We explore two different data analysis approaches: supervised and unsupervised. Supervised learning is utilized when labeled data are available, while UL is employed in the absence of labels where physics of the inverse problem is used. In addition, we leverage the SeLU activation function for its demonstrated stability and efficiency. To automatically optimize the hyperparameters and reduce network training time, we utilize

BO. We used poststack seismic data to demonstrate the results. The results show significant improvements compared to existing methods, as evidenced by metrics, such as MSE, PCC, and coefficient of determination in estimating AI.

#### REFERENCES

- [1] B. H. Russell, *Introduction to Seismic Inversion Methods*. Houston, TX, USA: SEG Books, 1988.
- [2] J. Pendrel, "Seismic inversion—the best tool for reservoir characterization," *CSEG Recorder*, vol. 26, no. 1, pp. 18–24, 2001.
- [3] M. K. Sen, *Seismic Inversion*. Houston, TX, USA: Society of Petroleum Engineers, 2006.
- [4] R. B. Latimer, R. Davidson, and P. Van Riel, "An interpreter's guide to understanding and working with seismic-derived acoustic impedance data," *Leading Edge*, vol. 19, no. 3, pp. 242–256, 2000.
- [5] B. Wu, D. Meng, and H. Zhao, "Semi-supervised learning for seismic impedance inversion using generative adversarial networks," *Remote Sens.*, vol. 13, no. 5, 2021, Art. no. 909.
- [6] M. K. Sen and P. L. Stoffa, *Global Optimization Methods in Geophysical Inversion*. Cambridge, U.K.: Cambridge Univ. Press, 2013.
- [7] R. Zhang, M. K. Sen, S. Phan, and S. Srinivasan, "Stochastic and deterministic seismic inversion methods for thin-bed resolution," *J. Geophys. Eng.*, vol. 9, no. 5, pp. 611–618, 2012.
- [8] R. Nunes, L. Azevedo, and A. Soares, "Fast geostatistical seismic inversion coupling machine learning and fourier decomposition," *Comput. Geosciences*, vol. 23, no. 5, pp. 1161–1172, 2019.
- [9] R. J. Ferguson and G. F. Margrave, "A simple algorithm for band-limited impedance inversion," *CREWES Res. Rep.*, vol. 8, no. 21, pp. 1–10, 1996.
- [10] K. Maynard, P. Allo, and P. Houghton, "Coloured seismic inversion, a simple, fast and cost effective way of inverting seismic data: Examples from clastic and carbonate reservoirs, Indonesia," in *Proc. 29th Annu. Conv.*, 2003, vol. 1, pp. 1–13.
- [11] J. Helgesen et al., "Comparison of constrained sparse spike and stochastic inversion for porosity prediction at Kristin field," *Leading Edge*, vol. 19, no. 4, pp. 400–407, 2000.
- [12] M. K. Sen, A. Datta-Gupta, P. Stoffa, L. Lake, and G. Pope, "Stochastic reservoir modeling using simulated annealing and genetic algorithm," *SPE Formation Eval.*, vol. 10, no. 1, pp. 49–56, 1995.
- [13] D. P. Hampson, B. H. Russell, and B. Bankhead, "Simultaneous inversion of pre-stack seismic data," in *Proc. SEG Tech. Prog. Expanded Abstr. Soc. Exploration Geophysicists*, 2005, pp. 1633–1637.
- [14] B. VerWest, R. Masters, and A. Sena, "Elastic impedance inversion," in *Proc. SEG Int. Expo. Annu. Meeting*, 2000.
- [15] A. Buland and H. Omre, "Bayesian linearized AVO inversion," *Geophysics*, vol. 68, no. 1, pp. 185–198, 2003.
- [16] R. Zhang, M. K. Sen, and S. Srinivasan, "A prestack basis pursuit seismic inversion," *Geophysics*, vol. 78, no. 1, pp. R1–R11, 2013.
- [17] G. Huang, X. Chen, J. Li, C. Luo, H. Wang, and Y. Chen, "Pre-stack seismic inversion using a Rytov-WKB approximation," *Geophysical J. Int.*, vol. 227, no. 2, pp. 1246–1267, 2021.
- [18] X. Wu, Y. Shi, S. Fomel, and L. Liang, "Convolutional neural networks for fault interpretation in seismic images," in *Proc. SEG Int. Expo. Annu. Meeting*, 2018.
- [19] O. M. Saad and Y. Chen, "Deep denoising autoencoder for seismic random noise attenuation," *Geophysics*, vol. 85, no. 4, pp. V367–V376, 2020.
- [20] Z. Geng, X. Wu, Y. Shi, and S. Fomel, "Deep learning for relative geologic time and seismic horizons," *Geophysics*, vol. 85, no. 4, pp. WA87–WA100, 2020.
- [21] H. Chen, J. Gao, W. Zhang, and P. Yang, "Seismic acoustic impedance inversion via optimization-inspired semisupervised deep learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5906611.
- [22] V. Das, A. Pollack, U. Wollner, and T. Mukerji, "Convolutional neural network for seismic impedance inversion CNN for seismic impedance inversion," *Geophysics*, vol. 84, no. 6, pp. R869–R880, 2019.
- [23] J. Zhang, J. Li, X. Chen, Y. Li, G. Huang, and Y. Chen, "Robust deep learning seismic inversion with a priori initial model constraint," *Geophysical J. Int.*, vol. 225, no. 3, pp. 2001–2019, 2021.
- [24] Q. Li and Y. Luo, "Using gan priors for ultrahigh resolution seismic inversion," in *Proc. SEG Int. Expo. Annu. Meeting*, 2019.
- [25] R. Biswas, M. K. Sen, V. Das, and T. Mukerji, "Prestack and poststack inversion using a physics-guided convolutional neural network," *Interpretation*, vol. 7, no. 3, pp. SE161–SE174, 2019.

- [26] M. Alfarraj and G. AlRegib, "Semisupervised sequence modeling for elastic impedance inversion," *Interpretation*, vol. 7, no. 3, pp. SE237–SE249, 2019.
- [27] M. Alfarraj and G. AlRegib, "Petrophysical property estimation from seismic data using recurrent neural networks," in *Proc. SEG Tech. Prog. Expanded Abstr. Soc. Exploration Geophysicists*, 2018, pp. 2141–2146.
- [28] R. Biswas, A. Vassiliou, R. Stromberg, and M. K. Sen, "Stacking velocity estimation using recurrent neural network," in *Proc. SEG Int. Expo. Annu. Meeting*, 2018.
- [29] B. Wu, D. Meng, L. Wang, N. Liu, and Y. Wang, "Seismic impedance inversion using fully convolutional residual network and transfer learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 12, pp. 2140–2144, Dec. 2020.
- [30] Y. Wang, Q. Wang, W. Lu, and H. Li, "Physics-constrained seismic impedance inversion based on deep learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 7503305.
- [31] Y.-Q. Wang, Q. Wang, W.-K. Lu, Q. Ge, and X.-F. Yan, "Seismic impedance inversion based on cycle-consistent generative adversarial network," *Petroleum Sci.*, vol. 19, no. 1, pp. 147–161, 2022.
- [32] A. Cai, H. Di, Z. Li, H. Maniar, and A. Abubakar, "Wasserstein cycle-consistent generative adversarial network for improved seismic impedance inversion: Example on 3D SEAM model," in *Proc. SEG Tech. Prog. Expanded Abstr. Soc. Exploration Geophysicists*, 2020, pp. 1274–1278.
- [33] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
- [34] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
- [36] G. S. Martin, S. Larsen, and K. Marfurt, "Marmousi-2: An updated model for the investigation of AVO in structurally complex areas," in *Proc. SEG Int. Expo. Annu. Meeting*, 2002.



**Vineela Chandra Dodda** (Member, IEEE) received the B.Tech. degree in electronics and communication engineering (ECE) from Amritha Vishwa Vidyapeetham, Coimbatore, India, in 2015, and the M.Tech. degree in embedded systems from KL University, Guntur, India, in 2019. She is currently working toward the Ph.D. degree in seismic data processing, inversion and interpretation based on deep learning methods with the Department of ECE, SRM University, Amaravathi, India, under the guidance of Dr. Karthikeyan Elumalai.

From 2015 to 2017, she was an Associate Software Engineer with Robert Bosch Engineering and Business Solutions Ltd., Coimbatore, India. Her research interests include neural networks, machine learning, and seismic data processing.



**Lakshmi Kuruguntla** (Member, IEEE) received the B.Tech. degree in electronics and communication engineering (ECE) from Jawaharlal Nehru Technological University (JNTU), Hyderabad, India, in 2008, and the M.Tech. degree in digital electronics and communication systems (DECS) from JNTU, Kakinada, India, in 2012.

She is currently an Assistant Professor with Koneru Lakshmaiah Education foundation (KL University), Guntur, India. Her research interests include seismic signal processing, compressive sensing, and dictionary learning.



**Anup Kumar Mandpura** (Member IEEE) received the B.Tech. degree in electronics engineering from Banaras Hindu University, Varanasi, India, in 2006, the M.Tech. degree in digital signal processing from the Indian Institute of Technology Guwahati, Guwahati, India, in 2009, and the Ph.D. degree in wireless communication from the Indian Institute of Technology Delhi, New Delhi, India, in 2018.

He is currently an Assistant Professor with the Department of Electrical Engineering, Delhi Technological University, New Delhi, India. His research interests include the performance analysis of wireless communication systems, energy harvesting communication systems, microwave/millimeter-wave systems, and signal processing.



**Karthikeyan Elumalai** (Member, IEEE) received the B.E. degree in electronics and communication engineering and M.E. degree in communication systems from Anna University, Chennai, India, in 2007 and 2009, respectively, and the Ph.D. degree from Indian Institute of Technology Delhi, New Delhi, India, in 2018.

He is currently an Assistant Professor with the Department of ECE, SRM University, Amaravathi, India. Also, he received a research grant from the Department of Science and Technology, India. His research interests include seismic signal processing, machine learning, and seismic modeling and inversion.



**Mrinal K. Sen** (Member, IEEE) received the M.Sc. degree in applied geophysics from the Indian Institute of Technology (Indian School of Mines), Dhanbad, Dhanbad, India, in 1979, and the Ph.D. degree in theoretical seismology from the Hawaii Institute of Geophysics, University of Hawai'i at Mānoa, Honolulu, HI, USA, in 1987.

He is currently a Professor of geophysics and the holder of the Jackson Chair in Applied Seismology, Department of Geological Sciences, and the Institute for Geophysics, The University of Texas at Austin, Austin, TX, USA. From 2013 to 2014, he was the Director of the National Geophysical Research Institute, Hyderabad, India.

Dr. Sen was a recipient of the 2018 Virgil Kauffman Gold Medal of the Society of Exploration Geophysicists for making significant advancements in the sciences of exploration geophysics in the last five years.



# DeepFake Detection using Transfer Learning

Rahul Thakur<sup>1</sup>, Amit kumar samanta<sup>2</sup>, Amrit<sup>3</sup>, Daksh Garg<sup>4</sup>

<sup>1,2,3,4</sup>Department of Electronics and Communications Engineering, Delhi Technological University, Delhi, India

<sup>1</sup>rahulthakur@dtu.ac.in, <sup>2</sup>amitkumarsamanta\_2k19ec013@dtu.ac.in, <sup>3</sup>amrit\_2k19ec014@dtu.ac.in, <sup>4</sup>dakshgarg\_2k19ec050@dtu.ac.in

**Abstract**—A deepfake is a computer-generated fake image or video that combines images to create a new image or video that depicts an event, comment, or activity that did not actually occur. This has become a real problem nowadays to decide the originality of a video. For the same reason we are trying to create a machine learning model using transfer learning which will help us to distinguish between different videos to decide which video is real and which one is fake. For that we are using four different models and comparing their results. The overall best model is InceptionResNetV2 considering its training time and accuracy. In our results the InceptionResNetV2 performs best and gives an accuracy of 97.1 percent.

**Keywords**—ResNet50V2, InceptionResNetV2, NASNet, Deepfake, InceptionV3, Transfer Learning

## I. INTRODUCTION

In actuality, 14968 phoney videos were published online or on social media, and nearly 96% of them featured celebrities doing obscene actions [3]. The production of fraudulent videos is also rising quickly. Not only images but even video recordings are corrupted, demanded, and/or published publicly.

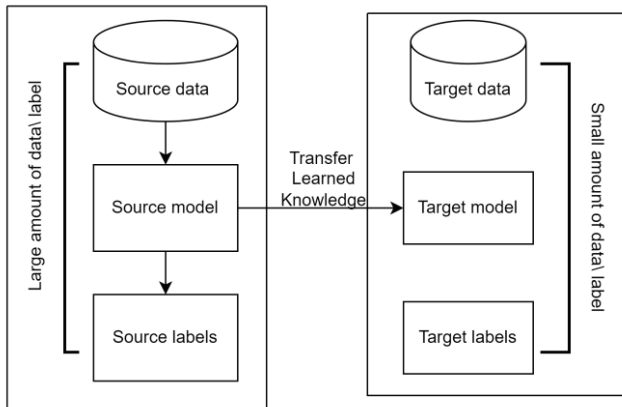


Fig. 1. The concept of transfer learning

DeepFakes are so convincingly fake that it is impossible to spot them with the unaided eye, and an average, uninformed individual would think they were real. The dataset utilised in this study, which consisted of fake photos and videos was taken from Kaggle [7]. View illustrations of genuine and to get the idea of deepfakes. In the dataset there are fictitious photos and videos [7]. The collection includes both false and non-fake photos and videos both of which can be recognised as such by the naked eye. Such false and actual picture identification can be aided by a properly trained deep learning model.

From fig.1 we can say reusing a model that has already been trained on a different issue is known as transfer learning. With transfer learning, a machine may use its understanding of one activity to better generalise about another [1]. In the context of image classification, the usage

of pretrained models is a manifestation of transfer learning. A pretrained model is one that has already been trained on a sizable benchmark dataset and is able to solve issues that are comparable to those that have just been discovered [2]. It is customary to employ models that have been tested and are widely accessible (e.g. InceptionResnetV2, Resnet50V2, Nasnet).

## II. LITERATURE REVIEW

The term "Deepfake" refers to modified photos or any other digital delegations that have created an unreal portion of it. It is a mix of the terms "deep learning" and "fake". AI is used in Deepfake [3]. DeepFakes may be created by anybody with access to computers. A DeepFake is a forgery made by carefully examining the target person's photos or videos and then replicating that person's actions by changing some of them or all of them [4].

TABLE I. THE COMPARATIVE STUDY AND LITERATURE SUMMARY OF DEEPFAKE DETECTION.

Author	Approach	Dataset	Accuracy (%)
Sinnott, R.O et al. , [21]	Mobile net and Xception	FaceForencics++	91
Jung, T et al. , [22]	DeepVision	Static deepfakes eye blinking images dataset	87
Lewis, J.K et al. , [23]	Multimodal network	Facebook deepfake challenge dataset	61
Zhuang, Y.X et al. , [24]	Dense Net and fake feature network	CelebA	90
Wen, Y et al. , [25]	Dense Net	annotated CT-GAN	80
Ismail, A et al. , [26]	XGBoost	CelebDF and FaceForencics++	90

## III. METHODOLOGY

For the process of data preprocessing we need images but our dataset is having videos so we created images from the frames of the videos. After each second, we select one frame. Now from these frames only the facial part for which we used the dlib library which returns the coordinates of the facial part recognized from the image. Upon getting coordinates we used OpenCV to crop out the image in size of 128x128.

Now we will be working on these images. After getting images we used pretrained models like InceptionResNet, NASNetMobile, InceptionV3 and ResNet50V2 to create our desired models for the image classification. After complete

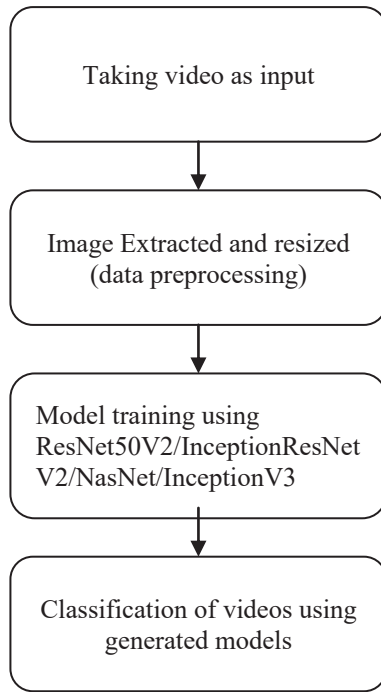


Fig. 2. Block diagram of Methodology

training of our model we tested these models using testing videos. To test the model first we extract the images of the facial part using dlib and opencv and then we use our model on those images. After getting the result of all images we can predict whether the video was fake or real.

#### A. ResNet50V2

One MaxPool layer, one Average Pool layer, and 48 Convolution layers make up the ResNet50 model. We have carefully investigated the ResNet50 design, which is a widely used ResNet model. A minor adjustment was made for ResNet50 and upper; the shortcut connections now skip three layers instead of only two [11].

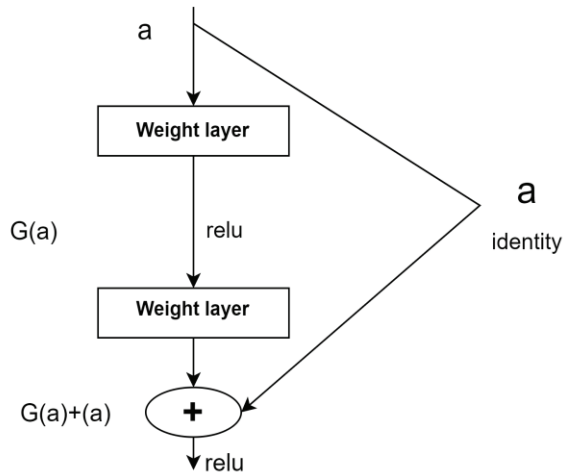


Fig. 3. Skip connection

In fig. 3 These layers are specifically allowed to approximate a residual function  $G(a) = H(a) - a$ . The original function thus becomes  $G(a) + a$ .

To a few hoard layers, we apply residual learning. In this paper, a construction block is thought of as [11]:

$$y = G(a, \{W_i\}) + a. \quad (1)$$

The function  $G(a, \{W_i\})$  exemplifies the residual mapping that has to be learnt. For the illustration in Fig. with two layers,  $G = W_2 \sigma(W_1 a)$  where  $\sigma$  stands for ReLU and the biases are left out to make the notation simpler.

The dimensions of  $a$  and  $G$  in Eqn must be equal to eqn.(1). If this is not the case (for example, when altering the input/output channels), the dimensions can be matched by executing a linear projection  $W_s$  by the skip connections:

$$y = G(a, \{W_i\}) + W_s a \quad (2)$$

In equation (2) a square matrix  $W_s$  is an additional option [11]. However, we shall demonstrate through experiments that the identity mapping is affordable and sufficient for dealing with the degradation problem, therefore  $W_s$  is only utilised when matching dimensions.

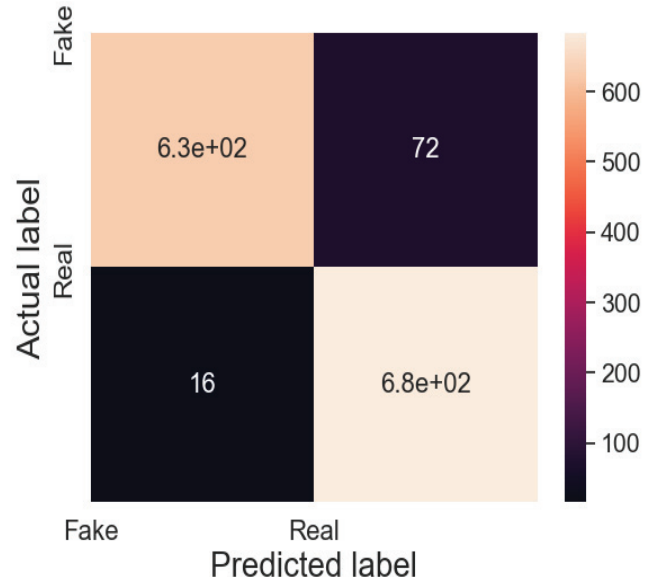


Fig. 4. Confusion matrix of ResNet50V2

After running the Resnet50V2 model, to conclude, we can say from the confusion matrix in fig. 4 that the ratio of correct prediction of real and fake is 680:630 and the ratio of wrong prediction of actual real and fake is 72:16.

#### B. InceptionResNetV2

A convolutional neural network called Inception-ResNet-v2 was used to train almost a million photos from the ImageNet collection [5]. Its 164-layer deep network The network therefore includes suitable feature representations for a diversification of image types. [13].

Its construction makes use of both the Residual link and the Inception formation. Convolutional filters of miscellaneous shapes are merged with residual connections in the Inception-Resnet block.. Utilising residual connections decreases and avoids deterioration problems caused by deep structures[16].

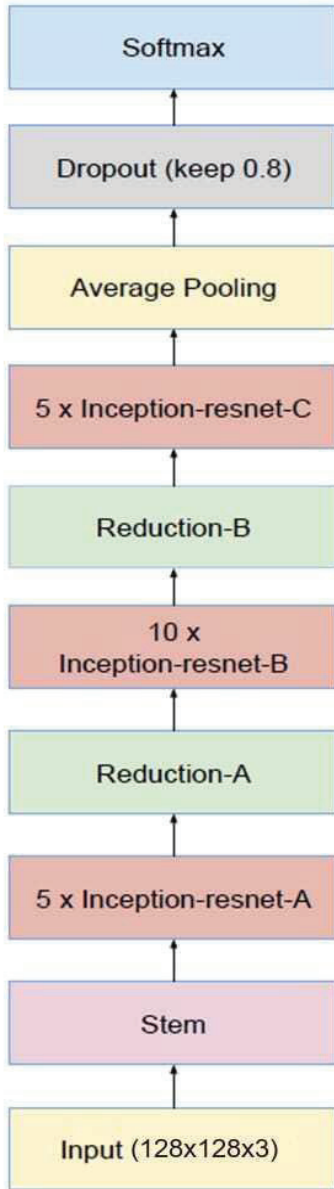


Fig. 5. Schema for InceptionResNetV2

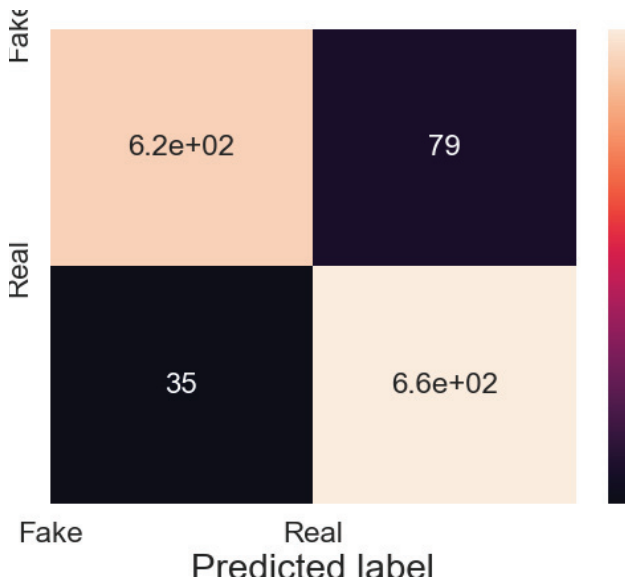


Fig. 6. Confusion matrix of InceptionResNetV2

We are using InceptiveResNetV2 for creating a model structure and added the layers which then finally pointed out to two final results as 0 for fake and 1 for real. On completion we can say from confusion matrix in fig. 6, that the ratio of correct prediction of real and fake is 660:620 and the ratio of wrong prediction of actual real and fake is 79:35.

### C. NASNet

The Google brain team developed the Mobile Neural Architecture Search Network (NASNet), which employs the two primary functions. 1) Normal cells 2) Reduction cells are seen in images 7, 8 [27].

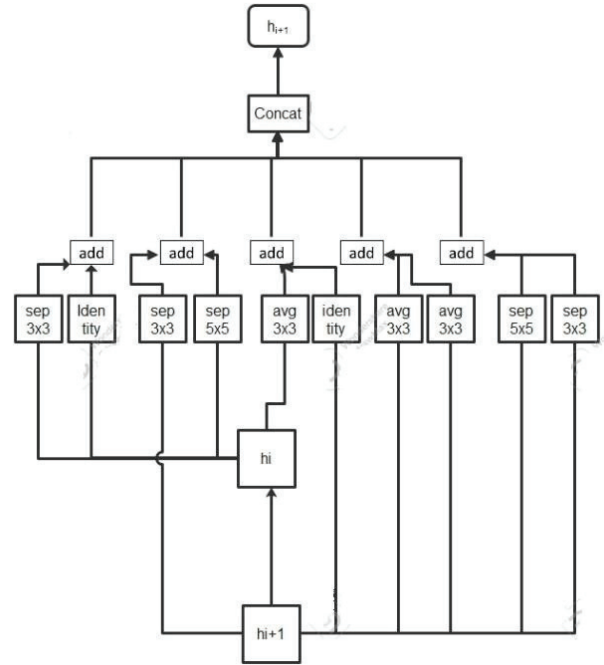


Fig. 7. Nasnet Normal cell

In order to attain a higher mAP, Nasnet first performs its operations on a small dataset before transferring its block to a large dataset. For better Nasnet performance, a customised drop path called Scheduled droppath for efficient regularisation is employed. In the original Nasnet Architecture, which is seen in Figures 7, 8, [27] normal and reduction cells are employed and the number of cells is not predetermined.

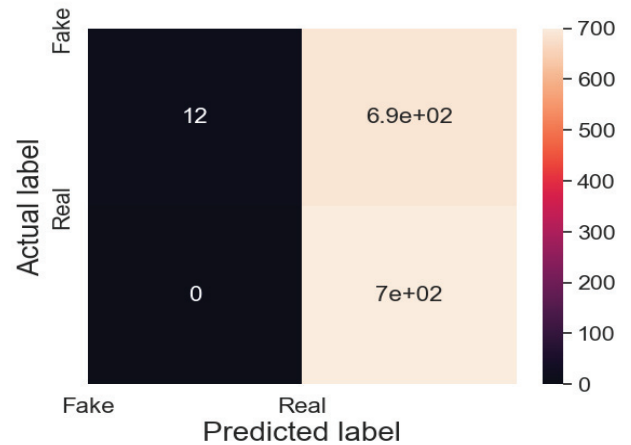


Fig. 8. Confusion matrix of NasNet



While reduction cells provide the feature map that has been reduced by a factor of two in terms of height and breadth, normal cells dictate the size of the feature map [27]. After the NasNet model's execution. Finally, we may state from confusion matrix in fig.8 , that the ratio of correct predictions of real and fake is 700:12 and the ratio of wrong predictions of actual real and fake is 690:0.

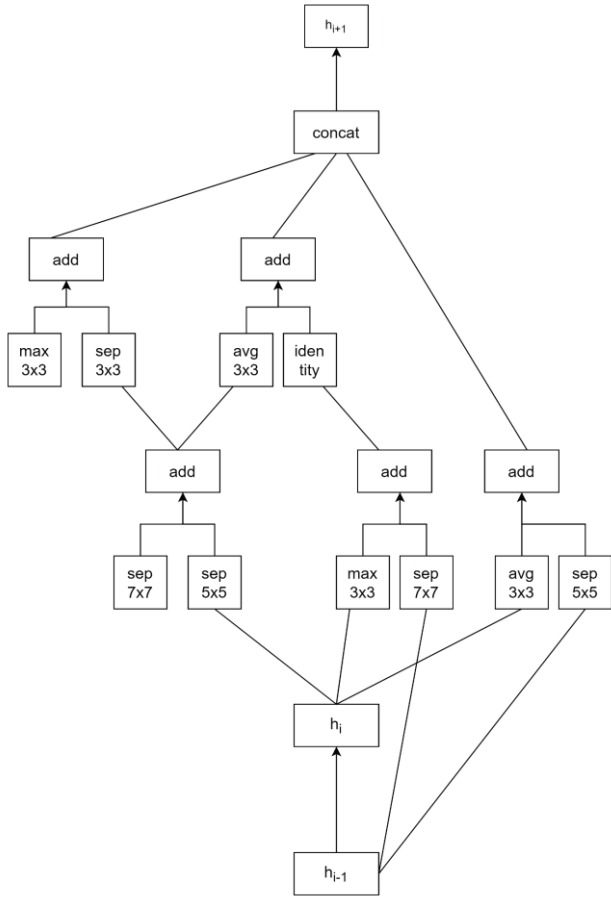


Fig. 9. Nasnet Reduced Cell

#### D. InceptionV3

The inception v3 model in Fig.11, which has 42 layers overall and a reduced error rate than its predecessors, was launched in 2015.[14]. The Inception V1 model has merely been improved and evolved within the Inception V3 model. Several network optimization methods were employed by the Inception V3 model to increase model adaptability. It has a deeper network and is more effective than the Inception V2 and V1 models, nonetheless, its pace is unrelenting.. Computing costs are lower [14]. It employs supporting classifiers as regularizers..It has also learned efficient feature representations over a wide range of images. After completion of running the InceptionV3 model we may state from the confusion matrix in fig. 10 , that the ratio of correct prediction of real and fake is 680:590 and the ratio of wrong prediction of actual real and fake is 110:25.

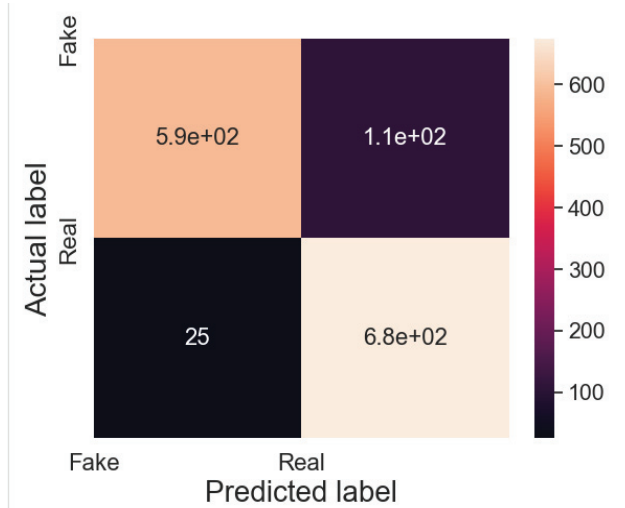


Fig. 10. Confusion matrix of InceptionV3

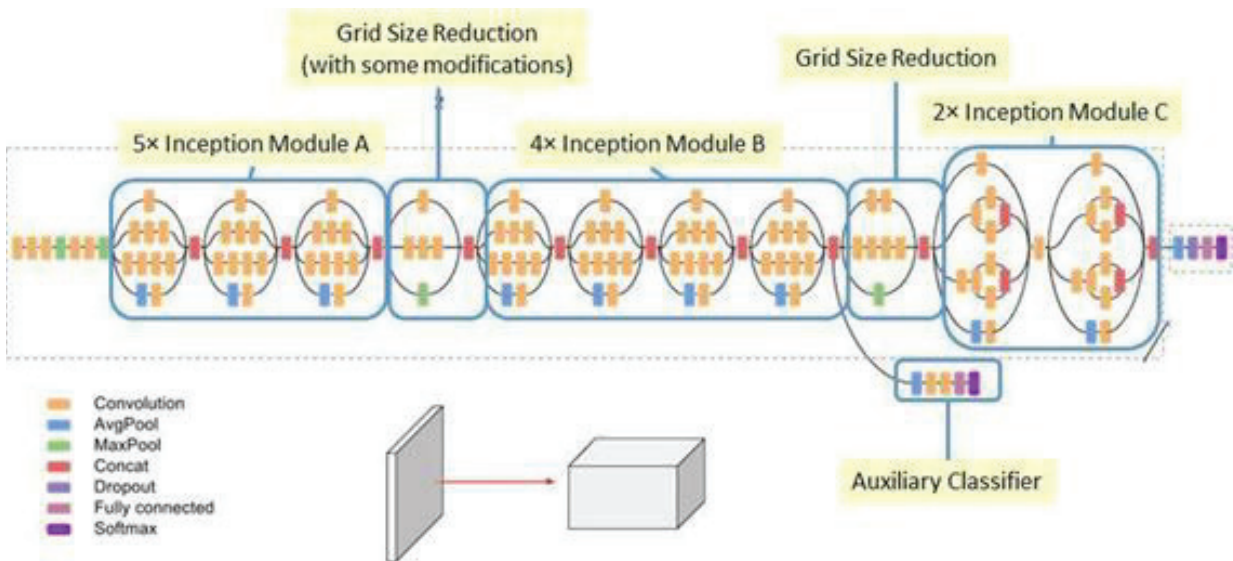


Fig. 11. Inception V3 architecture

#### IV. RESULT AND OBSERVATION

This paper is focused on comparing the different models to get the prediction whether the video is real or not and the time taken per epoch in case of each model.

TABLE II. RESULT COMPARISON

Model name	Time taken per epoch (ms)	Accuracy in model training (percent)
ResNet50V2	160	0.99982
InceptionResNet V2	140	0.97124
NASNet	750	0.9725
InceptionV3	90	0.99982

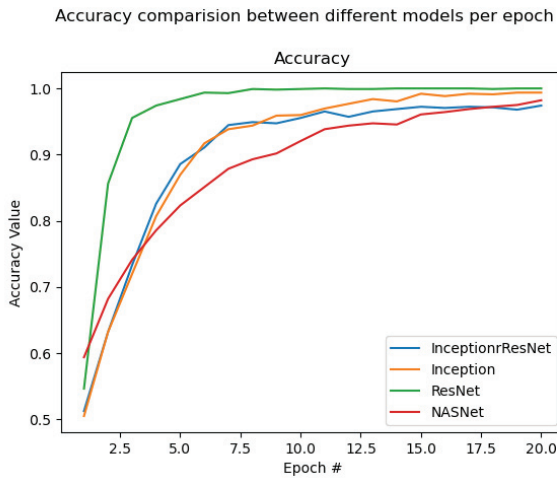


Fig. 12. Accuracy Graph

From table 2 and figure 12, the accuracy reported by InceptionResNetV2, InceptionV3, ResNet50V2, and NASNet is 97.12%, 99.23%, 99.82%, and 97.25%, respectively. According to Table 2, the training of the aforementioned models took 140ms, 90ms, 160ms, and 750ms each epoch, respectively. As a result, InceptionResNetV2 provides the greatest accuracy and requires less training time.

#### V. CONCLUSION

The primary Deepfake Images Detection methods have been reported in this project. In order to develop ever-more-sophisticated DeepFake detection algorithms capable of operating in every situation, we employed four different types of ResNet Model and a number of new datasets. Deep learning-based methods have produced the best outcomes in this field. All of the techniques described in this project may be viewed as starting points from which forensic researchers might begin to develop increasingly reliable and complex answers. As we can see from our observations, InceptionResNetV2 is working in the most effective way than others. It is faster for model training than NASNetMobile, InceptionResNet and it is having accuracy which is not giving the problem of overfitting like that in ResNet50 and InceptionV3.

In the future, research may be done with a larger dataset. Using better system settings with higher resolution (1024 X 1024) content, to build a robust and accurate system that can identify DeepFakes in videos or photographs.

#### ACKNOWLEDGMENT

We would like to express our gratitude to the Electronics and communication Department of Delhi technological University for guiding us throughout making this paper.

#### REFERENCES

- [1] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "video face manipulation detection through ensemble of CNNs: Semantic scholar," *2020 25th International Conference on Pattern Recognition (ICPR)*, 01-Jan-1970.
- [2] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Directory of open access journals," *IEEE Access*, 01-Jan-2022.
- [3] A. Jadhav, A. Patange, H. Patil, J. Patel, M. Mahajan, Deep residual learning for image recognition, *International Journal for Scientific Research and Development* 8 (2020) 1016– 1019.
- [4] T. Zhou, W. Wang, Z. Liang, and J. Shen, "Face forensics in the wild," *arXiv.org*, 30-Mar-2021.
- [5] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-V4, inception-resnet and the impact of residual connections on learning," *arXiv.org*, 23-Aug-2016.
- [6] D. Chaves, E. Fidalgo, E. Alegre, R. Alaiz-Rodríguez, F. Jánhez-Martino, and G. Azzopardi, "Assessment and estimation of Face Detection Performance Based on Deep Learning for Forensic Applications," *MDPI*, 11-Aug-2020.
- [7] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (DFDC) dataset," *arXiv.org*, 28-Oct-2020.
- [8] D. Guera and E. Delp, "[PDF] deepfake video detection using recurrent neural networks: Semantic scholar," *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Jan-2018.
- [9] L. Zhang, T. Qiao, M. Xu, N. Zheng, and S. Xie, "Unsupervised learning-based framework for Deepfake Video Detection: Semantic scholar," *IEEE Transactions on Multimedia*, 13-Jun-2022.
- [10] N. Messina, G. Amato, F. Carrara, F. Falchi, and C. Gennaro, "Testing deep neural networks on the same-different task: Semantic scholar," *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, Jan-2019.
- [11] K. He, X. Zhang, S. Ren, and J. Sun, "[PDF] deep residual learning for image recognition: Semantic scholar," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 10-Dec-2015.
- [12] M. Mahdianpari, B. Salehi, M. Rezaee, F. Mohammadimanesh, and Y. Zhang, "[pdf] very deep convolutional neural networks for complex land cover mapping using multispectral remote sensing imagery: Semantic scholar," *Remote. Sens.*, 14-Jul-2018.
- [13] Y. Bhatia, A. Bajpayee, D. Raghuvanshi, and H. Mittal, "Image captioning using Google's inception-resnet-V2 and recurrent neural network: Semantic scholar," *2019 Twelfth International Conference on Contemporary Computing (IC3)*, 08-Aug-2019.
- [14] C. Lin, L. Li, W. Luo, K. C. P. Wang, and J. Guo, "Transfer learning based traffic sign recognition using inception-V3 model," *Periodica Polytechnica Transportation Engineering*, 2019.
- [15] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for Deepfake Forensics," *CVF Open Access*, Jan-2020.
- [16] M. Koopman, A. M. Rodriguez, and Z. Geradts, "Detection of DeepFake video manipulation," in *The 20th Irish machine vision and image processing conference (IMVIP)*, 2018, pp. 133±136.
- [17] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," *arXiv preprint arXiv:2102.11126*, 2021.
- [18] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The DeepFake detection challenge (dfdc) preview dataset," *arXiv preprint arXiv:1910.08854*, 2019.
- [19] M. Patel, A. Gupta, S. Tanwar, and M. Obaidat, "Trans-Df: A transfer learning-based end-to-end deepfake detector: Semantic scholar," *2020*

- [20] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning Transferable Architectures for Scalable Image Recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 8697–8710, 2018.
- [21] D. Pan, L. Sun, R. Wang, X. Zhang, and R. Sinnott, "Figure 1 from Deepfake detection through deep learning: Semantic scholar," *2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)*, Dec-2020.
- [22] T.-hyun Jung, S. Kim, and K. Kim, "DeepVision: Deepfakes detection using human eye blinking pattern: Semantic scholar," *IEEE Access*, 20-Apr-2020.
- [23] J. Lewis, I. E. Toubal, H. Chen, V. Sandesera, M. Lomnitz, Z. Hampel-Arias, P. Calyam, and K. Palaniappan, "Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning: Semantic scholar," *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 10-May-2021.
- [24] C.-C. Hsu, Y.-X. Zhuang, and C.-Y. Lee, "Deep fake image detection based on pairwise learning," *MDPI*, 03-Jan-2020.
- [25] S. Solaiyappan, Y. Wen, Machine learning based medical image deepfake detection: A comparative study. *Machine Learning with Applications*. 2022 Jun 15;8:100298.
- [26] A. Ismail, M. Elpeltagy, M. S. Zaki, and K. Eldahshan, "A new deep learning-based methodology for video deepfake detection using XGBoost," *MDPI*, 10-Aug-2021.
- [27] SK. Addagarla, GK. Chakravarthi, P. Anitha. Real time multi-scale facial mask detection and classification using deep transfer learning techniques. *International Journal*. 2020 Jul;9(4):4402-8.

# Denoising using a Hybrid Filter Comprised of GGIF, WLS, and 2D Bilateral Filtering

Abhilasha Sharma

Department of Software Engineering  
Delhi Technological University  
New Delhi, India  
abhi16.sharma@gmail.com

Aryaman Dosajh

Department of Software Engineering  
Delhi Technological University  
New Delhi, India  
aryamandosajh\_2k19se018@dtu.ac.in

Moin Ahmad Chalkoo

Department of Software Engineering  
Delhi Technological University  
New Delhi, India  
moinahmadchalkoo\_2k19se075@dtu.ac.in

**Abstract-** Denoising helps improve image quality and recover important information from noisy pictures. This work introduces a novel denoising method. GGIF, WLS, and 2D Bilateral Filtering are used in a hybrid filter. The hybrid filter shown eliminates noise while preserving the picture's characteristics, edges, and textures. Denoising begins with GGIF, which preserves edges and reduces artefacts. Then, WLS reduces amplified noise and smooths the image adaptively according to the local gradient. Last, 2D bilateral filtering reduces noise while keeping structural integrity and edge information. The hybrid filter is tested using simulated and real-world noisy images. The trials reveal that the recommended strategy outperforms current denoising approaches in noise reduction, edge retention, and visual quality. The hybrid filter may solve image processing, computer vision, and remote sensing problems. It uses GGIF, WLS, and 2D Bilateral Filtering to denoise well and improved perceptual fog density.

**Keywords-** *Perceptual fog density, Image processing, MATLAB, Weighted least squares, Guided Image Filtering*

## I. INTRODUCTION

The technique of denoising a picture is a key step in the preprocessing phase of a wide variety of applications, including computer vision, medical imaging, remote sensing, and digital photography [1-3]. Denoising is a method that may be used to eliminate noise from an image while still maintaining significant characteristics of the picture, such as its edges, textures, and minute details [4]. This is the primary objective of the denoising process. The research that has been done on denoising has led to the creation of a very large number of distinct techniques, each of which has both benefits and drawbacks [5]. Recently, there has been a lot of interest in the development of hybrid filters, which combine a number of different approaches into a single process [6]. These filters are designed with the intention of achieving a greater level of performance by making the most of the advantages provided by each distinct strategy [7].

In this context, we present a hybrid filter for denoising that incorporates Globally Guided Image Filtering (GGIF), Weighted Least Squares (WLS), and 2D Bilateral Filtering. GGIF stands for Globally Guided Image Filtering, and WLS and 2D Bilateral Filtering are abbreviations for Weighted Least Squares Globally Guided Image Filtering is abbreviated as

GGIF, while Weighted Least Squares is abbreviated as WLS. Denoising was successfully completed by making use of each of these approaches, each of which has a unique set of benefits and drawbacks. We have high hopes that by integrating these many techniques into a single hybrid filter, we will be able to overcome the constraints that are unique to each method and increase the overall effectiveness of the denoising process.

Denoising an image is complex because it requires reducing noise in a manner that is effective while at the same time preserving the core characteristics that were there in the original picture [8]. This is the most challenging part of denoising an image. These components consist of edging, textures, and minute details in the design [9]. The existing denoising techniques often have difficulty striking the right balance because they either smooth the image too much, which results in the loss of important information, or they fail to reduce noise properly, which results in poor visual quality [10]. These two choices are not the best ones to make. As a consequence of this, there is a need for a technique of noise reduction that is effective at overcoming these limitations and enhancing the image quality by making use of the benefits provided by a variety of techniques of noise reduction [11].

This study's objective is to develop a hybrid filter for denoising that takes use of Globally

Guided Image Filtering (GGIF), Weighted Least Squares (WLS), and 2D Bilateral Filtering in equal measure. The need of creating a hybrid filter for denoising data was the impetus for this body of work. The purpose of the proposed hybrid filter is to achieve higher performance in terms of noise reduction, edge preservation, and overall visual quality by overcoming the limitations of the separate approaches by combining them into a single method. This is done in order to overcome the constraints of the separate approaches [12-14]. The development of a filter of this sort may be of significant value to a range of applications, including computer vision, medical imaging, remote sensing, and digital photography [15-17]. These applications all need high-quality images for accurate processing and interpretation of the data they collect.

This cutting-edge hybrid filter combines the GGIF, WLS, and 2D Bilateral Filtering techniques in an attempt to provide a denoising solution that is comprehensive in nature. This is where the hybrid filter gets its unique characteristics from.

Since it utilises all of these different strategies at the same time, the hybrid filter is able to make the most of the positive aspects of each one while mitigating the negative aspects of the others, which eventually results in improved noise reduction capabilities. The implementation of the hybrid filter and its evaluation on synthetic and real-world noisy images demonstrate its effectiveness in noise reduction and edge preservation, outperforming state-of-the-art denoising techniques [18]. This is demonstrated by the fact that the hybrid filter outperforms traditional denoising methods. This is shown by the fact that the hybrid filter works better than the conventional techniques of noise reduction. This innovative approach to denoising offers a different perspective on image processing and has the potential to enhance a broad variety of applications that rely on high-quality photographs [19-20].

The GGIF filtering technique is one that keeps the edges of a picture while decreasing noise by taking into consideration the local structure of the guiding image. This approach was developed by Google. It is well renowned for the adaptive smoothing capabilities that the WLS approach has. This is due to the fact that the approach adjusts the level of smoothing based on the local gradient in the image. This leads to a decrease in noise without significantly altering the appearance of major components of the picture. On the other hand, the 2D Bilateral Filtering method is a well-established strategy that displays outstanding performance in the preservation of edge information while effectively suppressing noise. This is accomplished by the use of two filters that operate in opposite directions.

This paper presents the design, implementation, and evaluation of the proposed hybrid filter, which combines the strengths of GGIF, WLS, and 2D Bilateral Filtering to achieve superior denoising performance. GGIF stands for Globally Guided Image Filtering, and WLS and 2D Bilateral Filtering stand for Two-dimensional Bilateral Filtering. In the context of two-dimensional bilateral filtering, the abbreviations GGIF and WLS stand for globally guided image filtering and 2D Bilateral Filtering, respectively. In order to illustrate the filter's effectiveness in terms of noise reduction, edge retention, and overall increase in visual quality, it is applied to a wide range of noisy photographs, some of which were created digitally while others were shot in the real world. The results of the experiments are compared to the most cutting-edge methods that are presently being used for the process of denoising. This comparison serves to emphasize the benefits of the proposed hybrid filter as well as its potential applications in a wide range of industries.

## II. LITERATURE REVIEW

In recent years, a number of different denoising approaches have been developed, all with the goal of enhancing noise reduction, edge preservation, and overall visual quality. This section provides a high-level summary of a selection of pertinent research that were published after 2019 and that have led to the development of various noise-reduction techniques, such as hybrid filters.

Chen, Y. et al. [1] proposed a work to provide a new adaptive guided image filtering-based denoising approach that enhances both edge retention and noise reduction performance. The advantages of guided image filtering and approaches for bilateral filtering have been combined in the method that has been suggested.

Li et al. [2] suggested that for the purpose of picture denoising a hybrid filter that combines directed filtering with non-local means filtering. The approach achieves higher performance in both noise reduction and edge preservation as a result of its efficient use of the strengths that are inherent in both methods. Luo, X. et al. [3] suggested that the purpose of this work is to offer a deep learning-based denoising approach that makes use of hybrid filters. These filters combine classic filtering methods with convolutional neural networks. In comparison to the approaches that are already in use, the method that was developed displays superior performance in terms of reducing noise and maintaining picture attributes.

Wang et al. [4] suggested a hybrid filtering approach for the purpose of picture denoising. This method combines guided filtering with block-matching and 3D filtering, and it is referred to as BM3D. The technique is very successful in removing noise while maintaining the image's borders and small features in their original state.

Zhang et al. [5] provided a deep plug-and-play super-resolution approach that makes use of hybrid filters for the purpose of denoising the image. Deep learning and conventional filtering approaches are brought together in this method, which results in a versatile framework that may be used for a variety of different denoising applications.

These experiments demonstrate the significance of integrating a number of different denoising approaches in order to get higher performance in terms of noise reduction, edge preservation, and visual quality. The goal of the hybrid filter that combines GGIF, WLS, and 2D Bilateral Filtering that has been presented is to build upon these recent achievements and further contribute to the development of successful denoising techniques.

TABLE I. COMPARISON ANALYSIS

Ref,year	Technique Used	Advantages	Disadvantages
[6],2018	G-GIF	Produce sharper and well-preserved images	Visual inconsistencies in inpainted regions
[8],2021	Laplacian and Gaussian Pyramids	The object search is faster using a coarse-to-fine strategy	Image resolution is reduced
[13],2021	Adaptive Airlight Refinement and Non-Linear Color Balancing	It produces visually pleasing images without halo artifacts maintaining the naturalness of the image	Loss of original image fidelity and overcorrection
[14],2022	Non Linear Transformation	Proposed method transforms the minimum filtering	Risk of overfitting in particularly in



		on superpixels of a hazy image into the minimum filtering on superpixels of a haze-free image which prevents over enhancement in the long-range regions	areas with complex structures and high frequency details
[19],2019	Fast Adaptive Bilateral Filtering	It attempts to eliminate the meaningless texture while preserving dominant structure as well as possible	Smoothering blur out intricate textures & introduce halo artifacts around high contrast images

### III. IMPLEMENTATION

The following steps need to be taken in order to accomplish denoising and dehazing utilizing a hybrid filter that is comprised of GGIF, WLS, and 2D Bilateral Filtering using MATLAB software with a changed intensity of GGIF:

- Generate up the picture: With the imread function, bring the picture into MATLAB so that it may be processed. In order to continue processing the picture, use the im2double function to convert the image to values with double precision.
- Remove the haze from the picture by using a dehazing technique like the dark channel previous approach to estimate the transmission map and the amount of ambient light. Make use of these settings to restore the picture without the haze.
- The updated Globally Guided Image Filtering (GGIF) should be applied as follows: Adjust the settings so that the updated GGIF has the appropriate parameters, such as the changed intensity parameter (epsilon) and the window size for the filter. Apply the altered GGIF by making use of a specialized function that, upon receiving the dehazed picture as well as the parameters as input, returns the image that has been filtered.
- Implement the filtering known as Weighted Least Squares (WLS): Adjust the settings for the WLS filtering, including the values for lambda and alpha. WLS filtering may be applied to the image that was produced from the GGIF stage by using the edge preserving filter function of MATLAB in conjunction with the 'wls' parameter.
- Setting the settings for the 2D Bilateral Filtering, including the domain and range standard deviations, is the first step in using this filtering method. Applying 2D Bilateral Filtering on the picture that was produced from the WLS stage requires the use of a custom

MATLAB function or an implementation provided by a third party.

- Postprocessing: A postprocessing phase that adjusts the contrast and brightness of the picture may be used to further improve the quality of the image that is produced. The visual quality of the final denoised and dehazed picture may be improved by the use of techniques such as histogram equalization and adaptive contrast enhancement, amongst others.

We used MATLAB to create a denoising and dehazing hybrid filter by following these steps. This filter will combine the benefits of GGIF, WLS, and 2D Bilateral Filtering, which will result in an improvement in picture quality. The photographs that were used as input are shown in figure 1. The method described here is applicable to any kind of picture; however, for demonstration purposes, we will utilize six photos. The majority of the pictures taken outside exhibit favorable outcomes.

### IV. RESULTS

MATLAB was used as the platform for the development of the denoising and dehazing hybrid filter, which is a combination of the modified Globally Guided Image Filtering (GGIF), Weighted Least Squares (WLS), and 2D Bilateral Filtering. The findings that were obtained using this strategy indicate that there was a substantial improvement in both the picture quality and the perception fog density.

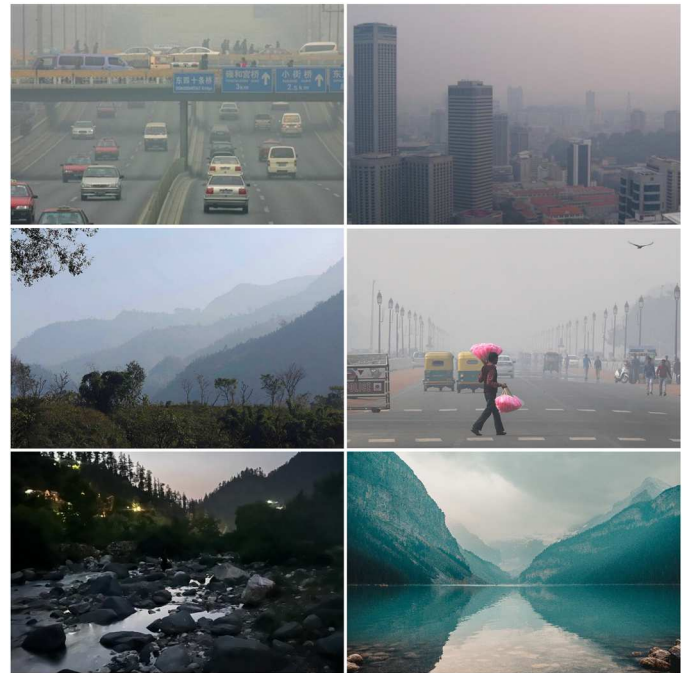


Fig. 1. Input Images Hazy and Noisy



TABLE II: RESULTS COMPARISON

PFD	GGIF	GGIF with WLS & Bilateral Filter
Image 1 PFD	0.800454	0.409022
Image 2 PFD	2.47031	0.844825
Image 3 PFD	1.75596	0.853404
Image 4 PFD	1.8298	0.814872
Image 5 PFD	1.03638	0.555204
Image 6 PFD	1.32298	0.688661

The dark channel prior approach that was used for the purpose of preprocessing was successful in removing the haze that was present in the input picture, which resulted in an image that was both clearer and more aesthetically attractive. This action made a considerable contribution towards the overall improvement of the perceived fog density.

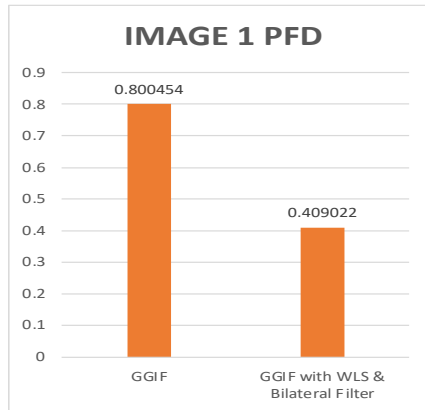


Fig. 2. Image 1 PFD Result

Improved denoising: The updated GGIF, which featured a changed intensity parameter, efficiently reduced noise while maintaining the picture structure. This was accomplished without altering the original GGIF. With manipulation of the filter's intensity parameter, edge information was maintained, and the filter was able to conform to the regional differences present in the picture, resulting in an output of improved quality. Fig. 2 to Fig. 7 shows the graph outputs respectively.

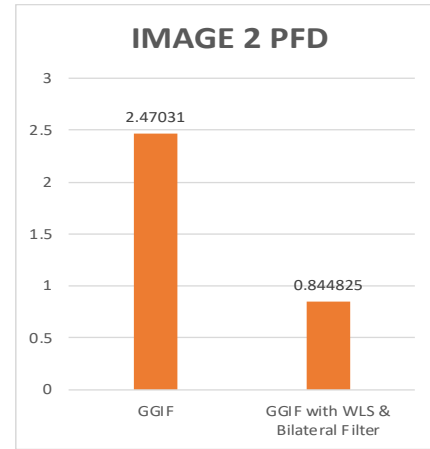


Fig. 3. Image 2 PFD Result

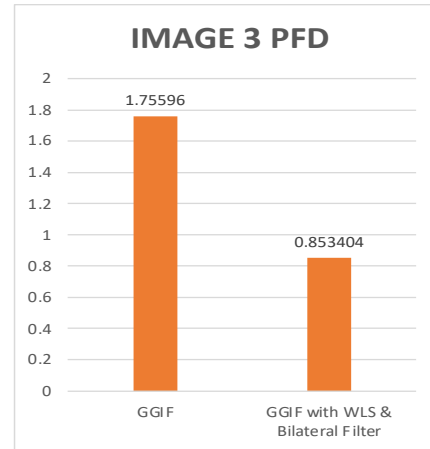


Fig. 4. Image 3 PFD Result

The WLS filtering approach further improved the picture by limiting noise amplification and adaptively smoothing the image depending on the local gradient. This allowed for the retention of the image's edges. This led to an improvement in the preservation of edges and a reduction in noise, both of which contributed to an increase in the apparent density of the fog.

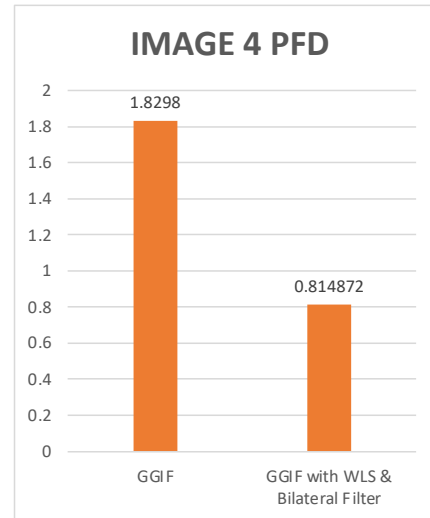


Fig. 5. Image 4 PFD Result

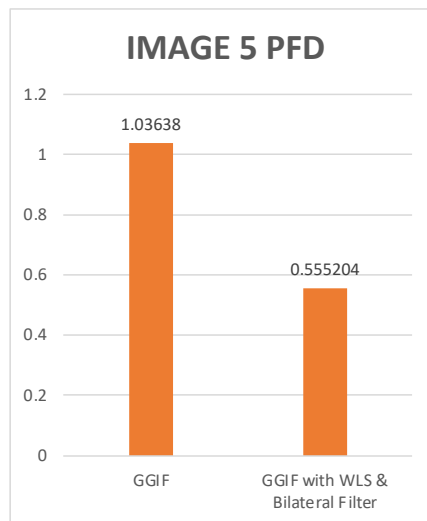


Fig. 6. Image 5 PFD Result



Fig. 7. Image 6 PFD Result

The output results of the images are shown in figures screenshots from fig 8 to fig 13. Noise suppression and structural integrity: The 2D Bilateral Filtering further enhanced the picture by efficiently suppressing noise while maintaining the image's structural integrity and edge information. This was accomplished without sacrificing the image's integrity. This filtering method was crucial in the production of a final output picture that had increased perceptual fog density as well as lower levels of image noise.

Postprocessing: The contrast and brightness of the picture's final output were improved using histogram equalisation or adaptive contrast enhancement algorithms, which resulted in an image that was more aesthetically attractive.

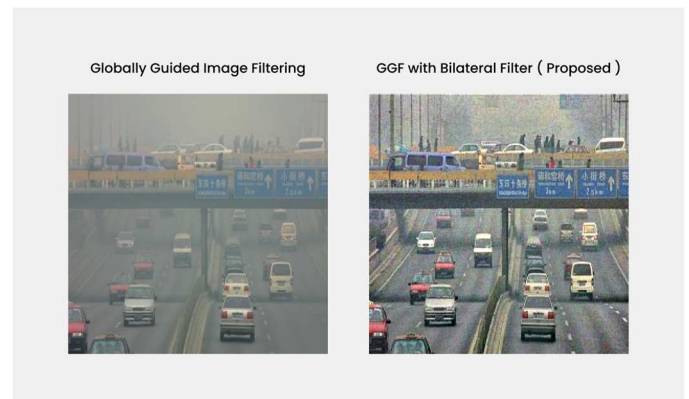


Fig. 8. Image 1 Output

In the first image, the filter performed a remarkable denoising job on the image, significantly reducing the noise levels and enhancing its overall clarity. It effectively preserved fine details and textures while smoothing out unwanted noise artifacts, resulting in a clean and visually appealing image.

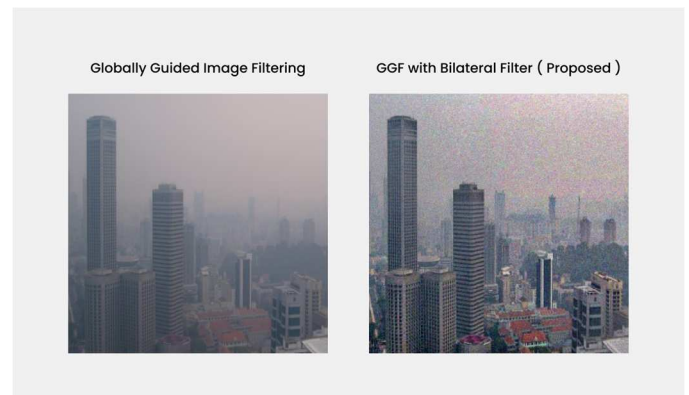


Fig. 9. Image 2 Output

In image 2, the filter achieved impressive dehazing results on the image, substantially reducing noise levels and improving overall clarity. It successfully preserved intricate details and textures while effectively eliminating unwanted haze, resulting in a visually pleasing and clean image.

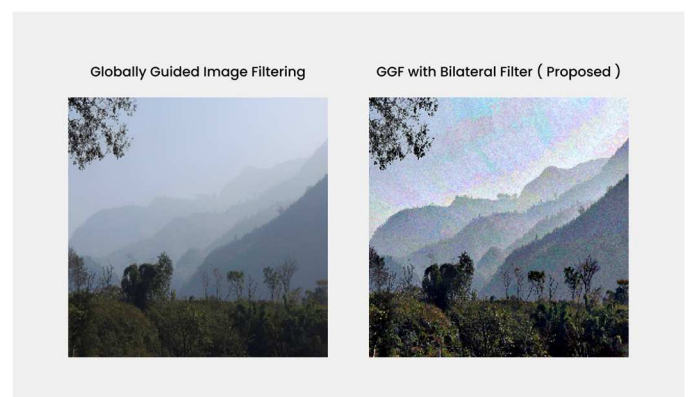


Fig. 10. Image 3 Output

In the third image, the filter successfully cleared and improved my image, effectively eliminating unwanted artifacts and

enhancing its overall quality. It significantly reduced noise levels, resulting in a cleaner and more refined appearance. The filter's application resulted in a noticeable improvement, with enhanced clarity and improved visual appeal.

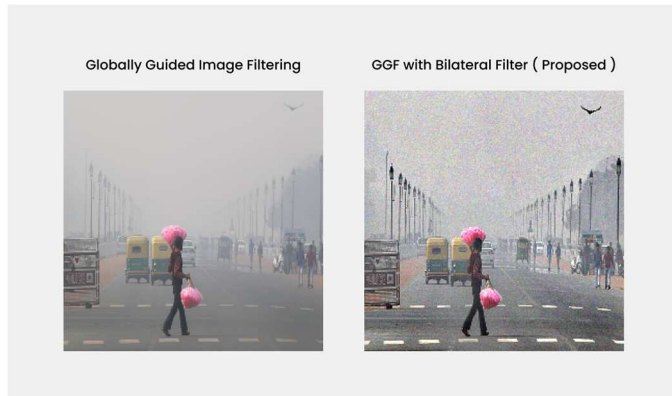


Fig. 11. Image 4 Output

In the fourth image, the filter efficiently dehazed my image, effectively removing atmospheric haze and improving visibility. It restored the lost details and enhanced the overall clarity of the scene, resulting in a significantly clearer and more vibrant image.

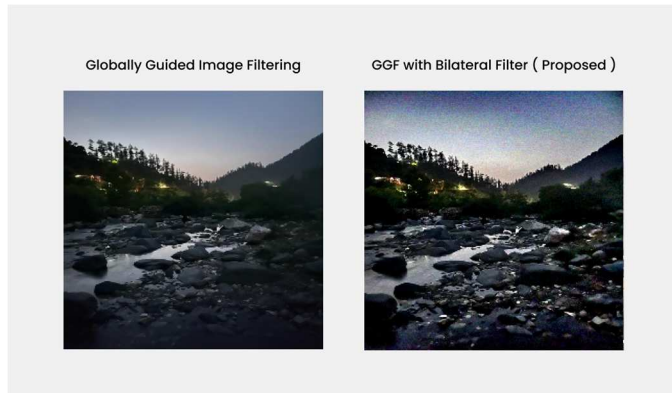


Fig. 12. Image 5 Output

The filter effectively cleared and enhanced my image by eliminating undesirable artifacts and improving its overall quality. It successfully reduced noise levels, resulting in a more polished and refined look. Applying the filter noticeably improved the image, enhancing its clarity and visual appeal.

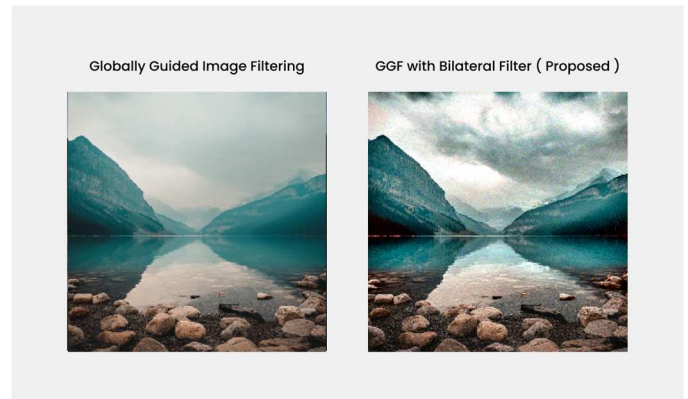


Fig. 13. Image 6 Output

The filter significantly enhanced the clarity of my image, bringing out sharper details and improving overall definition. It effectively reduced blurriness and enhanced the crispness of the visual elements. The application of the filter resulted in a noticeably clearer and more distinct image with improved visual clarity.

## V. CONCLUSION

In conclusion, an enhanced GGIF, WLS, and 2D Bilateral Filtering are effectively merged into a hybrid filter, which results to a large improvement in both the image quality and the perceived fog density. This is achieved by a combination of the improved GGIF, WLS, and 2D Bilateral Filtering. This method is able to effectively resolve the problems of denoising and dehazing, providing a powerful solution for enhancing photos in a range of applications including photography, computer vision, and surveillance systems. For the future scope we will implement this technique using deep learning approaches.

## REFERENCES

- [1] Y. Chen, Y. Zhang, & S. Zhang, (2019). An adaptive guided image filtering-based method for image denoising. *Multimedia Tools and Applications*, 78(20), 28839-28858.
- [2] H. Li, Q. Xie, & Y. Li, (2020). Image denoising using a hybrid filter based on guided filtering and non-local means filtering. *Journal of Visual Communication and Image Representation*, 66, 102661.
- [3] X. Luo, M. Hu, & Y. Zhang, (2020). A deep learning-based method for image denoising using hybrid filters. *Signal, Image, and Video Processing*, 14, 141-148.
- [4] R. Wang, Y. Wang, & X. Zhang, (2020). A novel hybrid filtering method for image denoising using guided filter and BM3D. *Multimedia Tools and Applications*, 79(5-6), 3451-3468.
- [5] K. Zhang, W. Zuo, & L. Zhang, (2020). Deep plug-and-play super-resolution for arbitrary blur kernels. *IEEE Transactions on Image Processing*, 29, 2970-2985.
- [6] Zhengguo Li "Single Image De-Hazing Using Globally Guided Image Filtering" *IEEE Transactions on Image Processing*, Vol. 27, No. 1, January 2018.
- [7] W. Wang and X. Yuan, "Recent advances in image dehazing," *in IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 3, pp. 410-436, 2017, doi: 10.1109/JAS.2017.7510532.
- [8] Z. Li, H. Shu and C. Zheng, "Multi-Scale Single Image Dehazing Using Laplacian and Gaussian Pyramids," *in IEEE Transactions on Image Processing*, vol. 30, pp. 9270-9279, 2021, doi: 10.1109/TIP.2021.3123551.

- [9] L. Shen, Y. Zhao, Q. Peng, J. C. -W. Chan and S. G. Kong, "An Iterative Image Dehazing Method With Polarization," in *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1093-1107, May 2019, doi: 10.1109/TMM.2018.2871955.
- [10] Q. Wu, W. Ren and X. Cao, "Learning Interleaved Cascade of Shrinkage Fields for Joint Image Dehazing and Denoising," in *IEEE Transactions on Image Processing*, vol. 29, pp. 1788-1801, 2020, doi: 10.1109/TIP.2019.2942504.
- [11] S. Guo, Z. Liang and L. Zhang, "Joint Denoising and Demosaicking With Green Channel Prior for Real-World Burst Images," in *IEEE Transactions on Image Processing*, vol. 30, pp. 6930-6942, 2021, doi: 10.1109/TIP.2021.3100312.
- [12] M. Ju, C. Ding, C. A. Guo, W. Ren and D. Tao, "IDRLP: Image Dehazing Using Region Line Prior," in *IEEE Transactions on Image Processing*, vol. 30, pp. 9043-9057, 2021, doi: 10.1109/TIP.2021.3122088.
- [13] S. Kanti Dhara, M. Roy, D. Sen and P. Kumar Biswas, "Color Cast Dependent Image Dehazing via Adaptive Airlight Refinement and Non-Linear Color Balancing," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 2076-2081, May 2021, doi: 10.1109/TCSVT.2020.3007850.
- [14] S. C. Agrawal and A. S. Jalal, "Dense Haze Removal by Nonlinear Transformation," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 593-607, Feb. 2022, doi: 10.1109/TCSVT.2021.3068625.
- [15] W. Liu, P. Zhang, Y. Lei, X. Huang, J. Yang and M. Ng, "A Generalized Framework for Edge-Preserving and Structure-Preserving Image Smoothing," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6631-6648, 1 Oct. 2022, doi: 10.1109/TPAMI.2021.3097891.
- [16] J. Xu, Z. -A. Liu, Y. -K. Hou, X. -T. Zhen, L. Shao and M. -M. Cheng, "Pixel-Level Non-local Image Smoothing With Objective Evaluation," in *IEEE Transactions on Multimedia*, vol. 23, pp. 4065-4078, 2021, doi: 10.1109/TMM.2020.3037535.
- [17] W. -C. Tu and S. -Y. Chien, "Two-Way Recursive Filtering," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 11, pp. 4255-4268, Nov. 2021, doi: 10.1109/TCSVT.2021.3049833.
- [18] W. Liu, P. Zhang, X. Chen, C. Shen, X. Huang and J. Yang, "Embedding Bilateral Filter in Least Squares for Efficient Edge-Preserving Image Smoothing," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 23-35, Jan. 2020, doi: 10.1109/TCSVT.2018.2890202.
- [19] R. G. Gavaskar and K. N. Chaudhury, "Fast Adaptive Bilateral Filtering," in *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 779-790, Feb. 2019, doi: 10.1109/TIP.2018.2871597.
- [20] R. G. Gavaskar and K. N. Chaudhury, "Fast Adaptive Bilateral Filtering," in *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 779-790, Feb. 2019, doi: 10.1109/TIP.2018.2871597.

# Design and Analysis of LLC Resonant Converter for Electric Vehicle Battery Charging

Shreyas

CoE for Electric Vehicle and Related Technologies  
Department of Electrical Engineering  
Delhi Technological University, Delhi, India  
shreyask628@gmail.com

Mayank Kumar (Senior Member IEEE)

CoE for Electric Vehicle and Related Technologies  
Department of Electrical Engineering  
Delhi Technological University, Delhi, India  
mayankkumar@dtu.ac.in

**Abstract**— In order to achieve wide input range, minimal switching losses, and stable performance designing of resonant converters are required. This paper suggests effective methodology of battery charging using resonant converter. Resonant power converters have become quite popular recently in applications requiring solid-state transformers, solar PV fed electric vehicle (EV) charging infrastructures, etc. An LLC DC/DC resonant converter is designed and proposed in this paper for the utilization of battery charging. The battery is charged using closed loop constant current charging technique. This approach indicates the requirements of fine-tuning of PI controller for constant current application, which is used for EV battery charging application.

**Keywords**— Constant current battery charging, LLC resonant converter(inductor-inductor-capacitor), PI controller, zero voltage switching.

## I. INTRODUCTION

Internal combustion (IC) engines have been the only type of engines used in international transportation networks for a very long period. Diesel, gasoline serve as the primary sources of energy for the vehicle engine in conventional transportation propulsion systems [1]. Due to the hazardous CO<sub>2</sub> emissions emitted during combustion, the usage of such fuels results in health issues. As a result, recent research and attention have focused heavily on electric and hybrid vehicles (HEV). It is done to reduce the hazardous gases emitted during the combustion of these fuels and create the foundation for a clean and energy-efficient transportation system. In order to fulfil the requirement of power in the electric vehicle transportation, rechargeable batteries are essential in hybrid and electric automobiles. The rechargeable batteries are changing the way of designing Electric vehicles. As the batteries are dc operated that always require dc-dc converter for controlling the power stages. Developing countries are replacing fuel driven vehicles with traditional, less expensive EVs which are driven by lead acid and Li-ion batteries.

The fast-charging mechanism is the current demand of EV nowadays. To decrease the charging time an EV, fast charging stations are installed and used suitably [2]. for reducing charging time of EV, fast charging stations and choice of battery packs are important. Modern lithium ion and lead acid batteries packs, which use fast charging that providing longer travel distances with shorter duration of charging time. Thus, effective and quick charging method is always a topic of research area for EV batteries.

The inductor-inductor-capacitor (LLC) resonant converter offers numerous advantages, including the capacity to maintain the output parameters constant over a large range of extension of load and line fluctuations with only a modest variation in switching frequency. Over the full operational range, it can accomplish zero voltage switching (ZVS). All

necessary parasitic elements of semiconductor devices are used to achieve soft switching [3]. It is easier to charge the batteries through full-bridge inductor-inductor-capacitor (LLC) resonant dc-dc converter. This converter provides moderate frequency variation, low value of current/voltage stress at switches and achieves ZVS & ZCS operation to mitigate switching losses [4]. These parameters define the different topology of resonant power converter.

The proportional integral (PI) controller and pulse frequency modulation (PFM) techniques are utilized to manage the battery's voltage and current at output side of converter. LLC resonant converter receives power from DC voltage and transferred to load. The output voltage and output current of the converter is controlled by the PI-controller [4]. PFM controller executes desired switching signal, that provides gating signal to converter for switching. The overcharging situations in rechargeable batteries can be avoided by constant current charging method [5]. The phenomenon of overcharging is quite common nowadays and requires instant solution to avoid the heating of batteries during running conditions. As the temperature varies place to place at different interval of time, so designing of these converters are important that avoid overcharging.

## II. CONVERTER DESIGN

### A. Resonant Power Converter

Resonant power converter mainly comprises three stages including square wave generator, resonant tank circuit and bridge rectifier with filter capacitance. By alternately applying a pulse width modulated gating signal using PI controller (approx. 50% duty cycle) to switches S<sub>1</sub>, S<sub>4</sub>, and S<sub>2</sub>, S<sub>3</sub>, the square wave generator creates a voltage that is square in nature. Typically, a little dead period is included in between each subsequent transition. The square wave is generated by a full-bridge converter [6].

Magnetizing inductance component of transformer, leakage inductance and capacitor make up the resonant network. The magnetizing branch of the inductance, which serves as a shunt inductor is represented by L<sub>m</sub>, resonant inductor and resonant capacitor is represented by L<sub>r</sub> and C<sub>r</sub> respectively. This resonant network helps to achieve resonance condition. Resonant network filters out higher harmonic currents and helps to achieve resonance and receives a square nature of voltage waveform through inverter end. The resonance condition is quite crucial in designing the converter that maintains the switching losses. This resonance condition is achieved by the suitable values of resonant inductor and resonant capacitor with shunt inductance of transformer connected in parallel. At rectifier end of the resonant converter creates dc voltage from the ac input. The output filter capacitance of the rectifier is used to remove high

frequency ripples [4], [7]. At the rectifier end is connected with battery pack for the application of charging [6].

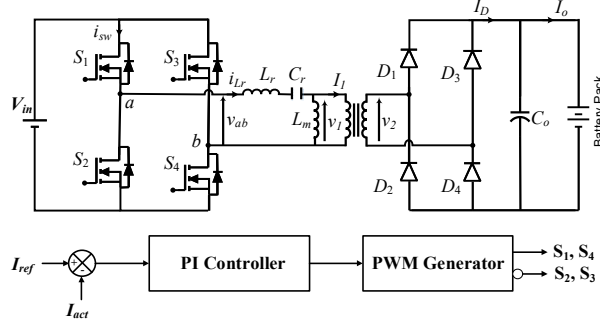


Fig 1. LLC resonant converter with rechargeable battery.

#### A. Operating frequency region

*Case 1:*  $f_s = f_r$ : Each cycle of switching delivers power, therefore each half switching cycle is involved [1]. At half cycle of switching, rectifier current is zero, and the magnetising current is achieved due to the presence of resonant inductor. Resonant tank has unity gain and designed for most optimal operation. The efficiency at this point and the turns ratio of isolation transformer is maintained that specific converter can function properly for nominal values of input and output voltages [8].

*Case 2:*  $f_s > f_r$ : Each cycle of switching involves small amount of power delivery action, which is analogous to resonant frequency operation. Additionally, before resonant half cycle is finished, second half cycle of switching starts. Secondary rectifier diodes are harmed by strong commutation, whereas primary side MOSFETs experiences a great turn-off loss. This case of operation is intended for buck operation due to a higher input voltage rating.

*Case 3:*  $f_s < f_r$ : Power delivery action is completed during each half of the switching cycle. When the switching half cycle is completed, the freewheeling action starts and lasts until the magnetising current is reached by the resonant inductor current  $I_{Lr}$ . Circulating current increases the value of conduction losses of converter. Converter runs in this mode when boost operation is required due to a lower input voltage.

#### B. Soft switching

Converters can implement soft switching in a number of different ways. The goal is to produce a forced swing using LC transients. As a result, soft switching activates and deactivates the electrical switch using an LC resonant circuit. The current and voltage waveform intersection is minimized by controlling the switching timing [9]. It is crucial to eliminate power losses to improve efficiency. Moreover, it helps in lowering inductance, switching losses, and diode losses. ZVS and ZCS are switched as part of the process. In fact, the electronic switch utilizes the resonance phenomenon to switch on and off under soft switching conditions. The ability of switches to turn on and off at zero (or almost zero) voltage or current reduces switching losses and improves converter efficiency. Fig. 2 represents switching power loss in IRF530 during turn-ON. As a result, to obtain precise coordination between the multiple waveforms, soft switching approaches require more complicated control circuits [10].

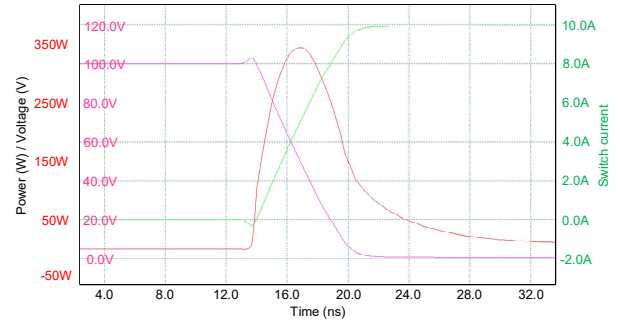


Fig. 2. Switching power loss in IRF530 during switch-ON condition.

#### C. Design Calculation

Fig. 3 represents equivalent circuit of LLC resonant converter. The equivalent circuit comprises of  $R_{ac}$  which depends on the turns ratio of isolation transformer and value of equivalent load resistance. This equivalent circuit is designed by considering the effect of transformer and load resistance. Due to this equivalent circuit simplification to resonant tank circuit becomes easy. The value of  $R_{ac}$  is calculated and presented in Eq (3).

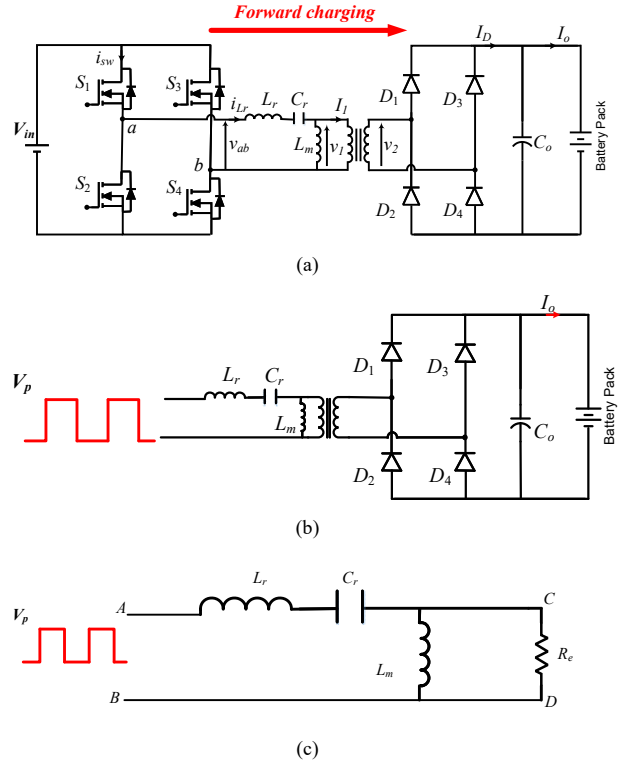


Fig 3. LLC resonant converter (a) connected with rechargeable battery; (b) output of square wave connected with resonant tank fed rectifier circuit; (c) ac equivalent circuit with  $R_{ac}$  load

For designing resonant converter, The minimum and maximum converter voltage gain values, along with the appropriate transformer turns ratio are chosen. The nominal value of voltage gain is considered to be unity.

$$M_{nom} = 1; \quad \frac{N_p}{N_s} = 1.25 \quad (1)$$

where;

$M_{nom}$  = nominal voltage gain,

$N_p = 120$  and  $N_s = 96$  are the primary and secondary turn ratio.



$$M_{\max} = \left( \frac{V_{in\_nom}}{V_{in\_min}} \right) M_{nom} = 1.04 \quad (2a)$$

$$M_{\min} = \left( \frac{V_{in\_nom}}{V_{in\_max}} \right) M_{nom} = 0.96 \quad (2b)$$

where;

$M_{\max}$  = maximum voltage gain,

$V_{in\_nom}$ ,  $V_{in\_min}$ ,  $V_{in\_max}$  are nominal, minimum and maximum input voltages respectively.

The proper values of gain are obtained by selecting the nominal voltage, min nominal voltage and max nominal voltage of the converter specifications. The values of resonant tank circuit are calculated by selecting quality factor ( $Q_{max}$ ) and gain ( $m$ ) value. The value of  $Q_{max}$  is considered as 0.4 and represented by Fig. 4. The value of  $m$  is considered high due to following reasons –

- Higher the value of efficiency
- Higher magnetizing inductance
- Lower magnetizing circulating current

The proper values of  $Q_{max}$  and  $m$  specifies the performance of converter as well as soft switching behaviour. The converter's switching determines power losses, which have an impact on the converter's efficiency. Higher the switching losses, decreases the efficiency and performance at higher frequency operation. Converters at such high frequency of 100 kHz require fast and instant switching which enables the resonant mode of operation.

Step1: Selecting the  $Q_{max}$  value; i.e.  $Q_{max} = 0.4$

Step 2: Selecting the  $m$  value; i.e.  $m = 6.3$ .

Step3: Calculating resonant components value

$$R_e = \left( \frac{8}{\pi^2} \right) \times \left( \frac{N_p}{N_s} \right)^2 \times \left( \frac{V_o^2}{P_o} \right) = 22.52 \text{ ohms} \quad (3)$$

Step 4: Calculation of resonant capacitance:

$$C_r = \frac{1}{2\pi Q_{max} R_e} = 19.03 \text{ nF} \quad (4)$$

Step 5: Calculation of resonant inductance:

$$L_r = \frac{1}{4\pi^2 f_o^2 C_r} = 120 \mu\text{H} \quad (5)$$

The specification of resonant converter is depicted in Table I.

TABLE I. SPECIFICATIONS OF RESONANT CONVERTER

Parameters	Variables	Values
DC-input Voltage	$V_d$	120 V
Resonant Frequency	$f_r$	89.7 kHz
Switching Frequency	$f_{sw}$	100 kHz
Resonant Inductance	$L_r$	120 $\mu\text{H}$
Resonant Capacitance	$C_r$	22.52 nF
Filter Capacitance	$C_o$	470 $\mu\text{F}$
Output current	$I_o$	15.625 A
Battery capacity		15.625 Ah
Battery type		Lithium ion (C1 Type)
Battery nominal voltage	$V_b$	96
Initial state of charge		45%

Switching frequency is higher than the resonant frequency which forces the desired converter to operate in buck mode. Thus, battery rating of 96 volts is charged by the dc input

voltage of 120 volts. The filter capacitance is considered at high value of 470 $\mu\text{F}$ ; which eliminates the higher order harmonics at output voltage. The resonance condition is also achieved to mitigate the losses due to switching of the converter at such high frequency of 100 kHz.

The gain of designed resonant converter is crucial in the designing part. Gain of the converter is compared with respect to range of frequencies for operation. This defines the range of operating frequency at which highest quality factor is achieved.

Converter gain is the function of frequency and value of  $m$  respectively.

Transfer function of full bridge LLC resonant converter

The transfer function of Full Bridge LLC resonant converter is represented by the ac equivalent circuit with  $R_{ac}$  load as mentioned in Fig. 3(c). It is represented by the voltage across CD terminals and voltage across AB terminals of the given circuit diagram and represented by Eq (6).

Where  $X_{Lr} = \omega L_r$ ,  $X_{Cr} = 1/\omega C_r$ ,  $X_m = \omega L_m$

$$\frac{V_{CD}}{V_{AB}} = \frac{jX_m R_e}{jX_m R_e + (jX_{Lr} - jX_{Cr})(jX_m R_e)} \quad (6)$$

$$\frac{V_{CD}}{V_{AB}} = \frac{1}{1 + j \frac{X_{Lr}}{R_e} - j \frac{X_{Cr}}{R_e} + \frac{X_{Lr}}{X_m} - \frac{X_{Cr}}{X_m}}$$

$$\frac{V_{CD}}{V_{AB}} = \frac{1}{(1 + \frac{X_{Lr}}{X_m} - \frac{X_{Cr}}{X_m}) + j(\frac{X_{Lr}}{R_e} - \frac{X_{Cr}}{R_e})}$$

$$\frac{V_{o1}}{V_{in1}} = \frac{8/\pi^2}{\sqrt{1 + (\frac{\omega L_r}{\omega_{Lm}} - \frac{1/\omega C_r}{\omega_{Lm}})^2 + (\frac{\omega L_r}{R_e} - \frac{1/\omega C_r}{R_e})^2}}$$

$$\frac{V_{o1}}{V_{in1}} = \frac{8/\pi^2}{\sqrt{(1 + \frac{L_r}{L_m} - \frac{L_r}{\omega^2 C_r L_r L_m})^2 + (\frac{\omega L_r}{R_e} - \frac{1}{\omega C_r R_e})^2}} \quad (7)$$

Putting the value,

$$Q = \frac{\omega_o L_r}{R_e} = \frac{1}{\omega_o C_r R_e}$$

$$\omega_o = \frac{1}{\sqrt{L_r C_r}}, \omega_x = \frac{\omega}{\omega_o}$$

where  $\omega_x$  is normalized switching frequency that is defined by ratio of switching to resonant frequency-

$$\frac{V_{o1}}{V_{in1}} = \frac{8/\pi^2}{\sqrt{(1 + \frac{L_r}{L_m} (1 - \frac{1}{\omega_x^2}))^2 + Q^2 (\omega_x - \frac{1}{\omega_x})^2}}$$

$$\frac{V_{o1}}{V_{in1}} = \frac{(8/\pi^2) \omega_x^2}{\sqrt{(\omega_x^2 + (\frac{L_r}{L_m} (\omega_x^2 - 1))^2 + Q^2 \omega_x^2 (\omega_x^2 - 1)^2}}$$

$$\frac{V_{o1}}{V_{in1}} = \frac{(8/\pi^2)\omega_x^2}{\sqrt{(\omega_x^2 + (\frac{L_r}{L_m}(\omega_x^2 - 1))^2 + Q^2\omega_x^2(\omega_x^2 - 1)^2}} \quad (8)$$

Eq (8) shows the complete transfer function in terms of Quality factor, ratio of resonant to magnetizing inductance and operating frequency of the proposed converter. Based upon its number of curves are drawn at different values of  $m$  to decide the range of the Gain. This also defines the operating region during operation.

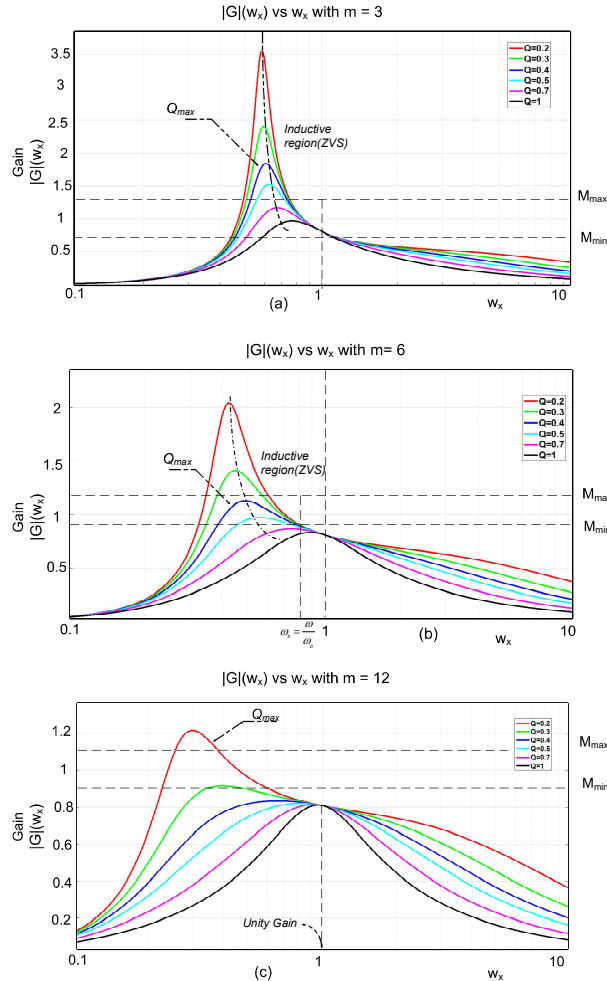


Fig 4. Gain versus frequency curve of resonant converter at (a)  $m = 3$ ,  $Q = 0.7$  (b)  $m = 6$ ,  $Q = 0.4$  (c)  $m = 12$ ,  $Q = 0.2$ .

The curve in Fig. 4 is plotted at different values of  $m$  that depicts the operating region of the converter at a specific value of  $Q$ . Inductive region operation is being used by the converter to achieve ZVS condition. The maximum value of  $Q$  is represented by  $Q_{max}$ . As the  $Q$  and frequency changes, the operating region of converter also changes. Thus, it is very crucial to optimize the quality factor value of  $Q_{max}$  for better operating condition and achieving resonance. The steps to design the LLC converter is presented in Fig. 5.

Load independent gain characteristics offered by LLC-resonant converter in lagging region helps to sustain ZVS and constant switching frequency operation. The parallel inductor placed parallel to transformer is crucial for maintaining ZVS under light load conditions and maintaining the magnitude of inrush current at the beginning.

This section describes how design factors influence voltage regulation and efficiency performance of resonant converter. It simplifies the design and helps to choose the values of converter that performs resonance. The final design objective is to meet gain requirement for all line and load regulations during charging the battery. These steps are executed to design the converter which is suitable to operate at wide range of frequencies as well as different loading conditions that made it suitable for EV charging applications. Soft switching and resonance condition is obtained by resonant tank circuit. The PI controller is going to be tuned perfectly for achieve constant value of current at the output side of converter during charging.

#### Modes of operation of proposed converter

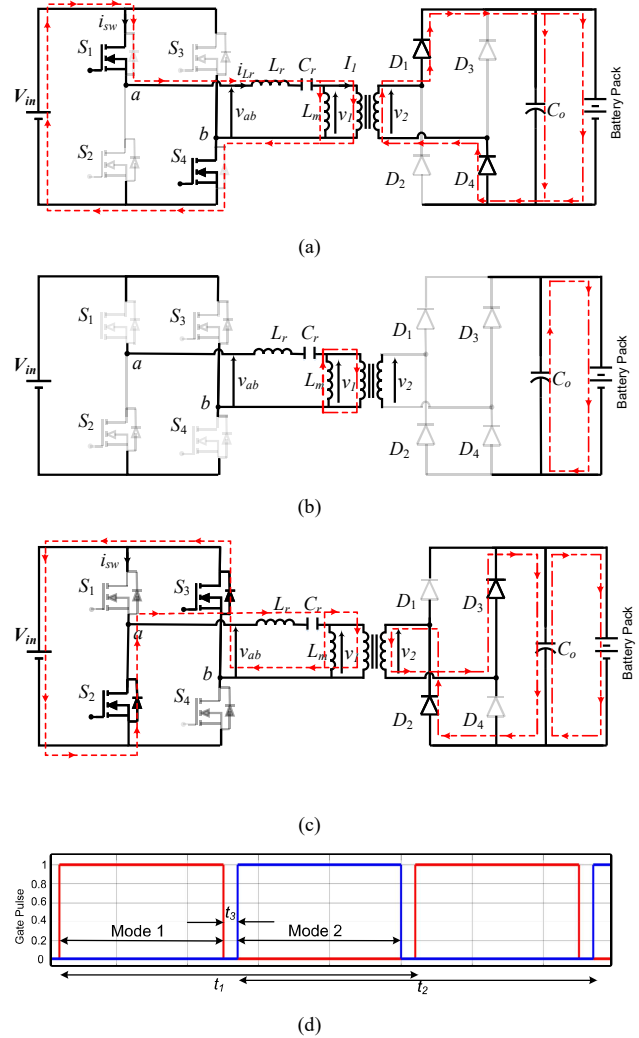


Fig 5. (a) Switching condition of  $S_1, S_4$  (Mode1) (b) Switching condition of  $S_2, S_3$  (Mode2) (c) Dead band region

As shown in Fig 5(a),  $S_1, S_4$  switch on at  $t = t_1$  that allows the input current to pass to load through resonant tank circuit and achieves ZVS at the end.

As shown in Fig 5(b),  $S_1, S_4$  are cut off at  $t_3$  while  $S_2, S_3$  still remain at the cut off state. This forms a dead time region where all the switches are in their off state.

As shown in Fig 5(c),  $S_2, S_3$  switches are operated at  $t = t_2$  and achieve ZVS while  $S_1, S_4$  remain at off state. The primary current reaches to load in the same manner that allows the battery to charge at the desired level of voltage.

### III. BATTERY CHARGING USING PI CONTROLLER

There are mainly two methods to charge a battery: constant voltage, constant current charging methods. Both charging methods are employed based on the requirement and application.

*A. Constant current charging method:* The value of current remains constant while charging the battery. The level of current signal is considered at approximately 9.85% of the maximum rating of battery during charging the batteries at constant value of current. The main drawback of long duration of charging is that as the battery is overcharged that may increase the temperature of battery which further overheats and require instant replacement of battery [11]. This constant current charging technique is executed with lithium ion and lead acid type of batteries.

*B. Constant voltage charging method:* In order to avoid overcharging, battery may also charged at constant voltage. The power supply maintains a constant voltage as long as, the charger provides a complete path to flow the full value of current through the battery. The value of current begins to decrease to the least min value of threshold voltage that is attained by the converter [11]. Lead acid batteries can be charged quickly with this methodology [12].

*C. The optimal charging is achieved by fine tuning of PI Controller.* The closed loop converter helps to stabilize the output values to a desired value which is further implemented to obtain charging conditions. PWM and a voltage controller are used, which is utilized by the standard LLC resonant converter. The PI controller processes the voltage error ( $V_{oe}$ ) that is produced by the difference created by the given reference output voltage and standard battery voltage. The iteration is executed till values near to the standard values is achieved. As it approaches to the standard values, the iterations are stopped and mark the values of PI tuning. All these values indicate best optimization of error signal which is generated by comparing the actual and standard signal. [13]

*PI Controller:* The PI controller mainly consists of function values of proportional and integral. PI controller which is implemented for constant current charging is represented by the following equation that represents the proportional and integral values of the function  $u(t)$ . The error signal is achieved by difference of output signal and reference signal for the converter. The error signal is directly fed back to the system that stabilizes the output at constant value. This makes the converter stable and provide a constant current or constant voltage operation [13].

The function  $u(t)$  is represented by controlling signal and  $e(t)$  is represented by error signal. The controlling signal is defined as the combination of proportional and integral parameters of the converter. P-term and I-term are all contributing to rectify the error output. The P-term is proportional to the error signal, the I-term is the integral to the error signal. The parameters of the controller are represented by proportional gain  $K_p$ , integral gain  $K_i$ .

$$u(t) = k \left[ e(t) + \frac{1}{T_i} \int_0^t e(T) dT \right] \quad (9)$$

where,  $T_i$  stands for integral time of the specified design of PI controller. The integral parameter of (9) represents previous value of errors and the proportional function estimates error corresponding to present value [14]. Thus, parameters of

proportional and integral are quite important to tune PI controller after number of iterations.

This term “current error ( $I_e$ )” refers to the discrepancy between a real current value and its standard current value of converter. The resulting current error ( $I_e$ ) signal is passed to PI controller that standardize the value of current for constant charging. Thus, it is important to check the error signal of current at each iteration that defines the accuracy of standardization method. Mathematically, it can be represented as follows [13]:

$$d(x) = d(x-1) + G_{pi} \{I_e(x) - I_e(x-1)\} + G_{ii} I_e(x) \quad (10)$$

where  $G_{pi}$  represents proportional gain and  $G_{ii}$  represents integral gain of the PI controller for constant charging.

The output is expressed as follows [13]:

$$f(x) = f(x-1) + G_{pg} \{V_{oe}(x) - V_{oe}(x-1)\} + G_{ig} V_{oe}(x) \quad (11)$$

Where  $G_{pg}$  represents proportional gain value and  $G_{ig}$  represents integral gain of the given converter.

Equations (9), (10), and (11) help to define the values of proportional and integral of the converter.

### IV. SIMULATION RESULTS

LLC resonant converter of 1500 watts is designed and modelled with MATLAB/SIMULINK software for output voltage rating of 96 V and constant current of 15.625 A. The converter is used for charging of Lithium-ion battery in constant current mode. System parameters are presented in Table I. Fig 6 represents the output current and voltage of LLC resonant converter. The ripple in the output voltage and output current are minimum which is suitable for battery charging application. This model can be applied in EV level 1 charger (upto 3.3kW) and upon extending the rating of converter can be applied in level 2 and level 3 chargers.

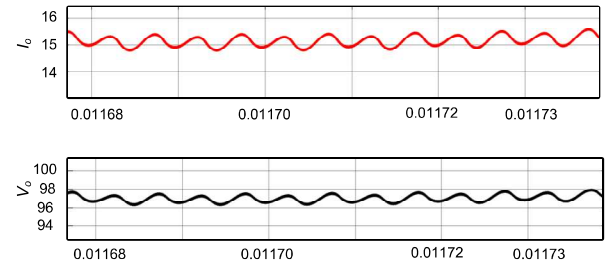


Fig. 6. Output current, output voltage

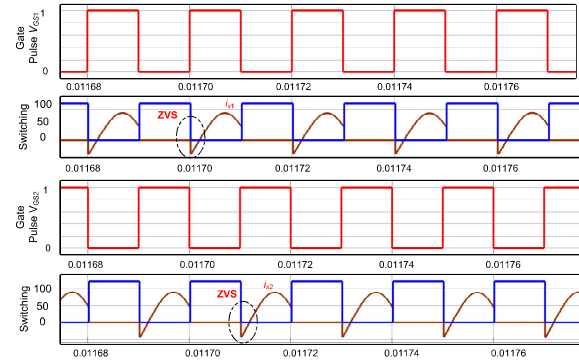


Fig. 7. Gate Pulse, Switch voltage and Switch current across MOSFET

Zero voltage switching behavior is achieved by switching the MOSFET voltage and current at a particular instant of time. The switching is performed to maintain the losses at their

minimum level at such high frequency operation and provides soft switching to the converter. This switching current is obtained through  $S_2, S_3$  switches and the same action is achieved with  $S_1, S_4$  switches during next cycle of gating pulse of converter as mentioned in Fig 7. The zero-voltage switching is obtained that mitigates the switching losses and improves the overall efficiency of converter for the application of EV charging.

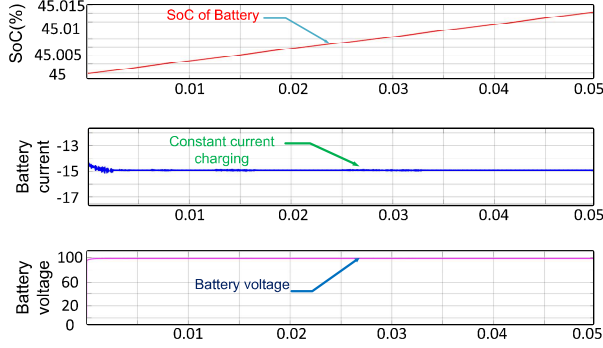


Fig. 8. SoC (%), constant charging current, battery voltage.

Fig. 8. representing SoC of battery which shows the charging case of the converter by the linearly increasing graph. The constant negative value of current represents behaviour of constant current charging method. Battery voltage and current values are maintained at constant level of 96 volts and 15.625 A.

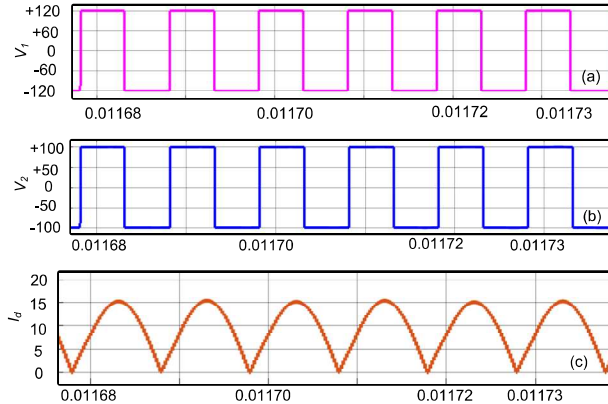


Fig. 9. (a) Input voltage at primary side( $V_1$ ) (b) Output voltage at secondary side( $V_2$ ) and (c) Diode current at secondary side ( $I_d$ )

Fig 9 represents the voltage at the primary and current at the secondary side of transformer. A high frequency operated transformer is used due to light weightage and comparatively small in size and can be implemented in the level 1 charging. Not just it transforms the input voltage and input current but also provides galvanic isolation.

## V. CONCLUSION

LLC full bridge resonant power converter is developed for the application of charging. The charging methodology is based on the constant current charging. Battery is charged at the charging current of value 15.625 A and battery voltage of 96 volts. This constant current charging method avoids the overheating problem. PI controller is tuned properly using closed loop configuration for the high switching operation and better efficiency.

## VI. ACKNOWLEDGEMENT

This research supported by the Science and Engineering Research Board (SERB), Department of Science & Technology, Government of India, under the SERB sanction order number SRG/2021/001640.

## REFERENCES

- [1]. R. -L. Lin and C. -W. Lin, "Design criteria for resonant tank of LLC DC-DC resonant converter," IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society, Glendale, AZ, USA, 2010, pp. 427-432, doi: 10.1109/IECON.2010.5674988.
- [2]. A. Bouach, S. Mariétoz and T. Delaforge, "Series Resonant Converter for DC fast-charging electric vehicles with wide output voltage range," 2019 21st European Conference on Power Electronics and Applications (EPE '19 ECCE Europe), Genova, Italy, 2019, pp. P.1-P.8, doi: 10.23919/EPE.2019.8914828.
- [3]. Y. Wei and A. Mantooth, "A Flexible Resonant Converter Based Battery Charger with Power Relays," 2021 IEEE Energy Conversion Congress and Exposition (ECCE), Vancouver, BC, Canada, 2021, pp. 1675-1680, doi: 10.1109/ECCE47101.2021.9595497.
- [4]. S. Wang, Y. Liu and X. Wang, "Resonant Converter for Battery Charging Applications With CC/CV Output Profiles," in IEEE Access, vol. 8, pp. 54879-54886, 2020, doi: 10.1109/ACCESS.2020.2981595.
- [5]. B. -H. Liu, J. -H. Teng and S. -S. Chen, "Novel H LLC Resonant Converter with Variable Resonant Inductor," 2022 IEEE IAS Global Conference on Emerging Technologies (GlobConET), Arad, Romania, 2022, pp. 327-331, doi: 10.1109/GlobConET53749.2022.9872360.
- [6]. R. Kodoth, T. Harikrishnan, K. R. Bharath and P. Kanakasabapathy, "Design and Development of a Resonant Converter Adapted to Wide Output Range in EV Battery Chargers," 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India, 2018, pp. 1018-1023, doi: 10.1109/RTEICT42901.2018.9012426.
- [7]. F. Musavi, M. Edington, W. Eberle, and W. G. Dunford, "Evaluation and efficiency comparison of front-end AC-DC plug-in hybrid charger topologies," IEEE Transactions on Smart Grid, vol. 3, no. 1, pp. 413-421, 2012.
- [8]. J. Deng, S. Li, S. Hu, C. C. Mi and R. Ma, "Design Methodology of LLC Resonant Converters for Electric Vehicle Battery Chargers," IEEE Transactions on Vehicular Technology, vol. 63, no. 4, pp. 1581- 1592, May 2014.
- [9]. G. Spiazzi, "Analysis and design of the soft-switched clamped-resonant interleaved boost converter," in CPSS Transactions on Power Electronics and Applications, vol. 4, no. 4, pp. 276-287, Dec. 2019, doi: 10.24295/CPSSPEA.2019.00026.
- [10]. B. Akhlaghi and H. Farzanehfard, "Family of Soft Switching Quasi-Resonant Interleaved Converters," 2022 13th Power Electronics, Drive Systems, and Technologies Conference (PEDSTC), Tehran, Iran, Islamic Republic of, 2022, pp. 473-478, doi: 10.1109/PEDSTC53976.2022.9767232.
- [11]. T. N. Gücin, M. Biberoğlu and B. Fincan, "A Constant-Current Constant-Voltage Charging based control and design approach for the parallel resonant converter," 2015 International Conference on Renewable Energy Research and Applications (ICRERA), Palermo, Italy, 2015, pp. 414-419, doi: 10.1109/ICRERA.2015.7418447.
- [12]. T. N. Gücin, M. Biberoğlu and B. Fincan, "Constant frequency operation of parallel resonant converter for constant-current constant-voltage battery charger applications," in Journal of Modern Power Systems and Clean Energy, vol. 7, no. 1, pp. 186-199, January 2019, doi: 10.1007/s40565-018-0403-7.
- [13]. C. Buccella, C. Cecati, H. Latafat and K. Razi, "Comparative transient response analysis of LLC resonant converter controlled by adaptive PI and fuzzy logic controllers," IECON 2012 - 38th Annual Conference on IEEE Industrial Electronics Society, Montreal, QC, Canada, 2012, pp. 4729-4734, doi: 10.1109/IECON.2012.6389483.
- [14]. M. I. Shahzad, S. Iqbal and S. Taib, "LLC series resonant converter with PI controller for battery charging application," 2014 IEEE Conference on Energy Conversion (CENCON), Johor Bahru, Malaysia, 2014, pp. 84-89, doi: 10.1109/CENCON.2014.6967481.
- [15]. R. Mssaurya and R. Saha, "Design and Simulation of an Half-Bridge LLC Resonant Converter for Battery Charger in EV," 2022 IEEE Delhi Section Conference (DELCON), New Delhi, India, 2022, pp. 1-9, doi: 10.1109/DELCON54057.2022.9753654.

# Design of 50 kW Two Stage off-Board EV Charger using CC-CV Algorithm

Kushank Singh  
Department of Electrical Engineering  
Delhi Technological University  
Delhi, India  
singhkushank87@gmail.com

Vanjari Venkata Ramana  
Department of Electrical Engineering  
Delhi Technological University  
Delhi, India  
venkat.vr90@gmail.com

**Abstract**— In this paper a 50 kW two stage off-Board EV charger is designed for charging a lithium ion battery using constant current (CC)-constant voltage (CV) algorithm. First stage includes three phase Vienna rectifier with power factor correction. The output of the first stage is the DC bus, which acts as an input to the second stage. Second stage includes full bridge LLC resonant converter and a lithium ion battery is connected to the output of second stage. To maintain constant DC bus voltage of 700V and to ensure unity power factor, dual loop control using d-axis and q-axis current control using space vector pulse width modulation (SVPWM) is adopted for controlling Vienna rectifier. Closed loop control for full bridge LLC resonant converter is designed using CC-CV control algorithm and pulse frequency modulation (PFM) to charge the rated 280 V/112 Ah lithium ion battery. MATLAB-Simulink is used for validating the designed system outcomes.

**Keywords**— Constant Current (CC)- Constant Voltage (CV) algorithm, Vienna rectifier, SVPWM control, Total Harmonic Distortion (THD), LLC converter, Zero Voltage Switching (ZVS), Gain plot, State of Charge (SOC), FFT analysis.

## I. INTRODUCTION

In today's growing market of EVs globally, unavailability of fast charging stations at certain distance is the main reason for the range anxiety among the EV users. Therefore, the main focus of various charging companies is towards the charging of electric vehicles at higher charging rates under the EV standards [1]. Fast charging requires efficient power converters which are capable of transferring higher power to achieve high C-rate. Most of researchers working on the new topologies to make grid more stable during charging, as the THD in the injected grid current without any power factor correction (PFC) control is more than 5% which cannot be considered based on Indian EV standards [2].

There are various level of charging: level 1, level 2 and level 3 charging [3]- [4]. Level 1 are mostly used in residential areas and have low ratings of up to 2 kW, which supports only slow charging and can take around 12 hours to charge to 100% SOC. Level 2 charging is adopted in residential as well as working premises, public places etc. which is capable of improving the C-rate of the battery and take around 6 hours to charge to 100% SOC. The ratings of level 2 charging goes up to 20 kW. Level 3 or DC fast charging are the most focus area nowadays due to their higher ratings ranging from 50 kW to several hundred kW. These charging station are provided among various areas of the cities and are expanding at a rising rate. Level 3 provides higher C-rates which helps the user to charge their vehicles up to 80% SOC in about 20 minutes and can take up to 1 hour depending on the rating of charging station. Level 1, level 2 chargers are characterized as on- board chargers and level 3 chargers as off-board chargers.

Various power factor corrections topologies are considered to make the grid current in phase with the voltage to minimize the THD below 5% based on the EV standards [5]. These topologies include interleaved boost PFC which is the simplified topology for power factor correction and features minimized inductor current ripples. In [6] the authors proposed a bridge less boost PFC topology for an EV charger which features no diode bridge rectifier and improved efficiency. A comparison was also made among PFC boost, interleaved boost, semi-bridge less boost and totem pole boost on the basis of the system complexity and efficiency.

For off-board EV charging, non- isolated dc-dc converters are not significant due to their limitations in power ratings and size. The full bridge LLC converter features higher power density, can be operated at higher switching frequency and also features ZVS of the primary side controlled switches based on the design consideration of the converter which minimizes the losses to a greater extent. The size of the converter can also be reduced when we are operating at higher switching frequency. But the designing of LLC converter is a tedious task due to various design consideration but its features attracts it for EV charging applications.

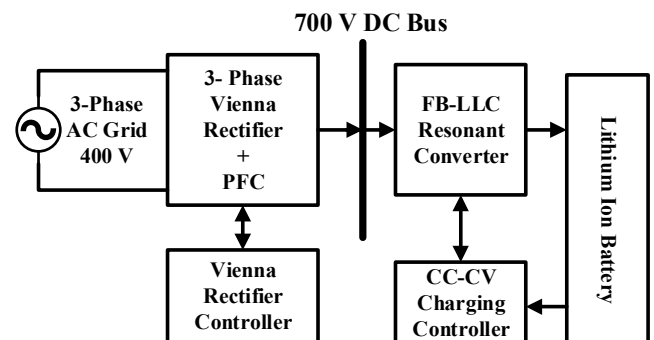


Fig. 1. Block diagram of designed off-board EV charger

There are various types of batteries which includes lead acid (Pb/PbO<sub>2</sub>) battery, lithium ion polymer (LiPo), nickel metal hydride (NiMH) batteries. Lithium ion batteries are preferred over other batteries in electric vehicles (EVs) due to their lighter weight and decent life cycle. To maximize the life of the lithium ion batteries CC-CV charging algorithm is preferred over CC or CV charging as it is more likely to mimic the chemistry involved in the battery. CC charging for fast charging can lead to temperature rise when the threshold voltage is reached. In CC-CV charging the CV mode is activated when the battery is charged up to a threshold voltage. In [7] the authors compared various charging algorithm including CC-CV, multi- stage CC, fuzzy logic etc. and conclusion are made on the basis of charging time and temperature effect on the battery.



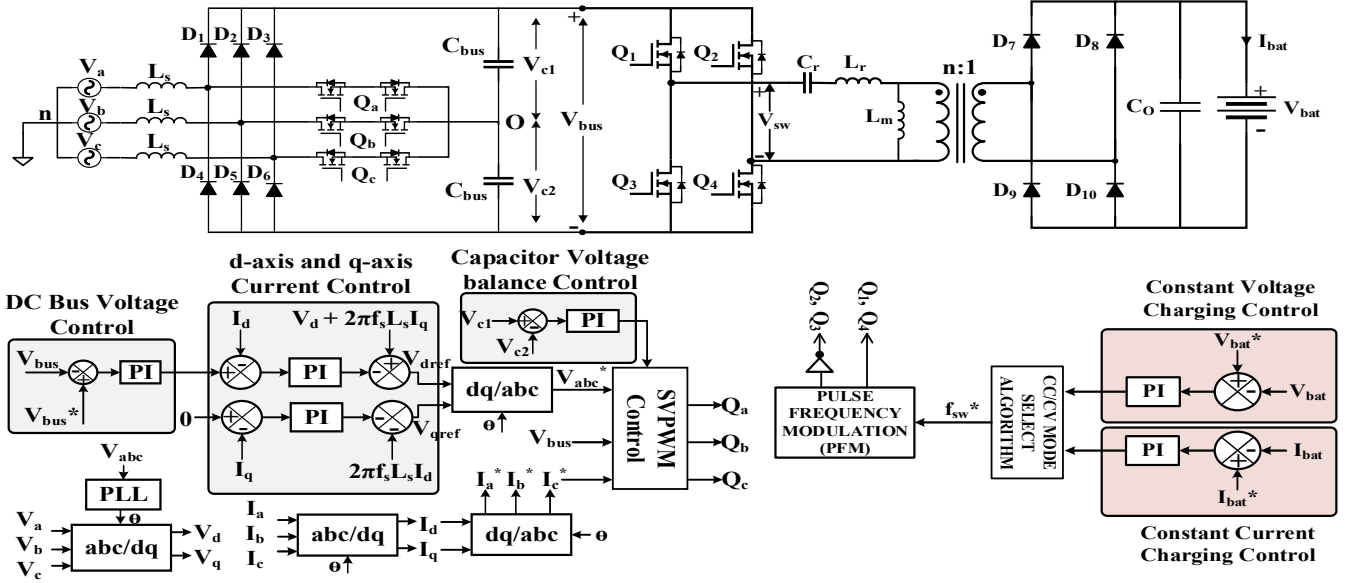


Fig. 2. Circuit diagram and control diagram of designed off-board EV charger

In this paper a 50 kW two stage off-Board EV charger is designed for charging a lithium ion battery using CC-CV algorithm and block diagram of designed system is shown in Fig. 1. First stage includes three phase Vienna rectifier with power factor correction. The output of the first stage is the DC bus, which acts as an input to the second stage. Second stage includes full bridge LLC resonant converter and a lithium ion battery is connected to the output of second stage. To achieve constant first stage bus voltage of 700V and to ensure unity power factor, dual loop d-axis and q-axis current control using SVPWM is adopted for controlling Vienna rectifier. FBLLC converter is designed and controlled using CC-CV control algorithm and pulse frequency modulation (PFM) to charge the rated lithium ion battery. MATLAB-Simulink is used for validating the designed system outcomes.

Organization of rest of the paper is defined as: section II describes the complete circuit of off-board EV charger, section III describes the modes of operation, section IV describes the design and control of AC/DC conversion stage and DC/DC conversion stage, results and performance of the system are evaluated in section V, and section VI contains the conclusion of this paper.

## II. CIRCUIT DESCRIPTION OF EV CHARGER

Vienna rectifier consists of boosting source inductor ' $L_s$ ' followed by a diode bridge rectifier (DBR) and then output bus capacitors ' $C_{bus}$ '. Vienna rectifier is a multilevel AC/DC converter due to three bidirectional switches  $Q_a$ ,  $Q_b$ ,  $Q_c$  connected between boosting source inductor and mid-point of bus capacitors. By using a control algorithm, mid-point of bus capacitors. By using a control algorithm, mid-point of bus capacitors. By using a control algorithm, mid-point of bus capacitors. The output of the three phase Vienna rectifier acts as an input supply to the LLC resonant converter. This voltage is provided to a switching circuit which is basically a full bridge inverter to generate a bipolar square wave waveform to excite the LLC tank circuit.  $L_r$ ,  $C_r$  resonates at resonant frequency  $f_o$  and behaves as series resonant circuit. After resonance, the LLC tank circuit provides a sinusoidal current which is resonating at resonant frequency. The magnitude of this current is changed using the turns ratio of

the high frequency transformer. This current is then transferred to the secondary side which includes the diode bridge rectifier for rectification and followed by the output filter capacitance which gives DC at the output. This DC is utilized to charge a lithium-ion battery. Fig. 2 shows the circuit diagram and control diagram for the designed off-board EV charger.

## III. MODES OF OPERATION OF EV CHARGER

### A. Vienna Rectifier Modes of Operation

Vienna rectifier consists of three bi-directional switches which makes eight switching states possible for operation. Based on grid current direction there are six sectors possible, which makes 48 possible states but some of them are redundant states which leads to 25 modes of operation. But in this paper all possible modes of operation for sector 1 is stated in TABLE I. and in Fig. 3 and Fig. 4.

TABLE I. SECTOR CLASSIFICATION BASED ON GRID CURRENT POLARITIES

<b>Sector 1</b>	$i_a, i_b, i_c = +, -, -$	<b>Sector 4</b>	$i_a, i_b, i_c = -, +, +$
<b>Sector 2</b>	$i_a, i_b, i_c = +, +, -$	<b>Sector 5</b>	$i_a, i_b, i_c = -, -, +$
<b>Sector 3</b>	$i_a, i_b, i_c = -, +, -$	<b>Sector 6</b>	$i_a, i_b, i_c = +, -, +$

TABLE II. VOLTAGE MAGNITUDE AND CAPACITOR MID-POINT CURRENT DURING EACH SWITCHING STATE

Switching State	$V_{ao}$	$V_{bo}$	$V_{co}$	$I_N$
000	$V_{bus}/2$	$-V_{bus}/2$	$-V_{bus}/2$	0
001	$V_{bus}/2$	$-V_{bus}/2$	0	$i_c$
010	$V_{bus}/2$	0	$-V_{bus}/2$	$i_b$
011	$V_{bus}/2$	0	0	$-i_a$
100	0	$-V_{bus}/2$	$-V_{bus}/2$	$i_a$
101	0	$-V_{bus}/2$	0	$-i_b$
110	0	0	$-V_{bus}/2$	$-i_c$
111	0	0	0	0

In TABLE II  $V_{ao}$ ,  $V_{bo}$  and  $V_{co}$  are the voltage magnitude between phases and capacitor mid-point and  $I_N$  is capacitor mid-point current.



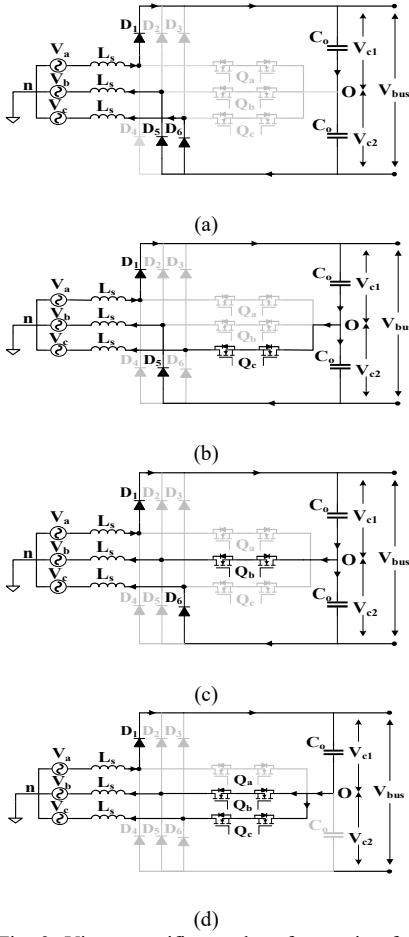


Fig. 3. Vienna rectifier modes of operation for switching states (a) 000 (b) 001 (c) 010 (d) 011

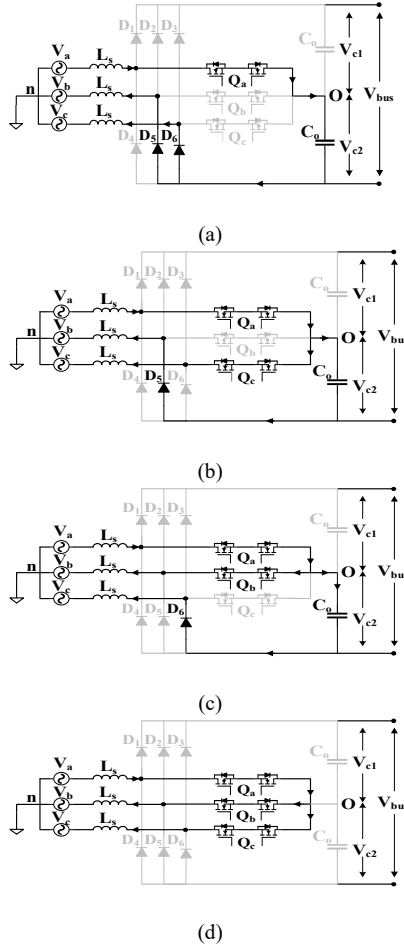


Fig. 4. Vienna rectifier modes of operation for switching states (a) 100 (b) 101 (c) 110 (d) 111

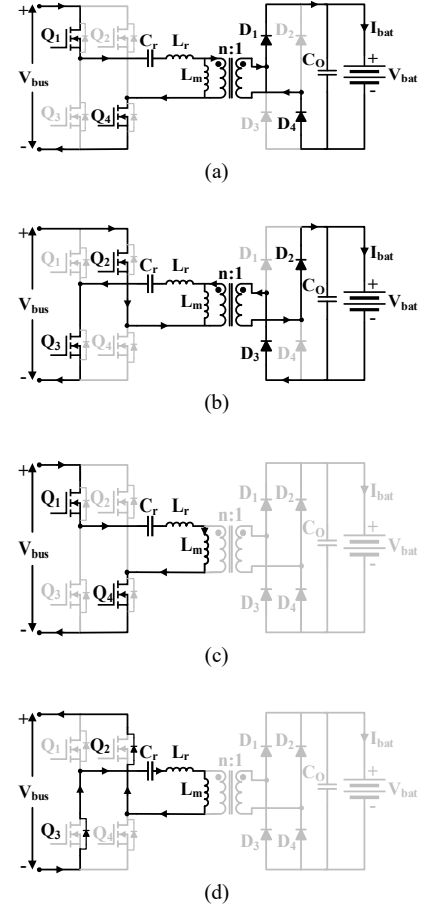


Fig. 5. FBLLC modes of operation for (a), (b) Power delivery (c), (d) No power delivery

### B. FBLLC Resonant Converter Modes of Operation

The various modes of operation of the LLC resonant converter involves positive cycle operation, negative cycle operation and freewheeling operation. In positive cycle operation, the gate pulse is provided to the controlled switches  $Q_1, Q_4$  which is responsible for positive half square wave at the input of resonating circuit as in Fig. 5(a). In negative cycle operation, the gate pulse is provided to the controlled switches  $Q_2, Q_3$  which is responsible for negative half square wave at the input of resonating circuit as in Fig. 5(b). These two modes are responsible to deliver power from primary side to secondary side. This is then utilized to charge the lithium ion battery. Fig. 5(c) describes the mode when the resonating current become equals to magnetizing current and no current flows to the secondary side of the converter and Fig. 5(d) shows the freewheeling mode for positive magnetizing current [9].

## IV. DESIGN AND CONTROL OF EV CHARGER

### A. Design of Vienna Rectifier

The Vienna rectifier designing parameters can be calculated using [10]. The boosting source inductance can be calculated using (1),

$$L_s = \frac{V_{bus/2}}{4 * f_{sw} * \Delta i_{ppmax}} \quad (1)$$

Where,  $V_{bus}$  is the DC bus voltage,  $f_{sw}$  is the operating switching frequency and  $\Delta i_{ppmax}$  is the maximum boosting inductor current ripple.

The bus capacitance can be calculated using (2),

$$C_{bus} = \left( \frac{1}{3} \right) \frac{P_{ac}}{4 * f_s * (V_{bus}^2 - (V_{bus} - \Delta V_{bus})^2)} \quad (2)$$

Where  $P_{ac}$  is the power rating,  $f_s$  is grid frequency,  $\Delta V_{bus}$  is the ripple in bus voltage and  $V_{bus}$  is the DC bus voltage. The design specifications for Vienna rectifier is tabulated in TABLE III.

TABLE III. DESIGN SPECIFICATIONS FOR VIENNA RECTIFIER

Parameters	Value
L-L rms grid voltage ( $V_{abc}$ )	400 V
Grid frequency ( $f_s$ )	50 Hz
Power rating ( $P_{ac}$ )	50 kW
DC bus voltage ( $V_{bus}$ )	700V
Bus capacitor ( $C_{bus}$ )	4300 $\mu$ F
Boosting source inductance ( $L_s$ )	41 $\mu$ H
Output current ( $I_o$ )	71.4285 A
Input current THD	< 3%
Switching frequency ( $f_{sw}$ )	200 kHz

$$V_o = M_g * \frac{V_{bus}}{n} \quad (9)$$

### B. Control of Vienna Rectifier

Vienna rectifier is controlled using d-axis and q-axis current control. In this algorithm, three loops are designed, output DC bus voltage control loop, d-axis and q-axis current control loop and capacitor voltage balance control loop as in Fig. 2. The pulses for controlled switches  $Q_a$ ,  $Q_b$  and  $Q_c$  are generated using SVPWM [11]. The control loops are designed using (3)-(7),

$$V_{qref} = \left\{ \begin{array}{l} - (K_{p_{iq}}(0 - I_q) + K_{i_{iq}} \int (0 - I_q) dt) \\ - 2\pi f_s L_s I_d \end{array} \right\} \quad (3)$$

$$I_{dref} = \left\{ \begin{array}{l} K_{p_{vbus}}(V_{bus}^* - V_{bus}) + \\ K_{i_{vbus}} \int (V_{bus}^* - V_{bus}) dt \end{array} \right\} \quad (4)$$

$$\Delta V_c = K_{p_{vc}}(V_{c1} - V_{c2}) + K_{i_{vc}} \int (V_{c1} - V_{c2}) dt \quad (5)$$

$$V_{dref} = \left\{ \begin{array}{l} -K_{p_{id}}(I_{dref} - I_d) - \\ K_{i_{id}} \int (I_{dref} - I_d) dt \\ + V_d + 2\pi f_s L_s I_q \end{array} \right\} \quad (6)$$

$$V_{bus} = V_{c1} + V_{c2} \quad (7)$$

Where  $V_{dref}$ ,  $V_{qref}$  and  $I_{dref}$ ,  $I_{qref}$  are the d-axis and q-axis reference voltages and currents respectively. In (3)-(6),  $K_{p_{iq}}$  and  $K_{i_{iq}}$  are PI controller gains for q-axis current loop,  $K_{p_{id}}$  and  $K_{i_{id}}$  are PI controller gains for d-axis current loop,  $K_{p_{vbus}}$  and  $K_{i_{vbus}}$  are PI controller gains for DC bus voltage loop,  $K_{p_{vc}}$  and  $K_{i_{vc}}$  are PI controller gains for bus capacitors voltage balance loop.  $V_{bus}^*$  is the reference DC bus voltage.  $V_{c1}$  and  $V_{c2}$  are DC bus capacitor voltages. Fig. 6 shows Vienna rectifier space vectors diagram.

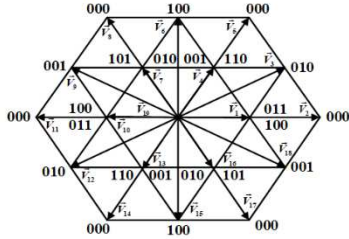


Fig. 6. Vienna rectifier space vectors diagram

### C. Design of Full Bridge LLC Resonant Converter

LLC resonant converter can be analyzed using First Harmonic Approximation (FHA) when not including the higher order harmonics using Fig. 7.

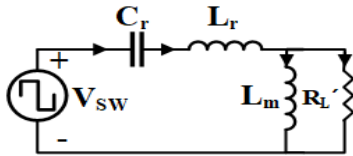


Fig. 7. LLC converter equivalent circuit

The designing steps followed for designing LLC converter can be summarized as in Fig. 8 using [12]. The gain is calculated as,

$$M_g = \left| \frac{jX_{Lm} || R_e}{(jX_{Lm} || R_e) + j(X_{Lr} - X_{Cr})} \right| \quad (8)$$

The output voltage depends on the gain magnitude,

The gain magnitude is the function of  $L_x$ ,  $f_x$  and  $Q$ . The optimum value of the  $Q$  and  $L_x$  is achieved using gain curve. So, the only controlled variable is the normalized frequency,  $f_x$ . The gain magnitude in terms of  $f_x$  can be written as,

$$M_g = \left| \frac{L_x * f_x^2}{[(L_x + 1) * f_x^2 - 1] + j[(f_x^2 - 1) * f_x * Q * L_x]} \right| \quad (10)$$

The quality factor is described as,

$$Q = \frac{1}{R_e} \sqrt{\frac{L_r}{C_r}} \quad (11)$$

The normalized frequency is described as,

$$f_x = \frac{f_{sw}}{f_o} \quad (12)$$

The turns ratio of transformer can be calculated as,

$$n = M_g * \frac{V_{bus}}{V_o} = \frac{V_{busnom}}{V_{onom}} = \frac{7}{3} \quad (13)$$

The maximum and minimum gain can be calculated as,

$$M_{gmin} = \frac{n * V_{o,min} + V_D}{V_{bus,max}} = 0.9 \quad (14)$$

$$M_{gmax} = \frac{n * V_{o,max} + V_D}{V_{bus,min}} = 1.11 \quad (15)$$

The  $L_x$  and  $Q$  can be selected from gain Vs normalized frequency as in Fig. 9 and  $f_{xmin}$  can be obtained as around 0.74 and minimum  $f_{sw}$  as around 150 kHz. In (14) and (15),  $V_D = 0.7$  V is the forward voltage drop across diodes. The capacitive and inductive region can also be identified using this gain curve to select the values of  $L_x$  and  $Q$  at which primary controlled switches can achieve ZVS as,  $Q = 0.6$ ,  $L_x = 3.3$ .

Equivalent load resistance can be calculated as,

$$R_e = \frac{8 * n^2}{\pi^2} * R_L = \frac{8 * n^2}{\pi^2} * \frac{V_o^2}{P_o} = 7.945 \Omega \quad (16)$$

Resonant circuit parameters can be calculated as,

$$C_r = \frac{1}{2 * \pi * f_{sw} * R_e * Q} = 167 \text{ nF} \quad (17)$$

$$L_r = \frac{1}{(2 * \pi * f_{sw})^2 * C_r} = 3.8 \mu\text{H} \quad (18)$$

$$L_x = \frac{L_m}{L_r} \Rightarrow L_m = L_x * L_r = 12.5 \mu\text{H} \quad (19)$$

The dead time to ensure ZVS can be calculated as,

$$\tau_{dead} \geq 16 * C_{eq} * f_{sw} * L_m \quad (20)$$

TABLE IV. DESIGN SPECIFICATIONS FOR LLC RESONANT CONVERTER

Parameters	Value
Bus voltage range ( $V_{busmin}$ - $V_{busmax}$ ), Nominal bus voltage ( $V_{busnom}$ )	675-725 V, 700 V
Output voltage range ( $V_{omin}$ - $V_{omax}$ ), Nominal output voltage ( $V_{onom}$ )	280-320 V, 300 V
Transformer turns ratio (n)	7:3
Quality factor (Q)	0.6
Resonating inductor ( $L_r$ )	3.8 $\mu\text{H}$
Resonating capacitor ( $C_r$ )	167 nF
Magnetizing inductance ( $L_m$ )	12.5 $\mu\text{H}$
Resonating frequency ( $f_o$ )	200 kHz
Inductance ratio ( $L_x$ )	3.3

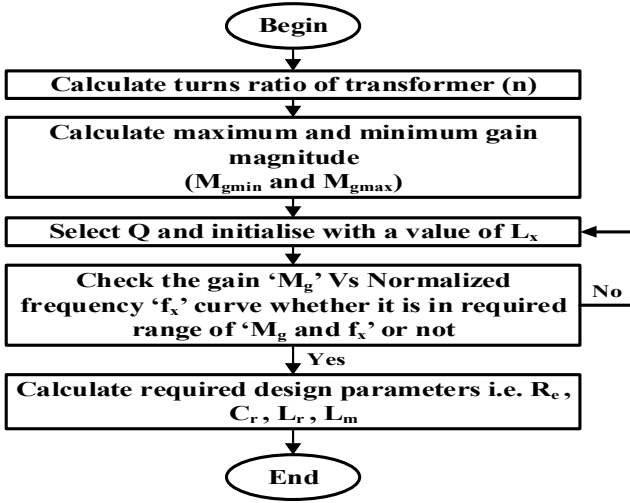


Fig. 8. Designing steps of LLC resonant circuit

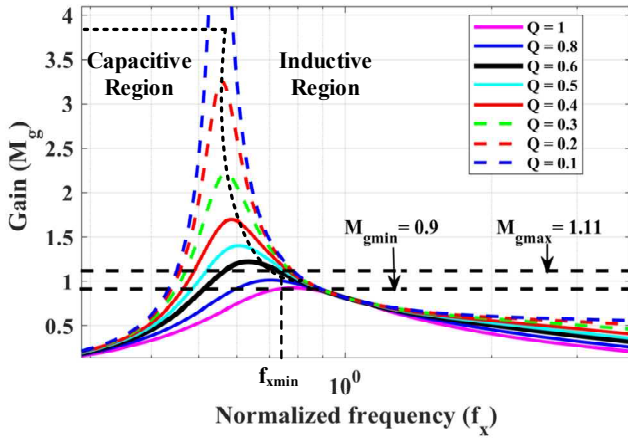


Fig. 9. Gain magnitude Vs normalized frequency curve

Design specifications of FBLLC converter is tabulated in TABLE IV.

#### D. FBLLC Resonant Converter Control

In constant voltage (CV) mode the sensed battery voltage  $V_{bat}$  is compared with the reference battery voltage  $V_{bat}^*$  and the produced error is provided to a PI controller to generate a reference frequency  $f_{sw}^*$ . In constant current (CC) mode the sensed battery current  $I_{bat}$  is compared with the reference battery current  $I_{bat}^*$  and the produced error is provided to a PI controller to generate a reference frequency  $f_{sw}^*$ . A CC-CV selector algorithm is designed to switch between CC mode and CV mode as in Fig. 11. The generated reference frequency  $f_{sw}^*$  is provided to the pulse frequency modulation (PFM) to generate pulses for switches  $Q_1, Q_4$  and complementary pulses for switches  $Q_2, Q_3$  as in Fig. 2. This variable frequency  $f_{sw}^*$  is responsible in controlling the output voltage as in (21)-(22),

$$f_{sw}^* = \left\{ \begin{array}{l} K_{p_{cv}}(V_{bat}^* - V_{bat}) + \\ K_{i_{cv}} \int (V_{bat}^* - V_{bat}) dt \end{array} \right\} \quad (21)$$

$$f_{sw}^* = \left\{ \begin{array}{l} K_{p_{cc}}(I_{bat}^* - I_{bat}) + \\ K_{i_{cc}} \int (I_{bat}^* - I_{bat}) dt \end{array} \right\} \quad (22)$$

Where  $K_{p_{cv}}$  and  $K_{i_{cv}}$  are the PI controller gains for CV charging and  $K_{p_{cc}}$  and  $K_{i_{cc}}$  are the PI controller gains for CC charging.

There is a lot of chemistry involved in a battery, various charging algorithms are adopted for charging different types of batteries. In EVs mostly lithium polymer batteries are in practice. For lithium-ion batteries constant current (CC) - constant voltage (CV) algorithm is preferred over other charging algorithms as in Fig. 10. Firstly, the battery is charged using constant current  $I_{bat}^*$ , till the battery voltage reaches a certain level,  $V_{bat}^{th}$  which is the threshold for transition from CC mode to CV mode. After  $V_{bat}^{th}$ , the battery current starts to decrease until the battery current reaches 10% of battery current which is denoted as  $I_{bat}^{lim} = 0.1I_{bat}^*$ , and battery voltage remains almost constant at  $V_{bat}^{th}$ .

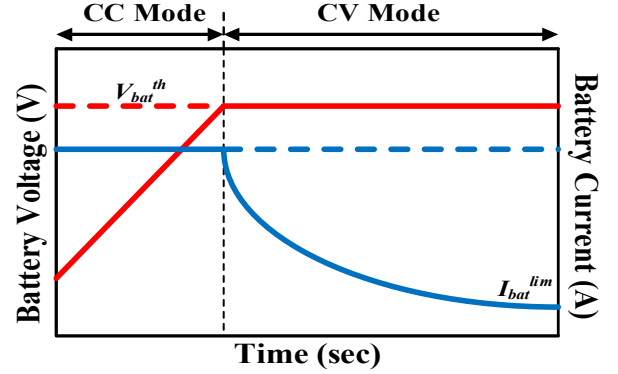


Fig. 10. CC-CV charging curve

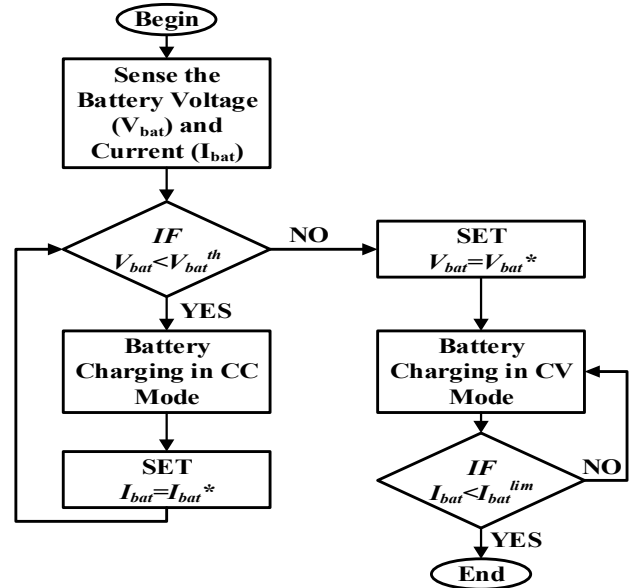


Fig. 11. Proposed CC-CV charging algorithm

## V. RESULTS AND PERFORMANCE EVALUATION

A 50 kW two stage off board EV charger is designed for charging a 280 V/ 112Ah lithium ion battery at high C-rate and the results and outcomes are validated using MATLAB-Simulink. Fig. 12(a) depicts the three phase line-to-line rms grid voltage. It can be clearly shown from Fig. 12(b) that three phase grid current is in phase with grid voltage. Fig. 12(c) shows that Vienna rectifier is able to maintain the 700 V bus voltage which acts as an input for LLC resonant converter. Fig. 12(d) shows that bus capacitor voltages are totally balance to provide constant 700 V bus voltage. From Fig. 12(e) the FFT analysis of phase 'a' grid current can be analyzed and THD comes out to be 2.35 %.

The LLC resonant converter is designed and outcomes are stated in Fig 13. Fig. 13(a) depicts the switching voltage  $V_{sw}$  having peak values as +700V and -700V which is the output of the full bridge inverter and this voltage is applied on the resonating circuit. Fig. 13(b) shows the output waveforms of the resonating inductor current,  $I_{lr}$  which resembles sinusoidal wave and magnetizing Inductor current,  $I_{lm}$ . Fig. 13(c) shows the resonating capacitor voltage,  $V_{cr}$ . The primary controlled switches are operated in ZVS and secondary diodes are operating in ZCS which can be proved using Fig. 13(d) and Fig. 13(e) respectively. From Fig. 13(d) it is clear that converter is operating in the inductive region and the current lags behind the voltage which leads to minimum losses and efficient operation.

Fig. 14(a) shows the battery voltage, which increases till the battery voltage attains threshold voltage  $V_{bat}^{th}=300\text{ V}$  at  $t=0.141\text{ s}$  and afterwards remains almost constant. Fig. 14(b) shows the battery current which remains constant at  $I_{bat}^*=150\text{ A}$  till the battery voltage reaches  $V_{bat}^{th}$  and after that starts decreasing to reach the limit current,  $I_{bat}^{lim}$ . But here we have demonstrated the charging of lithium ion battery for a limited time duration for the verification of the charging algorithm. Fig. 14(c) depicts the state of charge (SOC) of the battery. Here, at 80 % SOC the battery voltage is determined as 298.5 V and at  $t=0.2\text{ s}$  battery current  $I_{bat}$  is around 100 A based on control loops which will decrease further till  $I_{bat}^{lim}$ , which is used for validating the CC-CV control.

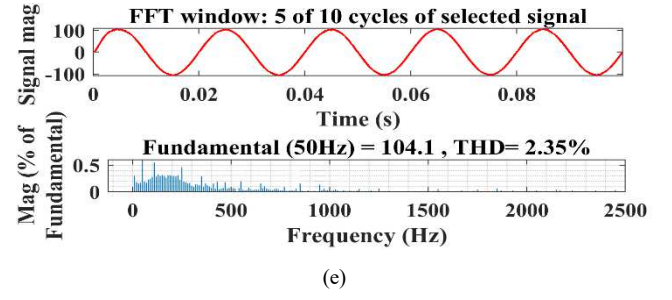
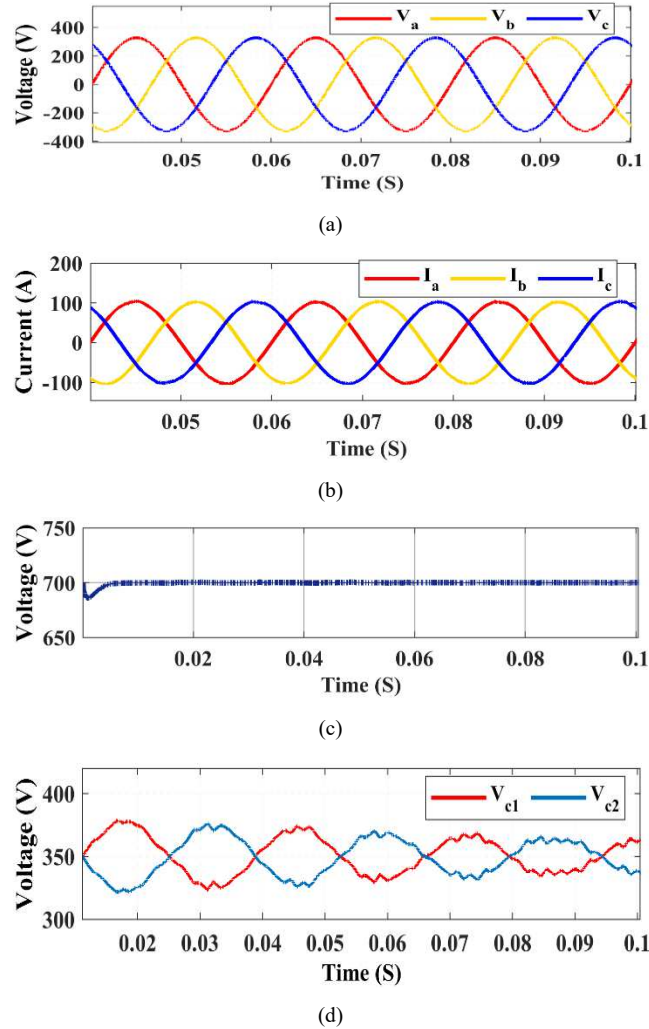


Fig. 12. Simulation results of Vienna rectifier (a) 3-phase grid voltage (b) 3-phase grid current (c) DC bus voltage (d) Bus capacitors voltage balance (e) FFT analysis of grid current

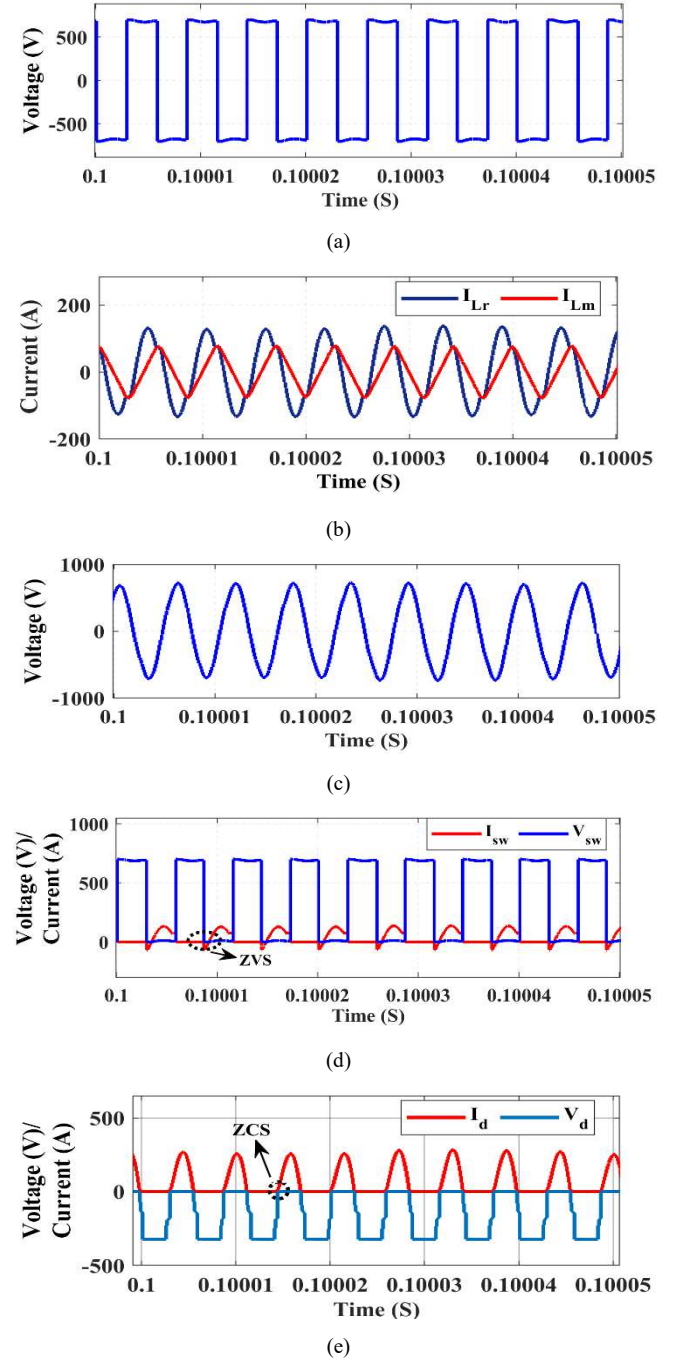


Fig. 13. Simulation results of FBLLC converter (a) Switching voltage (b) Resonating and magnetizing inductor current (c) Voltage across resonating capacitor (d) Voltage and current across primary side switch (e) Voltage and current across secondary side diode



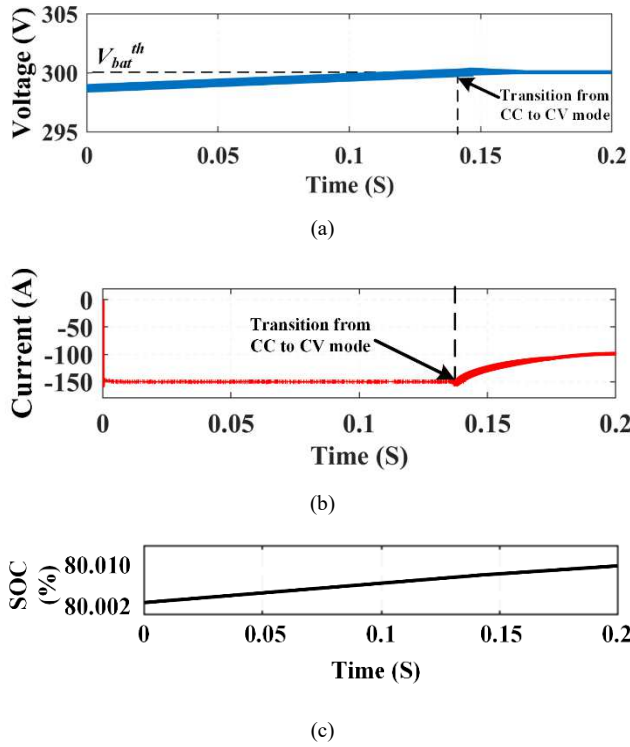


Fig. 14. Simulation results of (a) Battery voltage (b) Battery current (c) Battery state of charge (SOC)

## VI. CONCLUSION

In this paper, a complete designing of a 50 kW two stage off-board EV charger for charging a 280 V/ 112 Ah lithium ion battery is discussed. All the design considerations were taken into account to ensure ZVS and provide efficient charging. In first stage three phase Vienna rectifier was designed and it was able to minimize the THD to 2.35%, providing almost u.p.f. and capable of maintaining 700 V bus voltage. In second stage full bridge LLC converter was designed to ensure ZVS in the primary side controlled switches and ZCS in secondary side diodes. The DC voltage at the output of the LLC converter is utilized for charging the battery using designed CC-CV algorithm. The designed off-board charger is capable of charging a battery while mitigating its effects on the grid side. All the results and discussion are validated using MATLAB-Simulink and presented in this paper.

## ACKNOWLEDGMENT

The authors would thank the Centre of Excellence for Electric Vehicles and Related Technologies, Delhi Technological University for providing necessary facilities for performing our research work.

## REFERENCES

- [1] A. Ahmad, Z. Qin, T. Wijekoon and P. Bauer, "An Overview on Medium Voltage Grid Integration of Ultra-Fast Charging Stations: Current Status and Future Trends," *IEEE Open Journal of the Industrial Electronics Society*, vol. 3, pp. 420-447, 2022.
- [2] L. Wang, Z. Qin, T. Slangen, P. Bauer and T. v. Wijk, "Grid Impact of Electric Vehicle Fast Charging Stations: Trends, Standards, Issues and Mitigation Measures - An Overview," *IEEE Open Journal of Power Electronics*, vol. 2, pp. 56-74, 2021.
- [3] G. Rituraj, G. R. C. Mouli and P. Bauer, "A Comprehensive Review on Off-Grid and Hybrid Charging Systems for Electric Vehicles," *IEEE Open Journal of the Industrial Electronics Society*, vol. 3, pp. 203-222, 2022.
- [4] S. Pareek, A. Sujil, S. Ratra and R. Kumar, "Electric Vehicle Charging Station Challenges and Opportunities: A Future Perspective," in *2020 International Conference on Emerging Trends in Communication, Control and Computing (ICONC3)*, Lakshmangarh, India, 2020.
- [5] S. S. Sayed and A. M. Massoud, "Review on State-of-the-Art Unidirectional Non-Isolated Power Factor Correction Converters for Short-/Long-Distance Electric Vehicles," *IEEE Access*, vol. 10, pp. 11308-11340, 2022.
- [6] R. Pandey and B. Singh, "A Power Factor Corrected Resonant EV Charger Using Reduced Sensor Based Bridgeless Boost PFC Converter," *IEEE Transactions on Industry Applications*, vol. 57, no. 6, pp. 6465-6474, Nov.-Dec. 2021.
- [7] T. T. Vo, W. Shen and A. Kapoor, "Experimental comparison of charging algorithms for a lithium-ion battery," in *2012 10th International Power & Energy Conference (IPEC)*, Ho Chi Minh City, Vietnam, 2012.
- [8] J. W. Kolar, U. Drofenik and F. C. Zach, "VIENNA rectifier II-a novel single-stage high-frequency isolated three-phase PWM rectifier system," *IEEE Transactions on Industrial Electronics*, vol. 46, no. 4, pp. 674-691, Aug. 1999.
- [9] Abdel-Rahman, Sam, "Resonant LLC Converter: Operation and Design," Infineon Technologies North America (IFNA) Corp. , September 2012.
- [10] Texas Instruments, "Vienna Rectifier-Based, Three-Phase Power Factor Correction (PFC) Reference Design Using C2000™ MCU," 2016-2017.
- [11] A. Sunbul and V. K. Sood, "Simplified SVPWM Method for the Vienna Rectifier," in *2019 20th Workshop on Control and Modeling for Power Electronics (COMPEL)*, Toronto, ON, Canada, 2019.
- [12] Huang, Hong,, "Designing an LLC Resonant Half-Bridge Power," Texas Instruments (TI) Power Supply Design Seminar SEM1900, 2010.



# Design of a novel robust recurrent neural network for the identification of complex nonlinear dynamical systems

R. Shobana<sup>1</sup> · Bhavnesht Jain<sup>1</sup> · Rajesh Kumar<sup>2</sup>

Accepted: 29 August 2023

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

## Abstract

A novel fully connected recurrent neural network (FCRNN) structure is proposed for the identification of unknown dynamics of nonlinear systems. The proposed recurrent structure consists of internal feedback layers of adjustable weights which impart necessary memory property to the structure and improves its ability in handling the dynamical systems. The back-propagation algorithm (BP) is used to derive the weight update equations of the proposed model. The convergence of the proposed approach is proven in the sense of Lyapunov-stability analysis. A total of three examples are considered and the performance of the proposed structure is evaluated by comparing it with the results obtained from other popular neural network models such as feed-forward neural network (FFNN), Elman neural network (ENN), Jordan neural network (JNN), and the locally recurrent neural networks (LRNN). Experimental results obtained show that the FCRNN model has outperformed the other neural models in terms of identification accuracy and robustness.

**Keywords** Recurrent neural networks · Feed-forward neural network · Identification · Disturbance rejection

## 1 Introduction

An increase in the complexity of various industry-based processes has led to the use of intelligent controllers to control and stabilize various parameters that vary with time. Identification of dynamic models is fundamentally important to design a better controller or for a better understanding of the process (Sastry et al. 1994). This has led to various works on the identification of the best models for a dynamic process in the literature (Quaranta et al. 2020). Dynamic models are one whose output behavior depends over time. The problem of identification is generally solved using two

approaches. Modeling of any system can be done using the direct method or using the system identification approach. The direct approach deals with studying the interactions for a while by application of physical laws. The physical laws are usually expressed as differential equations. But, this method fails in the identification of the best model due to the absence of information about the system or the system being incomplete and unidentifiable. This has led to use of the system identification method for modeling and analyzing the behavior of the system (Ljung 2010; Moeller 2004). Among the system identification approaches, soft computing techniques form a practical approach for solving such complex problems. Based on the level of prior knowledge or input–output data at hand, one can classify identification models into two types (Haykin 2009):

1. **Grey box models:** Models are derived using first principles but still have a lot of unknown parameters.
2. **Black-box models:** Models are determined based on experimental data. It is designed with little or no prior insight into the system. It has a lot of unknown parameters.

Though the black-box identification methods such as block-structured, Volterra, nonlinear auto-regressive network with

✉ Rajesh Kumar  
rajeshmahindru23@gmail.com;  
rajeshmahindru23@nitkkr.ac.in

R. Shobana  
r.shobana@galgotiacollege.edu;  
shobanaramasubramanian@gmail.com

Bhavnesht Jain  
bhavneshtjain@dtu.ac.in; bhavneshtmk@gmail.com

<sup>1</sup> Department of Electrical Engineering, Delhi Technological University, Shahbad Daultpur, Main Bawana Road, Delhi 110042, India

<sup>2</sup> Department of Electrical Engineering, National Institute of Technology Kurukshetra, Kurukshetra 136119, India



exogenous inputs (NARX) model, and radial basis function networks are useful for nonlinear identification, they suffer from some limitations (Calin 2020). Among these methods, Artificial Neural network (ANN) and fuzzy logic methods have emerged as useful tools for identifying nonlinear systems (Haykin 2009). (ANN) possesses many advantages over fuzzy logic systems such as fault tolerance, robustness, and learning adaptivity to uncertainties and noisy data (Basheer and Hajmeer 2000). They are non-parametric methods that infer the characteristics of biological neurons. They consist of interconnected neurons between the input and output layers termed as hidden units with respective weights between them. They quickly adapt to process input–output behavior through learning. Hence, (ANN) is widely being used to implement nonlinear identification models and controllers that can self-adapt their parameters through training. The (ANN) is generally of two types (Elsheikh et al. 2019):

1. Multi-Layer Perceptrons (MLP)
2. Recurrent Neural Network (RNN)

MLPs are feed-forward neural networks with the output errors back-propagated using a standard BP algorithm. They cannot exhibit dynamic mapping unless they know the system's order or the use of tapped delay lines. By nature, they cannot retain their past information (Chen and Billings 1992). The availability of feed-forward and feedback connections of previous outputs to the hidden layer creates a memory in the recurrent neural networks (Sastry et al. 1994). The connections between the hidden layers can be through local or global feedback. In this work, a local feedback connection is considered within the hidden layer. Further, RNN can be of two types namely Partial Recurrent Neural Networks (PRNN) or a Fully Recurrent Neural Networks (FRNN). Hopfield is a form of a fully recurrent structure where every output is connected to the input. The network is designed symmetric and has no target to achieve as in supervised training, hence they suffer from memory limitation and inefficiency to learn new patterns. On the other side, Elman and Jordan's networks are also a form of fully RNN structures. A context layer is newly introduced in their structure to store the average of past outputs. In the Elman model, the outputs of the hidden layer are fed back as inputs to the context layer. Elman nets are very efficient due to the addition of an extra input layer but are not suitable for online identification (Ku and Lee 1995). In Jordan networks, delayed outputs of the network are fed back as inputs to the hidden layer. The size of the context layer depends on the size of the output layer. When more outputs are involved, the Jordan structure becomes very large and gives slow convergence. Hence, the original Elman and Jordan structures are being modified to improve the network performance. The addition of additional inputs between context and output nodes improves the dynamics and conver-

gence properties of modified Elman–Jordan structure over FFNN and the original Elman and Jordan structure (Thammano and Ruxpakawong 2010; Gao et al. 1996; Şen et al. 2020). Hybrid Elman–Jordan structures are widely used in literature for many applications as they result in an efficient and robust model (Pham and Karaboga 1999). The Local Recurrent Neural Network (LRNN), which belongs to the PRNN model category, shares the same fundamental structure as the FFNN model and also includes self-feedback which plays an important role in retaining the past information. These structures have dynamic neurons instead of static neurons like FFNN (Kumar et al. 2019). In this work, fully recurrent neural network (FCRNN) is proposed for identification of complex nonlinear dynamic systems. The efficiency and robustness of the proposed model are proved by comparing it with four different neural network structures namely FFNN, LRNN, ENN, and JNN. The gradient descent BP algorithm is used to update the weight equations. The convergence of the learning algorithm is also proved in the sense of Lyapunov-stability analysis.

## 1.1 Related works

Research works on the identification-based control and modeling of nonlinear dynamic systems are increasingly been carried out in past decades. Neural networks are being increasing used in the literature for identification and control of nonlinear dynamic systems that cannot be done using the conventional linear structures (Aggarwal 2018; Willis et al. 1992). The effectiveness of neural networks for identification of nonlinear dynamic system is cited in various related works (Noël and Kerschen 2017; Yu et al. 2019; Kroll and Schulte 2014). In Kumar et al. (2019), a Dynamic Recurrent Neural Network (DRNN) structure is proposed. The structure resembles NARX and Multi-Layer FFNN (MLFFNN) with recurrent self-weighted hidden neurons. The results show that DRNN performs better in terms of robustness and parameter variations due to the presence of memory in them. FFNN models such as MLP, radial basis functional neural networks (RBFN), and functional link networks are some of the universal architectures that are capable of identifying complex nonlinear systems as suggested in the literature. Feed-forward models suffer from the main disadvantage of being memoryless structures. Even with the usage of tapped delay lines, these structures suffer from slow learning and get trapped easily in local minima. Dynamic systems being dependent on past inputs and outputs, require structures with memory for learning the long-term dynamics of the system (Savran 2007). As given in Ge et al. (2009), RNN by nature is a dynamic structure with internal memory. The RNN structures are found to give accurate one step ahead prediction of complex nonlinear systems. In Sanchez (1994), the authors have proved that ANN can identify the dynamics of nonlin-

ear systems when they have a fading memory. They have proposed a form of PRNN structure called modified Hopfield with dynamic neurons for restoring past information and one step ahead of prediction. In Coban (2013), another novel form of PRNN, Context Layered LRNN (CLLRNN) is proposed. An extra context layer is included in the existing RNN structure and is trained with an adaptive learning rate. It has been found to retain past information and improve the performance of the structure better than other RNN networks in the literature. This structure is found to have rectified the drawback of FFNN making the network more dynamic and stable to identify complex dynamics. In Yazdizadeh and Khorasani (2002), four structures based on adaptive time delay neural networks is proposed for the identification of different classes of nonlinear systems. The proposed structures namely Time Delay Neural Network (TDNN) and Adaptive Time Delay Neural Network (ATDNN) are found to use lesser adjustable parameters and less prior information about system over FFNN models. Four ANN structures namely Elman NN, modified Elman NN, time-delayed ANN, and internal time-delayed RNN are proposed in Li et al. (2008). All four selected networks are trained using Genetic Algorithm (GA). The advantages and disadvantages of selected dynamic structures are discussed briefly. The results show that the RNN structures such as Elman NN, modified Elman NN and the internal time-delayed RNN have better identification precision than static time-delayed ANN. In Abdollahi et al. (2003) and Yazdizadeh and Khorasani (1997), different types of identification approaches commonly used in literature are briefly reviewed. A series-parallel identification approach is one where the past output of the plants are fed back as input to the network and parallel-based identification is one where the states of the identifier are fed back as input to the system. The series-parallel identification is found to perform with better convergence and stability and they do not require any order to be known before implementation as compared to parallel-based identification. In Pham and Karaboga (1999), the authors have evaluated the superiority of the GA over the BP algorithm to train the modified Elman and Jordan network against the standard Elman and Jordan structure by the addition of self-feedback connections for the context units with weights fixed between 0 and 1. Though GA does not get trapped in local minima like the BP algorithm yet updates weights on the entire population of a network. BP generally does for one layer at a time instant. In Deng (2013), a novel Dynamic Neural Network (DNN) is proposed. To avoid the mapping capability of DNN, a modified structure with series-parallel identification is carried out. It is found to provide good mapping capabilities for training and robustness for parameter variations of complex nonlinear systems. In Li (2001), an Extended Kalman Filter (EKF) is used to update the weight equations. Though EKF generates faster convergence than BP

gets trapped in instability caused by initial conditions during linearization. In Ogunmolu et al. (2016); Wang and A new concept using lstm neural networks for dynamic system identification, in, (2017), the authors have proposed linear and nonlinear Hammerstein models for overall system identification. Linear MLP-Hammerstein could determine uncertainties and disturbances as they occur over time but yet could not capture sequential data occurrence. Nonlinear Hammerstein models like Vanilla Long Short Term Memory (LSTM) are proposed by connecting three LSTM models with nonlinear activation functions in the first two layers and fully connected RNN at the last layer. Though the network found better models yet it suffers from a long training time. They were found to have faster convergence but are a little sensitive to delay. In Thammano and Ruxpakawong (2009), a multivalued connection weight depending on inputs involved for a better performance of modified recurrent structure over others is proposed. In Luttmann and Mercorelli (2021), various parametric and non-parametric methods of identification are compared. Global search methods such as Particle swarm optimization (PSO) and Differential Evolution (DE) are found to be more robust in parametric methods than GA. ANN is found to be widely used in many applications of nonlinear parametric methods. The Recurrent neural network structures are widely used for different applications for modeling and identification of nonlinear dynamical systems such as fuel cells, DC motors, chemical processes, tank systems, and fault detection (Bhat and McAvoy 1990). In Schubert et al. (1997), the criteria for the selection of the learning rate, initialization, and synchronization of the network is well surveyed. For successful convergence of RNN, the learning rate between actual and cost function should be selected as a minimum. The system has to be brought to linearized state before initialization to avoid the effect of nonlinear activation functions like tangent hyperbolic functions. These would help in modeling of system with a small number of neurons and training cycles. In Veerasamy et al. (2022), a PID-based controller for automatic load frequency control of the power system is designed. The combination of PSO and Gravitational Search Algorithm (GSA)-based recurrent Hopfield NN is used for the identification of the model and tuning of the parameters of the controller. The weight update equations are derived and checked for stability using the Lyapunov-based stability analysis. From the above survey, RNN is found to be more efficient in the identification of the best models for dynamic systems. This has motivated us to design a novel RNN structure that would improve the model's ability to extract the complex dynamic of nonlinear systems in a better way with very lesser inputs.

The main objectives of the paper are:

1. To design a novel recurrent structure that requires fewer inputs and trainable parameters for identifying complex dynamical systems.
2. To prove the convergence of the weight update equations and stability of the proposed structure.
3. To simulate and compare the performance of FCRNN with other selected RNN structures.
4. To simulate and compare the robustness of FCRNN with other selected RNN structures.

**The rest of the paper is organized as follows:** the introduction section discusses the advantage of RNN over other ANN networks and the primary objective of the work carried out. The related surveys conducted in the field of ANN by various researchers are also elaborated. Section 2 gives the problem statement and the main contribution of this paper. Section 3 briefs on the novel recurrent structure, and the derivation of the weight update rule (learning algorithm). The proof for convergence and stability for fixed learning rate is also discussed theoretically in this section. Section 4 will highlight the results of the simulation carried out to extract the dynamics of the system. Three examples are considered for the same. The results of FCRNN are compared with other considered structures of ANN and are tabulated. Section 5 focuses on the simulation results and the conclusion is formed based on the simulation results derived. This section proves the efficiency of the proposed FCRNN structure.

## 2 Problem statement

Let us consider a nonlinear plant with desired outputs such as  $y_p(k-1), y_p(k-2), \dots, y_p(k-m)$  and desired inputs such as  $r(k), r(k-1), r(k-2), \dots, r(k-n)$ .  $f$  is the nonlinear mapping function between them. The identification structure of the nonlinear plant can be mathematically given as follows:

$$y_p(k) = f[y_p(k-1), \dots, y_p(k-n), r(k-1), \dots, r(k-m)] \quad (1)$$

Here,  $f(\cdot)$  can be a neural network, wavelet, or sigmoid function. In this study, a neural network is considered for nonlinear approximation.  $n$  and  $m$  are the orders of the plant. If FCRNN is selected as an identifier, with  $\hat{f}$  being the unknown nonlinear and differentiable function, the mathematical differential equations are given by

$$y_{fcr}(k) = \hat{f}[y_p(k-1), r(k)] \quad (2)$$

where  $y_{fcr}(k)$  denotes the output of FCRNN. The structure uses the series–parallel configuration.

**From Universal Approximation Theorem proof** Calin (2020) Let  $L$  be the number of layers in the network, with

input  $r \forall R^n$ ,  $w_i$  and  $b_i$  are the weight and bias matrix, respectively,  $\sigma \forall R$  be the activation function and  $C(I_n)$  be the linear space of a continuous function with  $I_n = [0, 1]^n$ ,  $n$ -dimensional unit space. A function  $f : R \rightarrow [0, 1]^n$  is called sigmoidal if

$$\lim_{k \rightarrow \infty} y_p(k) = 0 \quad (3)$$

For a continuous arbitrary function  $\sigma$ , the finite sums of form are given by

$$G(x) = \sum_{i=1}^N \alpha_j \sigma(w_i r + b_i) \quad (4)$$

In other words, for function  $\forall f \in C(I_n)$  and  $\forall \epsilon > 0$ , there is a sum of  $G(x)$  of the above form such that

$$|G(x) - f(x)| < \epsilon, \forall r \in I_n \quad (5)$$

From the above proof, it can be concluded that one hidden layer can learn any continuous function  $\forall f \in C(I_n)$  with an  $\epsilon$  i.e error, by tuning the weights. The approximation capability of the network depends on hidden neurons and the layers in the neural structure. The main goal is to approximate the nonlinear function  $\hat{f} \simeq f$  using deep neural networks such as the proposed FCRNN to keep the number of hidden neurons at a minimum and achieve an accurate model as training progresses. In other words,

$$\lim_{k \rightarrow \infty} |y_p(k) - y_{fcr}(k)| \leq \epsilon \quad (6)$$

where  $\epsilon \rightarrow 0$ . To reach the requirement as given in Eq. (6), the trainable network weights are continuously updated using the standard BP algorithm.

## 3 Mathematical formulation of FCRNN structure

### 3.1 FCRNN model

The FCRNN is implemented using a series–parallel identification structure. The identification structure contains the blocks of the nonlinear plant and the FCRNN model. The identification being series–parallel mode, the past and present outputs  $y_p(k-1)$  and  $y_p(k)$ , respectively, becomes one of the inputs to the structure. The FCRNN identifier takes only two inputs. These include the present external input signal  $r(k)$  and one delayed output of the plant  $y_p(k-1)$ . This simplifies the structure and results in lesser parameter usage by the

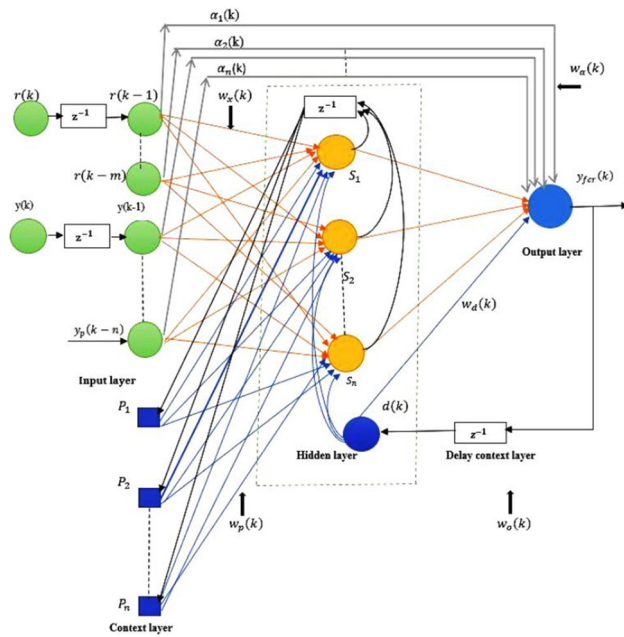


Fig. 1 Proposed FCRNN structure

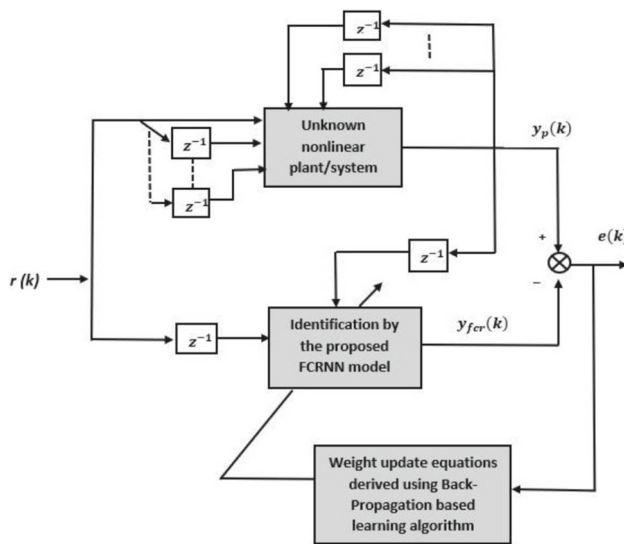


Fig. 2 Series-parallel identification model

structure. The identification error between the plant model and the network is used to update the weight equations and train the network. The identification structure of FCRNN is shown in Fig. 2. The FCRNN identifier takes the form of

$$y_{fer}(k) = \hat{f}[y_p(k-1), r(k)] \quad (7)$$

where  $\hat{f}$  is the nonlinear approximation function. Figure 1 shows the proposed FCRNN structure. The function of each layer used in Figure 1 is as follows:

- Input layer:** This layer 1 consists of the input signals. Here, the present and previous external inputs and previous outputs of the plant are considered as input signals to the structure, i.e.,  $y_{fer}(k-1)$  and  $r(k-1)$  are considered as inputs for the FCRNN structure. As the structure is recurrent with self-loops, they can retain their past values of inputs and outputs. The structure uses a minimum number of inputs for successful training of the network and the input signals are passed to the hidden neurons through weighted links  $w_x(k)$  (set of orange lines shown in Fig. 1).
- Hidden layer:** Layer 2 of the network receives three weighted connections of signals. One from the input layer and others from the context layer and the delayed context layer. A tangent hyperbolic activation function is used as a nonlinear function for the hidden layer neurons. The blue and black lines indicate the local feedback of signals received to the hidden layer from other layers in Fig. 1.
- Context layer:** Layer 3 is the additional input layer between input and the 1<sup>st</sup> hidden layer. The size of the context layer is the same as the size of the hidden layer. Each node stores the induced field values of its respective hidden neurons. They are self-feedback loops and transmit back the signals to the hidden layer through weights,  $w_p(k)$ .
- Output layer:** Layer 4 of the network structure receives the weighted connections from the hidden layer through weights,  $w_o(k)$  and  $w_d(k)$  plus an additional weighted link established between the input and the output layer (Grey lines in Fig. 1). These adjustable parameters are represented by the weight vector  $w_\alpha(k) = [\alpha_1(k), \alpha_2(k), \dots, \alpha_n(k)]$ . Each node of the output layer is executed with a linear or nonlinear activation function. In this case, the purelin is used as an activation function to extract the output.
- Delayed Context layer:** The layer refers to the layer between the output of the network and the hidden layer. This layer has a local feedback. The number of the delayed context layer is same as the size of the output layer. The previous output of the plant is fed back as one of the input to the hidden layer. The signals are passed through trainable weights.
- An additional weighted link:** An additional trainable weighted links between the input layer and the output layer are added in this work. This additional link makes the structure a fully recurrent neural network. Signals are passed to the output layer through weights,  $w_\alpha(k)$ . This addition makes the structure more robust as the inputs are directly connected to the output layer. With any further increase in the input range, the output quickly adapts to the new changes.



Mathematically, the relation between the input and output layers is

$$y_{fcr}(k) = \hat{f}\left(\sum_{i=1}^m X(k-i)w_x(k) + w_\alpha(k)X(k-i)\right) \quad (8)$$

where  $w_x(k)$  is the weight of input neurons, and  $w_\alpha(k)$  is the new dynamic trainable parameter added between input and output.  $X(k-i)$  is the matrix representing one delayed input of the plant. The hidden layer output at  $j^{th}$  instant is computed as

$$S_j(k) = g_1\left(\sum_{i=1}^m X(k-i)w_x(k) + b_x(k)w_b(k) + P_i(k)w_p(k)\right) \quad (9)$$

where  $b_x(k)$  is input bias neuron,  $w_b(k)$ ,  $w_p(k)$  are the weight associated with bias and context layer.  $g_1$  denotes the tangent hyperbolic activation function.  $P_i(k)$  denotes the context layer matrix. The output of the network is given by

$$y_{fcr}(k) = g_2\left(\sum_{i=1}^m S_j(k)w_o(k) + b_o(k)w_o(k) + d(k)w_d(k) + w_\alpha(k)X(k-i)\right) \quad (10)$$

where  $w_o(k)$  and  $w_d(k)$  are the weights associated with output bias and delayed output layer, respectively, and  $b_o(k)$  is the output bias neuron and  $g_2$  denotes the purelin activation function.  $d(k)$  denotes the delayed output layer matrix.

### 3.2 Learning algorithm

To update the tunable parameters of FCRNN, a gradient descent-based back-propagation algorithm is used. The tunable weight vectors of the proposed model include:  $[w_x(k), w_o(k), w_p(k), w_d(k), w_\alpha(k)]$  and each element in these weight vectors is modified during each epoch using the update equations that are derived using the BP method. To attain this, a cost function is defined in the first instance. Here, Mean Square Error (MSE) is chosen as the cost function to evaluate the efficiency of the training process. MSE is defined as the average squared difference of the output obtained from actual  $y_{fcr}(k)$  to the desired output of the plant  $y_p(k)$ . It is mathematically expressed as

$$E(k) = \frac{1}{2}[y_p(k) - y_{fcr}(k)]^2 \quad (11)$$

and

$$e(k) = y_p(k) - y_{fcr}(k) \quad (12)$$

where  $e(k)$  denotes the instantaneous error. To update the weights of the output layer, the errors are back-propagated from the output to the hidden layer using the chain rule as follows:

$$\frac{\partial E(k)}{\partial w_o(k)} = \frac{\partial E(k)}{\partial y_p(k)} \times \frac{\partial y_p(k)}{\partial V(k)} \times \frac{\partial V(k)}{\partial w_o(k)} \quad (13)$$

On simplification,

$$\frac{\partial E(k)}{\partial w_o(k)} = -e(k) \times S_j(k) \quad (14)$$

where  $S_j(k)$  indicates the induced field and derived using as given in Eq. (4). A linear activation function such as purelin is considered in the output layer. The output weights  $w_o(k)$  are updated using the formula as follows:

$$w_o(k+1) = w_o(k) + \eta e(k)S_j(k) \quad (15)$$

where  $\eta$  is the learning rate and its range is considered between 0 and 1. To calculate the weights associated with the hidden layer, the errors are further back-propagated from the hidden layer to the input layer. Using chain rule, they are calculated as follows:

$$\frac{\partial E(k)}{\partial w_x(k)} = \frac{\partial E(k)}{\partial y_p(k)} \times \frac{\partial y_p(k)}{\partial Z(k)} \times \frac{\partial Z(k)}{\partial S_j(k)} \times \frac{\partial S_j(k)}{\partial V(k)} \times \frac{\partial V(k)}{\partial w_x(k)} \quad (16)$$

where  $Z(k)$  and  $V(k)$  denote the induced field of the hidden layer and the induced field of the output layer, respectively. On simplification,

$$\frac{\partial E(k)}{\partial w_x(k)} = -e(k)w_o(k)(I - S_j^2)X(k) \quad (17)$$

where  $X(k)$  is the input matrix. The input layer weight update is as follows:

$$w_x(k+1) = w_x(k) + \eta e(k)(I - S_j(k)^2)w_o(k)X(k) \quad (18)$$

The weights between the hidden and context layer are similarly updated as done in Eq. (10) and Eq. (13) as

$$w_p(k+1) = w_p(k) + \eta e(k)(I - S_j(k)^2)w_o(k)P_i(k) \quad (19)$$

The weight update formula between the input and output layer is given by

$$w_\alpha(k+1) = w_\alpha(k) + \eta e(k)X(k) \quad (20)$$

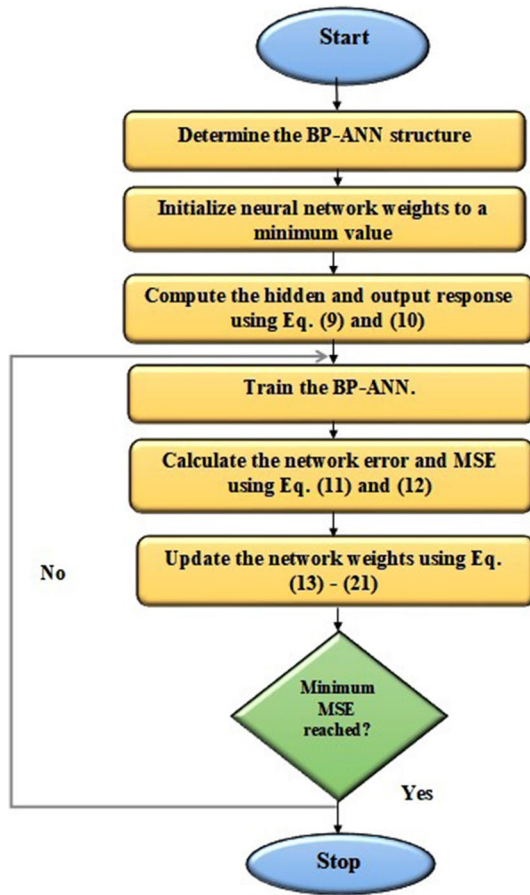


Fig. 3 Steps followed in training FCRNN structure

The weight update formula between the delay output layer and the hidden layer is as follows:

$$w_d(k+1) = w_d(k) + \eta e(k) d(k) w_o(k) (I - S_j(k)^2) \quad (21)$$

The MSE, MAE, and RMSE are calculated to evaluate the efficiency of the structures.

The various iterative steps followed in the execution of the FCRNN algorithm are also shown as a flowchart in Fig. 3.

### 3.3 Stability analysis for fixed learning rate

The convergence of the proposed structure is studied using the Lyapunov notion of stability. According to the Lyapunov-stability criteria, if there is any energy measure in the system, then the rate of change of error derives the stability of the system. The study involves deriving the weight update equations of the system and checking whether stability is achieved or not. The system is found to have achieved stability when the Lyapunov-based error function is minimum and positive. It can be expressed as

$$J = \min(E) \quad (22)$$

where  $E$  is the Lyapunov function. Once the condition is achieved, the system always will remain stable.

$$E(x) > 0 \text{ for } x > 0 \text{ and } E(x) = 0 \text{ for } x = 0 \quad (23)$$

Rate of change of error is given by

$$\frac{dE(k)}{dt} = \left( \frac{\partial E(k)}{\partial w_x(k)} + \frac{\partial E(k)}{\partial w_o(k)} + \frac{\partial E(k)}{\partial w_p(k)} + \frac{\partial E(k)}{\partial w_d(k)} + \frac{\partial E(k)}{\partial w_\alpha(k)} \right) \quad (24)$$

or

$$\frac{dE(k)}{dt} = \left( \frac{\partial E(k)}{\partial w_x(k)} \times \frac{dw_x(k)}{dt} + \frac{\partial E(k)}{\partial w_o(k)} \times \frac{dw_o(k)}{dt} + \frac{\partial E(k)}{\partial w_p(k)} \times \frac{dw_p(k)}{dt} + \frac{\partial E(k)}{\partial w_d(k)} \times \frac{dw_d(k)}{dt} + \frac{\partial E(k)}{\partial w_\alpha(k)} \times \frac{dw_\alpha(k)}{dt} \right) \quad (25)$$

Let the rate of change of weights with respect to time,  $\frac{dw_x(k)}{dt}$ ,  $\frac{dw_o(k)}{dt}$ ,  $\frac{dw_p(k)}{dt}$ ,  $\frac{dw_d(k)}{dt}$ , and  $\frac{dw_\alpha(k)}{dt}$  be taken as  $(x - y)^2$  and applying the weight update equations from Eqs. (10) to (17), the rate of change of error becomes

$$\begin{aligned} \frac{dE(k)}{dt} = & -x^2 \times (x - y)^2 (s + d(k) - 1 + y(k - 1) \\ & \times w_o(k) (I - S^2(k)) \\ & + r(k) \times w_o(k) (I - S^2(k)) - w_o(k) (I - S^2(k)) \\ & + T_1(k) \times w_o(k) (I - S^2(k)) \\ & + T_2(k) \times w_o(k) (I - S^2(k)) + T_3(k) \\ & \times w_o(k) (I - S^2(k))) \end{aligned} \quad (26)$$

where  $T_1(k)$ ,  $T_2(k)$ , and  $T_3(k)$  act as memory for past hidden context layer and  $d(k)$  acts as memory for past output context layer:

$$\frac{dE(k)}{dt} = -x^2 \times (x - y)^2 \left( \frac{\partial E(k)}{\partial w_x(k)} + \frac{\partial E(k)}{\partial w_o(k)} + \frac{\partial E(k)}{\partial w_p(k)} + \frac{\partial E(k)}{\partial w_d(k)} + \frac{\partial E(k)}{\partial w_\alpha(k)} \right) \quad (27)$$

Applying various ranges to  $x$  as given in Eq. (22) to the above equation, satisfies the condition. When the term becomes  $dE(k)/dt \leq 0$ , the system error converges to a minimum and it attains stability in terms of Lyapunov-stability analysis.



## 4 Simulation studies

In this section, a total of three examples of MISO (Multi-Input Single Output) nonlinear plant equations are considered to illustrate the efficiency of the FCRNN. Here,  $r(k)$  denotes the external input of the plant,  $y_{fcr}(k)$  denotes the output of FCRNN, and  $y_p(k)$  denotes the output of the plant. The maximum number of hidden neurons considered is 5 with a fixed learning rate of 0.001 for FCRNN, ENN, JNN, and LRNN. FFNN structure takes 6 hidden neurons with a fixed learning rate of 0.001 to match the desired model of the plant. The weights of the proposed model are updated at every iteration using a standard BP algorithm.

### 4.1 Example-1

Consider a nonlinear dynamic plant with a differential equation given as follows (Kumpati and Kannan 1990):

$$y_p(k) = \frac{y_p(k-1)}{1 + y_p^2(k-2)} + r^3(k-1) \quad (28)$$

The plant's output  $y_p(k)$  depends on both the previous input and output of the plant. The plant takes the following identification structure:

$$y_p(k) = f[y_p(k-1), y_p(k-2), r(k-1)] \quad (29)$$

The following variable input  $r(k)$  is applied to the plant:

$$r(k) = \begin{cases} \sin\left(\frac{\pi k}{45}\right), & \text{for } 0 < k \leq 250 \\ 0.1\sin\left(\frac{\pi k}{45}\right) - 0.1\cos\left(\frac{\pi k}{40}\right), & \text{for } 250 < k \leq 500 \\ -\sin\left(\frac{\pi k}{20}\right), & \text{for } 500 < k \leq 900 \end{cases} \quad (30)$$

The performance of FCRNN is compared with other network structures such as ENN, JNN, LRNN, and FFNN in terms of performance criteria such as MSE, MAE, and RMSE. The FCRNN takes the identification structure given as follows:

$$y_{fcr}(k) = \hat{f}[y_p(k-1), r(k-1)] \quad (31)$$

where  $r(k-1)$  and  $y_p(k-1)$  are two inputs considered for identification model. The ENN and JNN identification takes the following structure:

$$y_{ENN}(k) = \hat{f}[y_p(k-1), y_p(k-2), r(k-1)] \quad (32)$$

$$y_{JNN}(k) = \hat{f}[y_p(k-1), y_p(k-2), r(k-1)] \quad (33)$$

FFNN is trained with a hidden neuron of 6 to match the performance of recurrent structures. Mathematically, the FFNN identification structure is given by

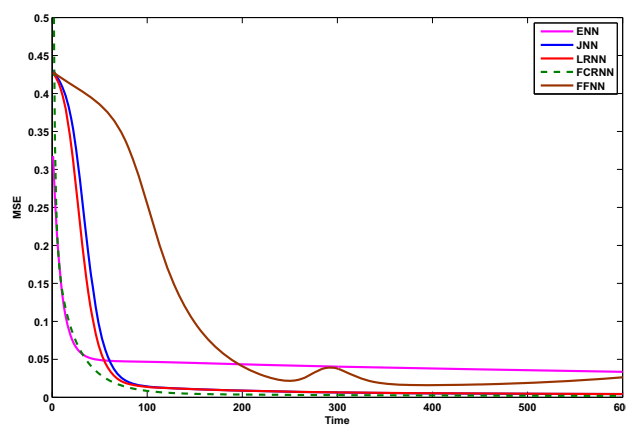


Fig. 4 Comparison of MSE plots of various structures [Example-1]

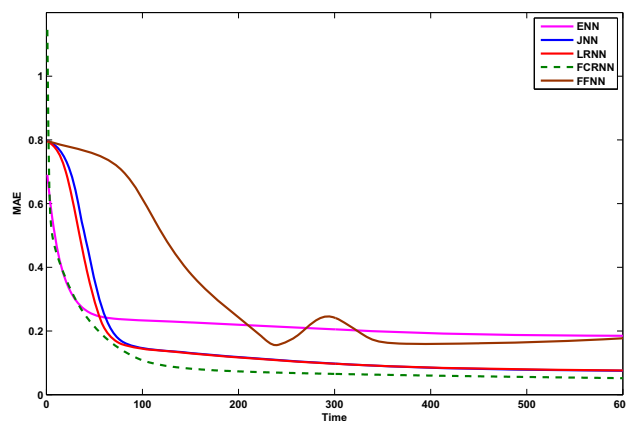


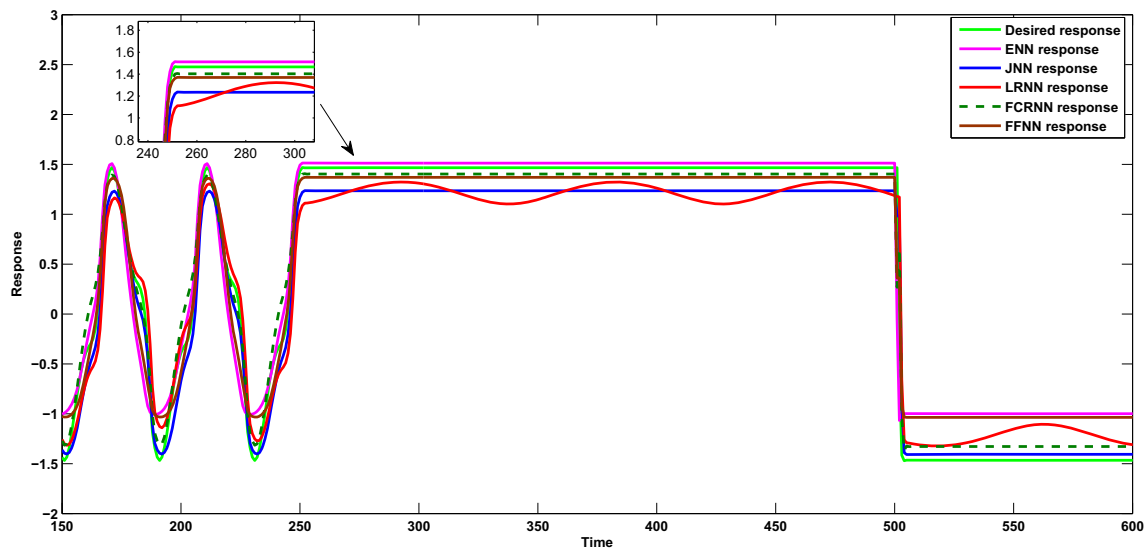
Fig. 5 Comparison of MAE plots of various structures [Example-1]

$$y_{FFNN}(k) = \hat{f}[y_p(k-1), y_p(k-2), r(k-1)] \quad (34)$$

Figure 4 shows the comparison of MSE obtained for various structures. Figure 5 shows the comparison of MAE obtained for various structures. Figure 6 shows the comparison of the response of FCRNN with other structures. Table 1 gives the comparison of the response of FCRNN with other structures chosen (best values obtained for error-based indicators are highlighted in bold). From the results, it is observed that the proposed structure requires fewer inputs to give a better prediction accuracy over other considered structures. Thus, in this example the performance comparison can be written as follows: FCRNN > ENN > JNN > LRNN > FFNN.

### 4.2 Disturbance rejection test [Example-1]

To check the disturbance recovering ability and robustness of the proposed FCRNN model, a disturbance signal is added to



**Fig. 6** Comparison of the response of FCRNN, ENN, JNN, LRNN, and FFNN at the end of the training [Example-1]

**Table 1** Comparison of performance of proposed structure with the other structures [Example-1]

Structure	No of Epochs	No of hidden neurons	No of samples in each epoch	MSE	MAE	RMSE
FCRNN	600	05	500	<b>0.000012</b>	<b>0.0021</b>	<b>0.0037</b>
ENN	600	05	500	0.0336	0.1848	0.1832
JNN	600	05	500	0.0043	0.0753	0.0656
LRNN	600	05	500	0.0042	0.0759	0.0646
FFNN	600	06	500	0.0143	0.1772	0.1195

the output of the network. A sine wave signal as a disturbance,

$$u(k) = \begin{cases} \sin\left(\frac{2\pi k}{15}\right), & \text{for } 250 < k \leq 450 \end{cases} \quad (35)$$

is introduced. This is shown in Fig. 7. The disturbance signals cause a rise in MSE value. The proposed identifier is found to bring back the increased MSE to zero, thus matching the desired performance. The comparison of disturbance rejection ability of FCRNN with other selected neural identifier is also shown in Fig. 8. Figure 8 shows that the FCRNN shows better disturbance rejection ability among others.

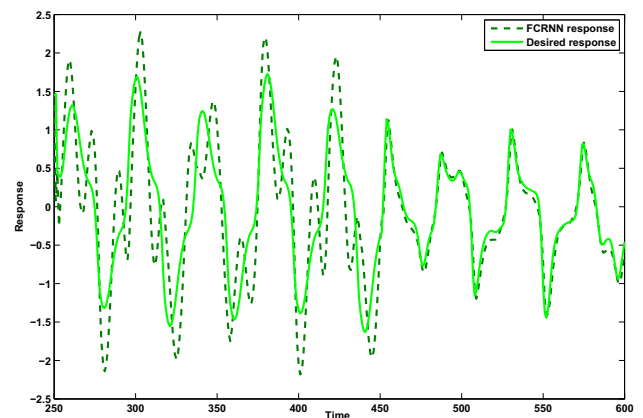
### 4.3 Example-2

In this example, a differential equation of nonlinear dynamic plant of degree 3 as given in Kumpati and Kannan (1990) is considered:

$$y_p(k) = 1.8398y_p(k-1) - 0.86070y_p(k-2) + 0.010688r(k-1) + 0.0101r(k-2) \quad (36)$$

The identification structure of the plant is as follows:

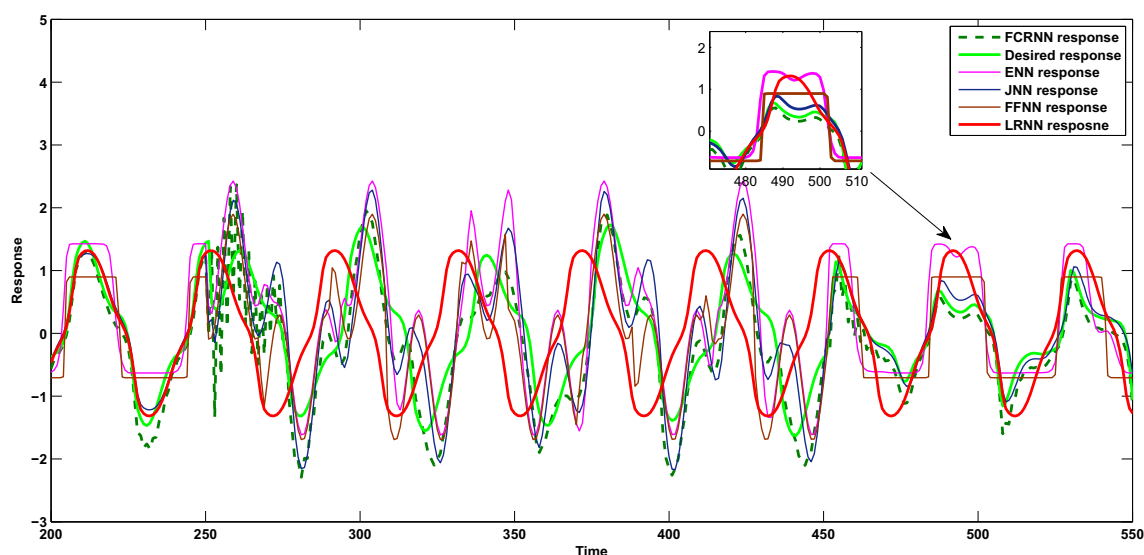
$$y_{fer}(k) = \hat{f}[y_p(k-1), y_p(k-2), r(k-1), r(k-2)] \quad (37)$$



**Fig. 7** Response of plant obtained on application of disturbance signal at the time instant [Example-1]

The output depends on previous inputs and outputs values of the plant. A variable input  $r(k)$  is supplied to the plant, where

$$r(k) = \begin{cases} \sin\left(\frac{\pi k}{45}\right), & \text{for } 0 < k \leq 250 \\ 0.1\sin\left(\frac{\pi k}{45}\right) - 0.1\cos\left(\frac{\pi k}{40}\right), & \text{for } 250 < k \leq 500 \\ -\sin\left(\frac{\pi k}{20}\right), & \text{for } 500 < k \leq 900 \end{cases} \quad (38)$$



**Fig. 8** Comparison of disturbance rejection ability of FCRNN with other selected neural structures [Example-1]

When FCRNN is considered as an identifier, it takes the identification structure as follows:

$$y_{fcr}(k) = \hat{f}[y_p(k-1), y_p(k-2), r(k-1)] \quad (39)$$

where  $y_p(k-1)$ ,  $y_p(k-2)$ , and  $r(k-1)$  are the inputs considered for identification. When ENN and JNN is considered as an identifier, they take the following identification structure, respectively:

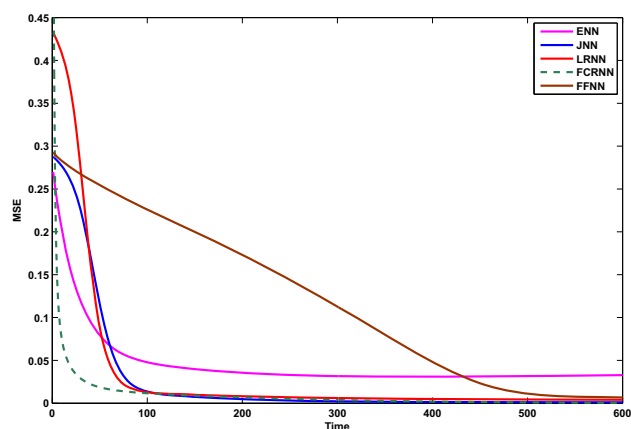
$$y_{ENN}(k) = \hat{f}[y_p(k-1), y_p(k-2), r(k-1), r(k-2)] \quad (40)$$

$$y_{JNN}(k) = \hat{f}[y_p(k-1), y_p(k-2), r(k-1), r(k-2)] \quad (41)$$

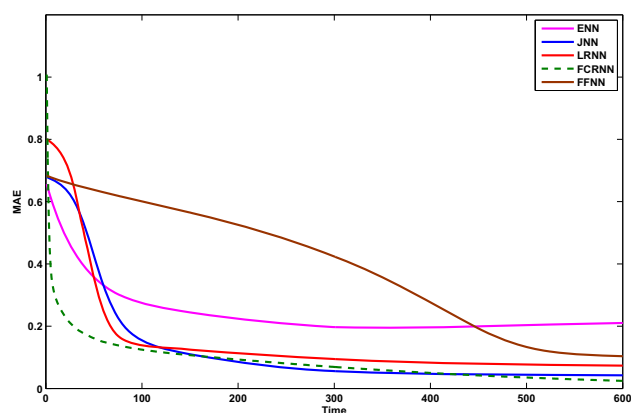
FFNN is trained with a hidden neuron of 6 to match the performance of recurrent structures. The FFNN takes the identification structure as follows:

$$y_{FFNN}(k) = \hat{f}[y_p(k-1), y_p(k-2), r(k-1), r(k-2)] \quad (42)$$

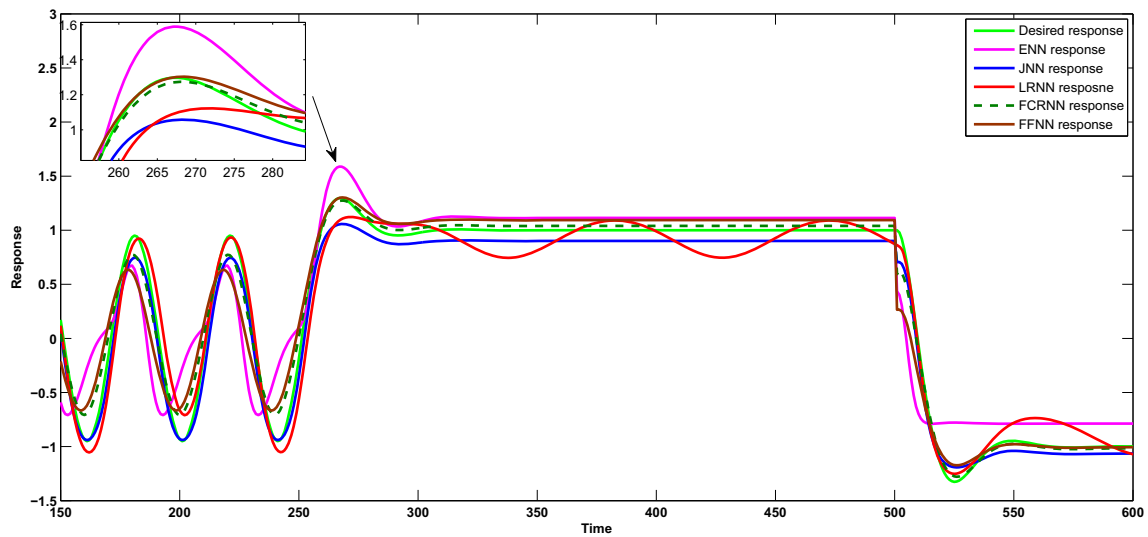
Figure 9 shows the comparison of MSE obtained for various structures. Figure 10 shows the comparison of MAE obtained for various structures. Figure 11 shows the performance of identified structure among others. Table 2 gives the comparison of FCRNN with other structures chosen (best values obtained for error-based indicators are highlighted in bold). The results from Table 2 shows that the proposed method has better prediction accuracy over other considered structures with lesser inputs.



**Fig. 9** Comparison of MSE plots of various structures [Example-2]



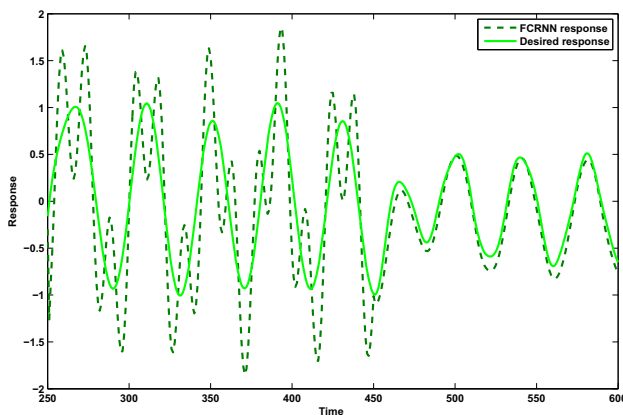
**Fig. 10** Comparison of MAE plots of various structures [Example-2]



**Fig. 11** Comparison of the response of FCRNN, ENN, JNN, LRNN, and FFNN at the end of the training [Example-2]

**Table 2** Comparison of performance of proposed structure with the other structures [Example-2]

Structure	No of epochs	No of hidden neurons	No of samples in each epoch	MSE	MAE	RMSE
FCRNN	600	05	500	<b>0.0000183</b>	<b>0.0024</b>	<b>0.0043</b>
ENN	600	05	500	0.0388	0.2266	0.1969
JNN	600	05	500	0.0011	0.0429	0.0339
LRNN	600	05	500	0.0012	0.0431	0.0340
FFNN	600	06	500	0.0105	0.1185	0.1026



**Fig. 12** Response of plant obtained on application of disturbance signal at the time instant [Example-2]

#### 4.4 Disturbance rejection test [Example-2]

The robustness of the proposed FCRNN model is tested by adding a disturbance signal as a sine wave introduced in a range of time instants between  $250 < k \leq 450$  at the output as

$$u(k) = \begin{cases} \sin\left(\frac{2\pi k}{15}\right), & \text{for } 250 < k \leq 450 \end{cases} \quad (43)$$

The corresponding response is shown in Fig. 12. The proposed identifier is robust enough to bring back the increased MSE to zero matching the desired plant response. The comparison of disturbance rejection ability of FCRNN with other selected neural identifier is also shown in Fig. 13. Figure 13 shows that the FCRNN shows better disturbance rejection ability among others.

#### 4.5 Example-3

Consider the following nonlinear plant with second-order differential equation as given in Kumpati and Kannan (1990):

$$y_p(k) = 0.72y_p(k-1) + 0.025y_p(k-2)r(k-1) + 0.001r^2(k-2) + 0.2r(k-3) \quad (44)$$

The identification structure of the plant will be

$$y_p(k) = \hat{f}[y_p(k-1), y_p(k-2), r(k-1), r(k-2), r(k-3)] \quad (45)$$

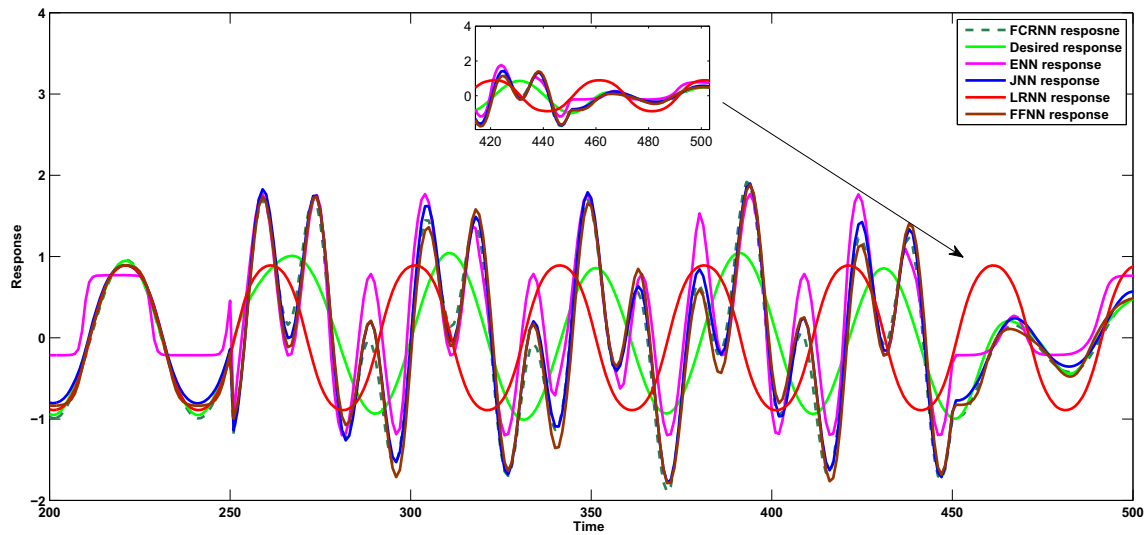


Fig. 13 Comparison of disturbance rejection ability of FCRNN with other selected neural structures [Example-2]

The above equation is supplied a variable input  $r(k)$ , where

$$r(k) = \begin{cases} \sin\left(\frac{\pi k}{45}\right), & \text{for } 0 < k \leq 250 \\ 0.1\sin\left(\frac{\pi k}{45}\right) - 0.1\cos\left(\frac{\pi k}{40}\right), & \text{for } 250 < k \leq 500 \\ -\sin\left(\frac{\pi k}{20}\right), & \text{for } 500 < k \leq 900 \end{cases} \quad (46)$$

The proposed FCRNN identification structure will be of the form given as follows:

$$y_{fcr}(k) = \hat{f}[y_p(k-1), r(k-1)] \quad (47)$$

where  $r(k-1)$  and  $y_p(k-1)$  are two inputs considered for the identification model. Mathematically, the ENN and JNN identification structure is given by

$$y_{ENN}(k) = \hat{f}[y_p(k-1), y_p(k-2), r(k-1), r(k-2), r(k-3)] \quad (48)$$

$$y_{JNN}(k) = \hat{f}[y_p(k-1), y_p(k-2), r(k-1), r(k-2), r(k-3)] \quad (49)$$

FFNN is trained with a hidden neuron of 6 to match the performance of recurrent structures:

$$y_{FFNN}(k) = \hat{f}[y_p(k-1), y_p(k-2), r(k-1), r(k-2), r(k-3)] \quad (50)$$

where  $r(k-1)$ ,  $r(k-2)$ ,  $r(k-3)$ ,  $y_p(k-1)$  and  $y_p(k-2)$  are the inputs considered for model identification. From the results, the proposed FCRNN model shows better performance and the performance can be given in the following order: FCRNN > ENN > JNN > LRNN > FFNN. Figure 14 shows the comparison of MSE obtained for plant

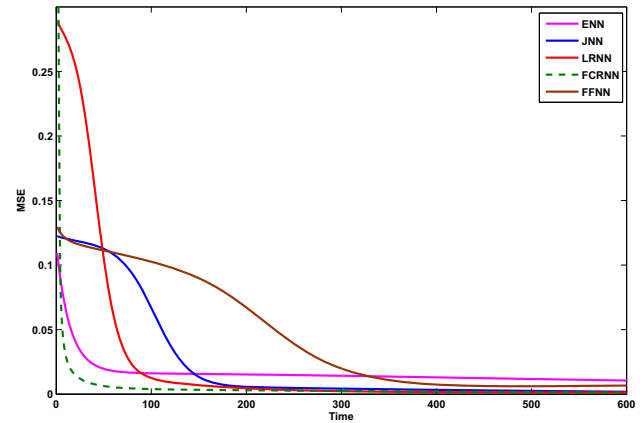


Fig. 14 Comparison of MSE plots of various structures [Example-3]

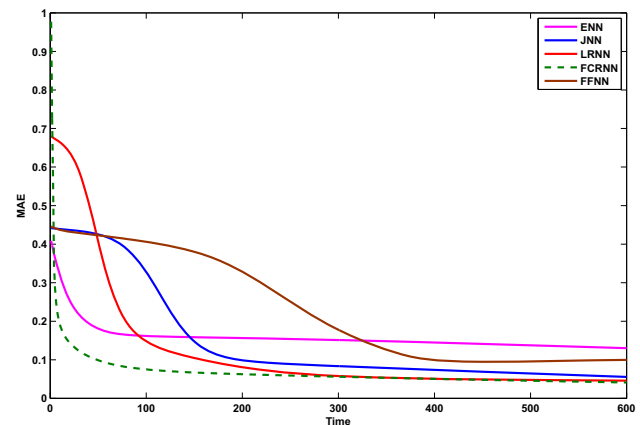


Fig. 15 Comparison of MAE plots of various structures [Example-3]

model Example-3. Figure 15 shows the comparison of MAE obtained for plant model Example-3. Figure 16 shows the performance of the proposed identifier among others. Table 3 also gives the comparison of FCRNN with other structures chosen in terms of error-based indicators (best values obtained for error-based indicators are highlighted in bold). It can be seen from Table 3 and Figs. 14, 15, and 16 that the proposed method gives a better prediction accuracy as compared to other considered structures with lesser inputs.

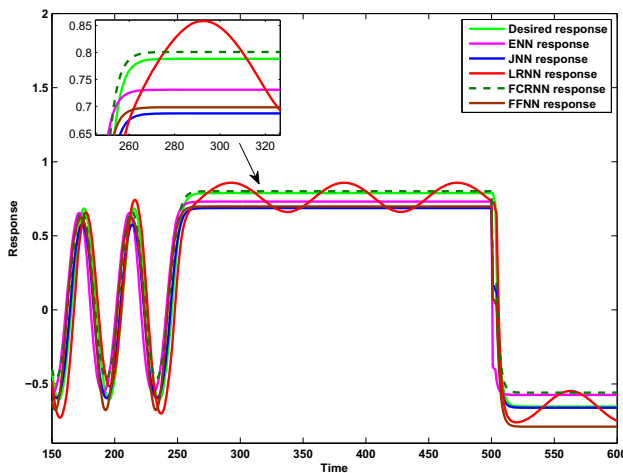
#### 4.6 Disturbance rejection test [Example-3]

The robustness and efficiency of the proposed FCRNN method are tested for Example-3. A disturbance signal in terms of sine wave,

$$u(k) = \left\{ \sin\left(\frac{2\pi k}{15}\right), \text{ for } 250 < k \leq 450 \right. \quad (51)$$

is introduced and added to the obtained output of the model. This is shown in Fig. 17. The proposed method is found to respond quickly and bring down the increased MSE to zero, hence tracking the plant's desired response.

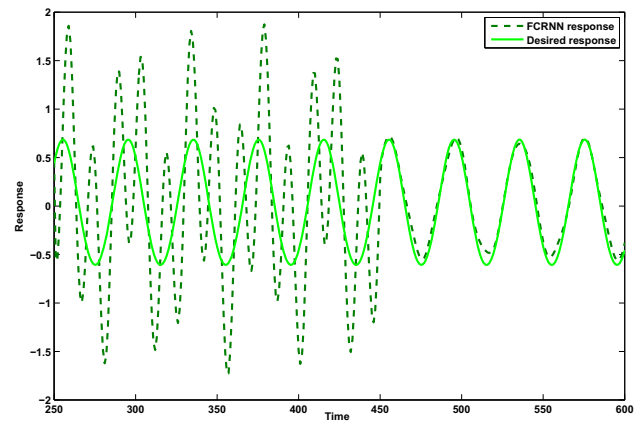
The comparison of disturbance rejection ability of FCRNN with other selected neural identifier is also shown in Fig. 18 and it can be seen from this figure that the FCRNN has shown



**Fig. 16** Comparison of the response of FCRNN, ENN, JNN, LRNN, and FFNN at the end of the training [Example-3]

**Table 3** Comparison of performance of proposed structure with the other structures [Example-3]

Structure	No of epochs	No of hidden neurons	No of samples in each epoch	MSE	MAE	RMSE
FCRNN	600	05	500	<b>0.00000869</b>	<b>0.0024</b>	<b>0.0029</b>
ENN	600	05	500	0.0127	0.1435	0.1128
JNN	600	05	500	0.0020	0.0569	0.0442
LRNN	600	05	500	0.0019	0.0558	0.0435
FFNN	600	06	500	0.0080	0.1112	0.0896



**Fig. 17** Response of plant obtained on application of disturbance signal at the time instant [Example-3]

better disturbance rejection ability as compared to other considered models.

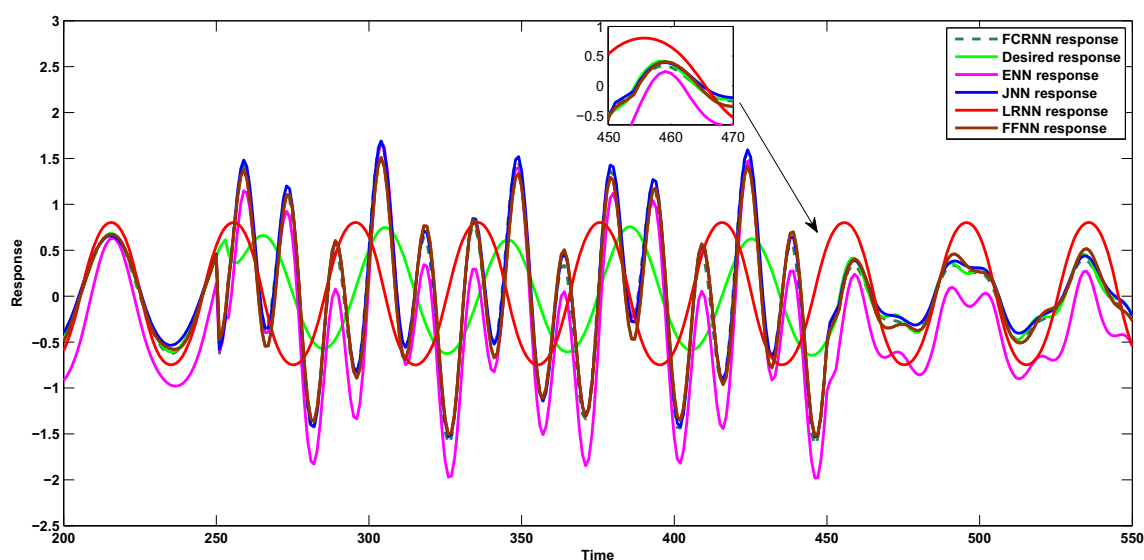
## 5 Conclusion

In this work, a novel recurrent neural network, known as the FCRNN model, is proposed for the identification of complex nonlinear dynamical systems. The weights of the proposed model are updated using the BP method. The convergence of the learning algorithm is proved using the Lyapunov-stability analysis. A total of 3 simulation examples are considered in the experimental study for testing the identification ability of the proposed model. The comparison is done in terms of MSE, MAE, RMSE, the number of hidden neurons, the number of input parameters, and the recovering ability of the structure. From the results and simulation, it can be seen that the proposed FCRNN performed better than the other selected neural models for all the considered examples. The ability of FCRNN to recover from any external disturbance is also found to be fast than ENN, JNN, LRNN and FFNN structures.

### 5.1 Future studies

The proposed structure proves to be a better model for the identification of nonlinear dynamic systems. Future research





**Fig. 18** Comparison of disturbance rejection ability of FCRNN with other selected neural structures [Example-3]

will concentrate on the use of meta-heuristic approaches to train and update parameters, such as the Genetic Algorithm (GA) (Ling et al. 2003; El-Shorbagy and El-Refaey 2020; Wang and Qing 2021), Particle Swarm Optimization (PSO) (Kang et al. 2014; Ge et al. 2007; Song et al. 2007) and Firefly Algorithm (FA) (Ariyaratne et al. 2020; Zhang et al. 2021; Yang and He 2018). Since these algorithms do not have the tendency to stuck in the local minimum, they can be integrated with Lyapunov-stability approach for most effective learning of the structure. Future studies will also focus on determining the optimal number of hidden layer neurons in the neural network, to extend the identification structure for Multi-Input Multi-Output (MIMO) system identification. Further, using the identification structure, this work might be extended to control a real-time nonlinear dynamic system.

**Author Contributions** The authors, Shobana. R, Bhavnes Jain and Rajesh Kumar, have contributed to the study, conception, design, material preparation, and analysis. The whole draft of the manuscript was written jointly by Shobana. R, Bhavnes Jain and Rajesh Kumar.

**Funding** The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

**Data availability** No dataset was used in the conducted study.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

**Ethical approval** This article does not contain any studies with human participants or animals performed by any of the authors.

**Informed consent** This research does not involve any human participants or animals.

## References

- Abdollahi F, Talebi H, Patel R (2003) A stable neural network-based identification scheme for nonlinear systems. In: Proceedings of the 2003 American control conference, 2003, vol 4, IEEE, pp 3590–3595
- Aggarwal CC et al (2018) Neural networks and deep learning. Springer, Berlin, p 3
- Ariyaratne M, Fernando T, Weerakoon S (2020) A self-tuning algorithm to approximate roots of systems of nonlinear equations based on the firefly algorithm. *Int J Swarm Intell* 5(1):60–96
- Basheer IA, Hajmeer M (2000) Artificial neural networks: fundamentals, computing, design, and application. *J Microbiol Methods* 43(1):3–31
- Bhat N, McAvoy TJ (1990) Use of neural nets for dynamic modeling and control of chemical process systems. *Comput Chem Eng* 14(4–5):573–582
- Calin O (2020) Deep learning architectures. Springer, Berlin
- Chen S, Billings SA (1992) Neural networks for nonlinear dynamic system modelling and identification. *Int J Control* 56(2):319–346
- Coban R (2013) A context layered locally recurrent neural network for dynamic system identification. *Eng Appl Artif Intell* 26(1):241–250
- Deng J (2013) Dynamic neural networks with hybrid structures for nonlinear system identification. *Eng Appl Artif Intell* 26(1):281–292
- Elsheikh AH, Sharshir SW, Abd Elaziz M, Kabeel A, Guilan W, Haiou Z (2019) Modeling of solar energy systems using artificial neural network: a comprehensive review. *Solar Energy* 180:622–639
- El-Shorbagy MA, El-Refaey AM (2020) Hybridization of grasshopper optimization algorithm with genetic algorithm for solving system of non-linear equations. *IEEE Access* 8:220944–220961
- Gao X, Gao X-M, Ovaska S, A modified elman neural network model with application to dynamical systems identification. In: (1996) IEEE international conference on systems, man and cybernetics. Information intelligence and systems (Cat. No. 96CH35929), vol 2. IEEE 1996, pp 1376–1381
- Ge H-W, Liang Y-C, Marchese M (2007) A modified particle swarm optimization-based dynamic recurrent neural network for identifying and controlling nonlinear systems. *Comput Struct* 85(21–22):1611–1622

- Ge H-W, Du W-L, Qian F, Liang Y-C (2009) Identification and control of nonlinear systems by a time-delay recurrent neural network. *Neurocomputing* 72(13–15):2857–2864
- Haykin S (2009) *Neural networks and learning machines*, 3/E. Pearson Education India
- Kang J, Meng W, Abraham A, Liu H (2014) An adaptive pid neural network for complex nonlinear system control. *Neurocomputing* 135:79–85
- Kroll A, Schulte H (2014) Benchmark problems for nonlinear system identification and control using soft computing methods: need and overview. *Appl Soft Comput* 25:496–513
- Ku C-C, Lee KY (1995) Diagonal recurrent neural networks for dynamic systems control. *IEEE Trans Neural Netw* 6(1):144–156
- Kumar R, Srivastava S, Gupta J, Mohindru A (2019) Comparative study of neural networks for dynamic nonlinear systems identification. *Soft Comput* 23(1):101–114
- Kumpati SN, Kannan P et al (1990) Identification and control of dynamical systems using neural networks. *IEEE Trans Neural Netw* 1(1):4–27
- Li S (2001) Comparative analysis of backpropagation and extended Kalman filter in pattern and batch forms for training neural networks. In: *IJCNN'01. international joint conference on neural networks*. In: *Proceedings (Cat. No. 01CH37222)*, vol 1, IEEE, pp 144–149
- Li X, Bai Y, Huang C (2008) Nonlinear system identification using dynamic neural networks based on genetic algorithm. In: 2008 international conference on intelligent computation technology and automation (ICICTA), vol 1, IEEE, pp 213–217
- Ling S-H, Leung FH-F, Lam H-K, Lee Y-S, Tam PK-S (2003) A novel genetic-algorithm-based neural network for short-term load forecasting. *IEEE Trans Ind Electron* 50(4):793–799
- Ljung L (2010) Perspectives on system identification. *Annu Rev Control* 34(1):1–12
- Luttmann L, Mercorelli P (2021) Comparison of backpropagation and Kalman filter-based training for neural networks. In: 2021 25th international conference on system theory, control and computing (ICSTCC), IEEE, pp 234–241
- Moeller DP (2004) Parameter identification of dynamic systems. In: *Mathematical and computational modeling and simulation*. Springer, pp 257–310
- Noël J-P, Kerschen G (2017) Nonlinear system identification in structural dynamics: 10 more years of progress. *Mech Syst Signal Process* 83:2–35
- Ogunmolu O, Gu X, Jiang S, Gans N (2016) Nonlinear systems identification using deep dynamic neural networks. [arXiv:1610.01439](https://arxiv.org/abs/1610.01439)
- Pham DT, Karaboga D (1999) Training Elman and Jordan networks for system identification using genetic algorithms. *Artif Intell Eng* 13(2):107–117
- Quaranta G, Lacarbonara W, Masri SF (2020) A review on computational intelligence for identification of nonlinear dynamical systems. *Nonlinear Dyn* 99(2):1709–1761
- Sanchez EN (1994) Dynamic neural networks for nonlinear systems identification. In: *Proceedings of 1994 33rd IEEE conference on decision and control*, vol 3, IEEE, pp 2480–2481
- Sastry P, Santharam G, Unnikrishnan K (1994) Memory neuron networks for identification and control of dynamical systems. *IEEE Trans Neural Netw* 5(2):306–319. <https://doi.org/10.1109/72.729193>
- Savran A (2007) Multifeedback-layer neural network. *IEEE Trans Neural Netw* 18(2):373–384
- Schubert M, Köppen-Seliger B, Frank PM (1997) Recurrent neural networks for nonlinear system modelling in fault detection. *IFAC Proc Vol* 30(18):701–706
- Şen GD, Günel GÖ, Güzelkaya M, Extended kalman filter based modified Elman-Jordan neural network for control and identification of nonlinear systems. In: (2020) *Innovations in intelligent systems and applications conference (ASYU)*. IEEE 2020, pp 1–6
- Song Y, Chen Z, Yuan Z (2007) New chaotic pso-based neural network predictive control for nonlinear process. *IEEE Trans Neural Netw* 18(2):595–601
- Thammano A, Ruxpakawong P (2009) Dynamic system identification using recurrent neural network with multi-valued connection weight. In: 2009 IEEE international conference on fuzzy systems, IEEE, pp 2077–2082
- Thammano A, Ruxpakawong P (2010) Nonlinear dynamic system identification using recurrent neural network with multi-segment piecewise-linear connection weight. *Memet Comput* 2(4):273–282
- Veerasamy V, Wahab NIA, Ramachandran R, Othman ML, Hizam H, Kumar JS, Irudayaraj AXR (2022) Design of single-and multi-loop self-adaptive pid controller using heuristic based recurrent neural network for alfc of hybrid power system. *Expert Syst Appl* 192:116402
- Wang Y (2017) A new concept using lstm neural networks for dynamic system identification. In: (2017) *American control conference (ACC)*. IEEE 2017, pp 5324–5329
- Wang Y, Qing D (2021) Model predictive control of nonlinear system based on ga-rbp neural network and improved gradient descent method. *Complexity* 2021:1–14
- Willis MJ, Montague GA, Di Massimo C, Tham MT, Morris AJ (1992) Artificial neural networks in process estimation and control. *Automatica* 28(6):1181–1187
- Yang XS, He XS (2018) Why the firefly algorithm works? *Nat Inspir Algor Appl Optim* 245–259
- Yazdizadeh A, Khorasani K (1997) Identification of a class of nonlinear systems using dynamic neural network structures. In: *Proceedings of international conference on neural networks (ICNN'97)*, vol 1, IEEE, pp 194–198
- Yazdizadeh A, Khorasani K (2002) Adaptive time delay neural network structures for nonlinear system identification. *Neurocomputing* 47(1):207–240. [https://doi.org/10.1016/S0925-2312\(01\)00589-6](https://doi.org/10.1016/S0925-2312(01)00589-6)
- Yu D, Wang Y, Liu H, Jermsittiparsert K, Razmjoooy N (2019) System identification of pem fuel cells using an improved Elman neural network and a new hybrid optimization algorithm. *Energy Rep* 5:1365–1374
- Zhang H, Zhang R, He Q, Liu L (2021) Variable universe fuzzy control of high-speed elevator horizontal vibration based on firefly algorithm and backpropagation fuzzy neural network. *IEEE Access* 9:57020–57032

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

# Detection of Sleep Apnea and its Intensity in Adults

Bhupinder Singh Saini

Department of Electronics and Communications  
Delhi Technological University,  
Delhi, India  
bhupindersingh.saini18@gmail.com

Ayussh Vashishth

Department of Electronics and Communications,  
Delhi Technological University,  
Delhi, India  
ayusshvashishth2208@gmail.com

Chirag Kaushik

Department of Electronics and Communications  
Delhi Technological University,  
Delhi, India  
ckaushik1799@gmail.com

Lavi Tanwar

Department of Electronics and Communications, Faculty,  
Delhi Technological University,  
Delhi, India  
lavi.tanwar@dtu.ac.in

**Abstract** - The following paper introduces a new method for identifying Obstructive Sleep Apnea (OSA), a widespread sleep disorder that impacts a large number of people globally. OSA is characterized by breathing pauses lasting from a few seconds to a minute or more. Our proposed approach utilizes audio signals for OSA detection. Existing studies require the use of ECG or EEG signals, which entail bulky equipment, electrodes, and instruments attached to the patient, resulting in a time-consuming and inconvenient signal extraction process. Conversely, our study uses audio signals due to their accessibility and convenience. To accurately detect OSA, we convert audio signals to time and frequency domains using FFT and DWT. Features are then extracted and used in the ANN model to obtain high accuracy and specificity in OSA detection. The proposed approach achieves high accuracy and specificity in detecting OSA. With the ANN model, we achieved an accuracy of 94.1%, sensitivity of 98.5%, and specificity of 88.7%. This indicates the potential of using audio signals for OSA detection, serving as a non-invasive and cost-effective method for OSA diagnosis.

**Keywords** - Obstructive Sleep Apnea, FFT (Fast Fourier Transform), Spectrogram, Kurtosis, ANN (Artificial Neural Network).

## I. INTRODUCTION

Sleep apnea is a condition where an individual experiences repeated disruptions in their breathing while sleeping, which may involve breaths that are not as deep as usual or can stop altogether for a few seconds to a minute or longer. These interruptions occur multiple times throughout the night, often following episodes of loud snoring. When breathing resumes, it might be accompanied by a snorting or choking sound. "OSA affects between 1% and 6% of individuals" [1]. Although sleep apnea can impact individuals of all age groups, it is more commonly observed in people between the ages of 55 and 60.

It is imperative to refer to scientific data that indicates a healthy individual typically exhibits a respiratory rate ranging between 12 to 20 breaths per minute. In other words, this means it takes about 3-5 seconds (for male patients) and 2-4 seconds (for female patients) for each breath. This information can serve as a baseline to compare with the breathing rate of patients. Generally, snoring can be observed in patients who have OSA. The average frequency at which patients snore is in the lower range of frequency and varies from 200-350 Hz.

## II. LITERATURE REVIEW

Several electrophysiological signals, including the Electromyogram (EMG), Electrocardiogram (ECG), Electroencephalogram (EEG), and ECG Derived Respiration (EDR) signals, were used to investigate and diagnose OSA. Several research papers have employed ECG [4-7], EEG [8-10], and SPO<sub>2</sub> [11-12] signals to diagnose this illness, and there were instances of using multiple signals for diagnosis [13-15]. Additionally, thoracic and abdominal signals were studied for diagnosis but didn't yield optimal results [16]. They were primarily detected during polysomnography (PSG). Statistics were derived for kurtosis, variance, mean, median, and standard deviation for constructing machine learning models to detect sleep apnea [2][3].

In [18], authors used wavelet transform and ANN on ECG signals to distinguish between normal and apneic patients achieving a specificity of 44% and a sensitivity of 70%. Using ECG, the authors of [17] proposed an automated method for apnea diagnosis. The segmented ECG sub bands obtained by DWT are divided into three sets of features. These characteristics had been used by RF classification to distinguish between normal and apnea ECG segments. The results showed that wavelet-based features can diagnose OSA patients by the suggested method's achieving 90% classification accuracy.

The study conducted in [19] utilized EEG, EMG, and ECG signals to differentiate between normal and sleep apnea patterns during sleep. The obtained characteristics were then fed into an MLP artificial neural network, which carried out linear and non-linear analyses on the signals.

In existing studies, OSA detection involves the use of heavy machinery, electrodes, and equipment attached to the subject, which makes signal extraction a time-consuming and inconvenient process. However, our study utilizes audio signals due to their accessibility and convenience in acquisition.

The following are the advantages of using audio signals in our study:

- Audio signals can be easily recorded and saved using a standard smartphone, eliminating the need for costly equipment.
- Compared to other types of signals, audio signals provide a practical and affordable option for both researchers and practitioners.

- Signal extraction from audio signals is a non-invasive process, making it comfortable for the patient.
- Audio signals can be recorded during the patient's natural sleep, ensuring a realistic and accurate representation of their sleep patterns.
- Using audio signals in OSA detection can lead to more widespread and convenient screening, ultimately improving patient outcomes.

### III. METHODOLOGY

The workflow for our study can be described as follows.

- Recorded audio samples were subjected to noise reduction.
- Time domain analysis was performed on the processed signals by converting them into the time domain and comparing the respiratory rates.
- For frequency domain analysis, the project was divided into two parts.
- The first part involved Fast Fourier Transform (FFT) analysis to detect episodes of apnea.
- The second part involved applying Discrete Wavelet Transform (DWT) to extract relevant features.
- Features extracted were then utilized to train and classify the signals using an Artificial Neural Network (ANN) model.

This comprehensive workflow allowed for a thorough analysis of the audio signals, which facilitated accurate diagnosis and quantification of the underlying respiratory disorder. There are the prerequisites for the study :

#### • Data Acquisition

For the analysis, the data is acquired manually by recording the sleeping audio of the apnea patient and healthy person. Each audio file is nearly 1-minute duration. The audio data is converted into amplitude signal values in MATLAB [20]. We have taken total a total of 118 numerous samples of apnea and healthy patients for the analysis, 59 each.

#### • Denoising of signals

The MATLAB software is utilized to convert the signals into amplitude form, following which they are subjected to a Savitzky-Golay FIR filter for processing. The Savitzky-Golay filter, which is also referred to as a digital smoothing polynomial filter or a least-squares smoothing filter, is utilized to eradicate signal interference from signals that have a broad spectrum of frequencies.

#### • Sampling rate

Humans can sense frequency from 20Hz to 20kHz, which simply means that sampling rate at its maxima will be 40kHz (20\*2kHz). Fs is kept at 44.1kHz (an extra 4.1kHz to improve audio quality) for recording purposes

The entire research is separated into three distinct sections:

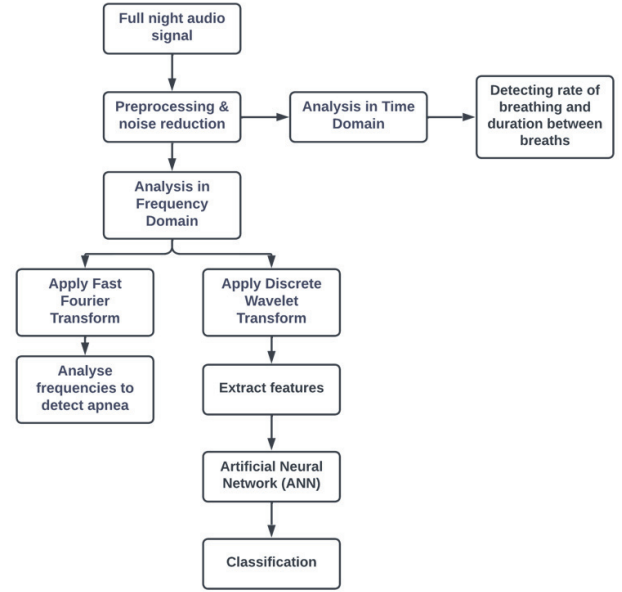


Fig. 1. Detection steps for sleep apnea

#### A. Time Domain

As mentioned in introduction, if the average duration of breaths is longer than 5 seconds, it can serve as a significant indication of the disorder. The audio signals were converted into discrete time signals, and then plotted to calculate the average time duration between the peaks obtained. This enabled us to determine the time duration between two consecutive breaths of the subject, which could then be compared with the normal time duration of a healthy person.

With observations, we can say if average breath rate is:

- 2-4 sec = patient has normal sleep
- 4-6 sec = patient has medium OSA symptoms
- > 6 sec = patient have strong signs of OSA and should take instant medical recommendations

#### B. Frequency Domain

This particular segment is further categorized into two distinct sub-divisions:

##### 1) FFT

If the frequency of snoring exceeds 350 Hz, then this implies that the patient is having a hard time breathing. Resulting in higher frequency of snores (crests exceeding 500 Hz). Which is also a referral to the disorder. With the analysis in frequency spectrum, we concluded about the quality of sleep of the patient. Along with this we also concluded the range of frequency in which patient's snores lie. This helped us in further analyzing the severity of the disorder.

##### 2) DWT

After being denoised, the signals are decomposed at various levels of the discrete wavelet transform (DWT). In functional and numerical analysis, the DWT [26] refers to a type of wavelet transform that involves discretely sampled wavelets. Similar to other wavelet transforms, it offers an advantage over Fourier transforms in terms of temporal



resolution. This is due to its ability to capture both frequency and temporal information, i.e., the location in time [23].

To analyze and study the results of different DWT levels, we calculated the **kurtosis** parameter [24-25] for the detailed and approximate coefficient vectors at every level. Kurtosis is a statistical measure used to assess the shape of a distribution, particularly its degree of peakedness and the presence of outliers. Tailedness is a factor that is taken into account when evaluating kurtosis. Mathematically, kurtosis is defined as [27]:

$$\text{Kurtosis} = \frac{\sum_{i=1}^n f_i(x_i - \bar{x})^4}{n\sigma^4}$$

where  $x_i$  is input,  $n$  is size of the data,  $\bar{x}$  is the mean and  $\sigma$  is the standard deviation [21][22].

### C. Artificial Neural Network

For classification of audio-signals into healthy signals and signals from apnea patients, four features are taken out from the audio-signal which are: kurtosis, standard deviation, variance and skewness.

$$\text{Variance} = \frac{\sum(x_i - \bar{x})^2}{N-1}$$

$$\text{Skewness} = \frac{\sum(x_i - \bar{x})^3}{(N-1)\sigma^3}$$

$$\text{Standard Deviation} = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

$$\text{Kurtosis} = \frac{\mu_4}{\sigma^4}$$

To detect sleep apnea, we have utilized an Artificial Neural Network (ANN) for machine learning purposes. ANN is a versatile and adaptive tool that is commonly employed to solve a wide range of problems. In this study, we have employed a single hidden layer comprising ten neurons and an output layer consisting of a single neuron. We have used ANN classifier where normal audio-signal is assigned 0 and apnea audio-signal is assigned 1 [17].

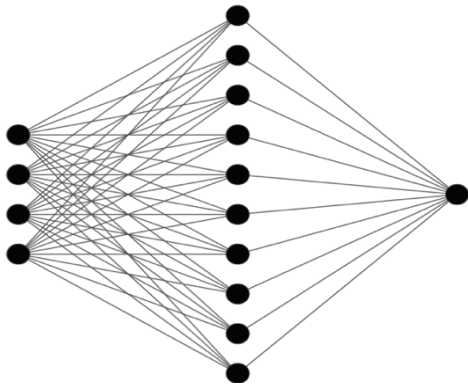


Fig. 2. Artificial Neural Network layers

## IV. RESULTS AND DISCUSSION

All the implementation is done in MATLAB R2020a software and 1.8 GHz Dual-Core Intel Core i5 processor is used.

### A. Time Domain Analysis

Here, a function is used to regulate the marking of every breath. This is done by keeping a minimum distance and a

min threshold respectively. Also mean is calculated to find the mean time of the breaths in the audio signals with the help of marked points. The amplitude vs time plots are shown in Fig 3 and 4.

Fig 5 and 6 shows that the mean respiratory rate of the patient during disrupted sleep was calculated to be 7.798 seconds, indicating the presence of obstructive sleep apnea compared to mean breath reate for normal sleep, i.e, 2.1877 seconds.

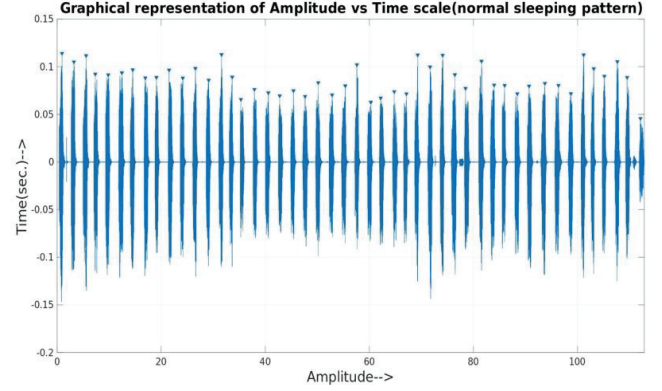


Fig. 3. Amplitude vs time plot for normal patient

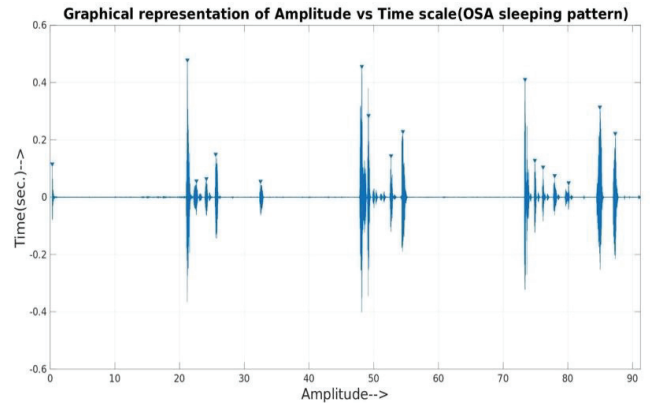


Fig. 4. Amplitude vs time plot for OSA patient

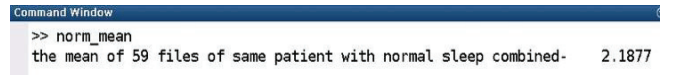


Fig. 5. Mean breath rate for normal sleep

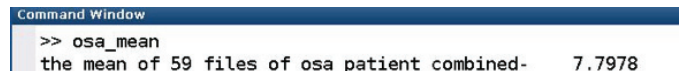


Fig. 6. Mean breath rate for OSA patient

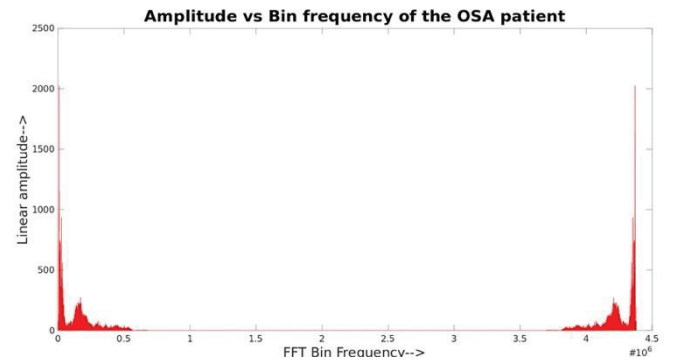


Fig. 7. Linear amplitude vs FFT bin frequency for normal breathing pattern

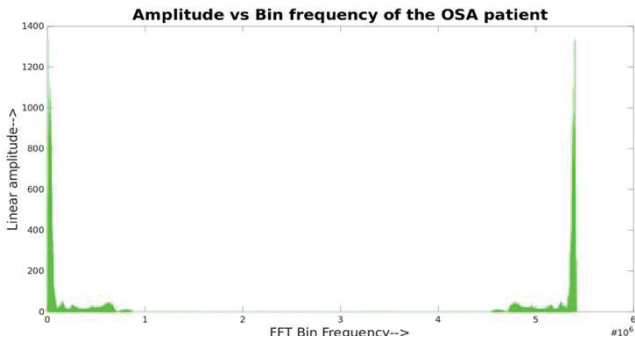


Fig. 8. Linear amplitude vs FFT bin frequency for OSA patient

## B. Frequency Analysis

### 1) FFT Analysis

After conducting FFT analysis on both the audio signals of a healthy person and a person with Obstructive Sleep Apnea (OSA), the following conclusions were made:

- The frequency range of most of the normal patients' audio signals lies below the 500Hz mark, indicating that frequency ranges below 500Hz are observed in normal sleeping patterns.
- In contrast, the frequency range of most of the OSA patient's audio signals lies well above the 500Hz mark, indicating that frequency ranges above 500Hz are observed in sleeping patterns of OSA patients.

### 2) DWT based Kurtosis Analysis

The kurtosis of the approximate and detailed coefficients of the signal is measured at various levels of DWT decomposition for both normal individuals and those with sleep apnea, as illustrated in Tables 1 and 2. The value of kurtosis for apnea patients is observed to be much higher than that of a healthy person. Therefore, by monitoring the kurtosis behavior in relation to the signal's different decomposition levels, one can potentially detect the presence of sleep apnea.

A comparison was made between the kurtosis of detailed and approximate coefficients as shown in Fig 9 and 10. The approximate coefficient kurtosis for a healthy individual is characterized by a smaller magnitude. The apnea patient's kurtosis is shown in the blue band and the healthy person's kurtosis is depicted in the red band. In both cases the kurtosis value first increases, then achieves a peak and then decreases with negative slope and finally becomes constant and with similar values after level nine.

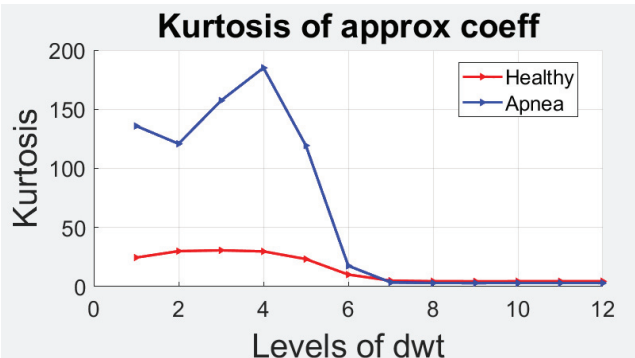


Fig. 9. Kurtosis vs levels of dwt (approximate coefficient)

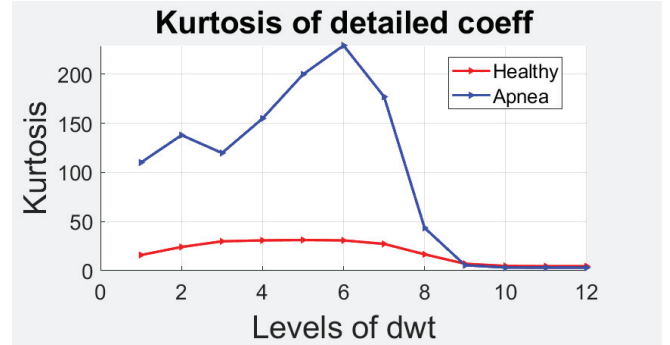


Fig. 10. Kurtosis vs levels of dwt (detailed coefficient)

### Mean kurtosis for approx coeff for osa patient

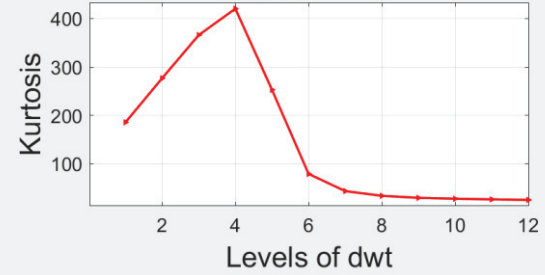


Fig. 11. Overall average kurtosis for approximate coefficient for OSA patient

### Mean kurtosis for detailed coeff for osa patient

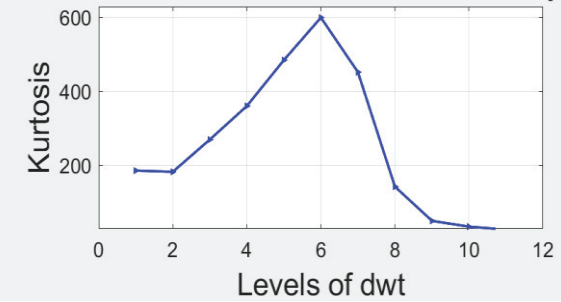


Fig. 12. Overall average kurtosis for detailed coefficient for OSA patient

To get a rough idea of detection of apnea with the audio sample of a patient, the average kurtosis value of all the DWT decomposition levels can be calculated. In this study, we have taken the mean of all the kurtosis values, for different DWT levels, and plotted for graphical representation. The results are shown in Fig 11 and 12. If the kurtosis values of the audio sample of a patient lies above this threshold band, then he/she has a high chance of having obstructive sleep apnea disorder.

TABLE I. THE KURTOSIS VALUES OF THE APPROXIMATE COEFFICIENTS FOR VARIOUS LEVELS OF DWT DECOMPOSITION.

DWT levels	Kurtosis value of the approximate coefficient for a normal individual	Kurtosis value of the approximate coefficient for OSA patient
1	24.5638	135.8807
2	29.9997	120.8092
3	30.6857	157.3536
4	29.7849	185.0657
5	15.9594	110.1359



6	10.7677	17.5913
7	5.0203	3.5448
8	4.4944	3.0145
9	4.4776	3.0344
10	4.4903	3.0440

TABLE II. THE KURTOSIS VALUES OF DETAILED COEFFICIENTS FOR VARIOUS LEVELS OF DWT DECOMPOSITION.

DWT levels	Kurtosis value of the detailed coefficient for a normal individual	Kurtosis value of the detailed coefficient for OSA patient
1	15.9594	110.1359
2	24.1574	138.0562
3	29.9142	119.8260
4	30.8500	154.9822
5	31.2855	199.9114
6	30.8434	229.3194
7	27.2823	117.1475
8	16.6353	43.5412
9	7.0301	5.5139
10	4.8079	3.2111

TABLE III. THE AVERAGE KURTOSIS VALUES OF DETAILED COEFFICIENTS FOR VARIOUS LEVELS OF DWT DECOMPOSITION.

DWT levels	Average Kurtosis of approximate coefficient	Average Kurtosis of detailed coefficient
1	185.7211	185.3485
2	276.7241	182.0624
3	365.9673	269.1619
4	420.4574	359.7925
5	252.9860	483.8299
6	79.0824	598.6875
7	43.5585	451.3092
8	33.8353	141.8838
9	29.6755	49.5226
10	27.8206	33.9098
Avg:	147.3027	233.6820

All Confusion Matrix		
Output Class	0	1
	64 54.2%	6 5.1%
	1 0.8%	47 39.8%
		Target Class
		98.5% 1.5%
		88.7% 11.3%
		94.1% 5.9%

Fig. 13. Confusion Matrix

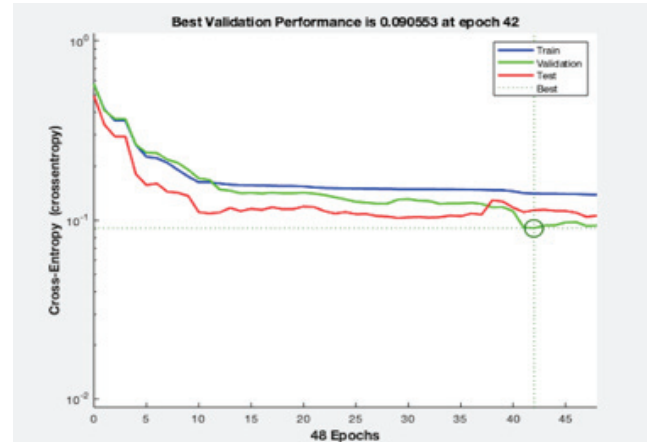


Fig. 14. ANN Performance

### C. Artificial Neural Network Results

The features extracted were used to train an ANN in MATLAB, with 64 samples for training purpose, 30 for validation, and 24 for testing. A total of 118 audio samples were utilized in the network. The results show that the accuracy achieved was 94.1%, with a sensitivity of 98.5% and a specificity of 88.7%, as illustrated in the confusion matrix represented in Figure 13.

The Artificial Neural Network showed best results at epoch 42, with 0.09 as Mean Squared Error (MSE) as shown in Fig 14. To keep the ANN simple, it was designed with only ten neurons in the hidden layer and four features as inputs. It is possible to achieve improved outcomes by augmenting either the quantity of extracted features or the amount of neurons present in the hidden layer.

### V. CONCLUSIONS

Polysomnography (PSG) is used to diagnose Obstructive Sleep Apnea (OSA), which is an uncomfortable and expensive approach. We diagnosed OSA by analyzing recorded audio signals. The promising results obtained from this method could potentially be used to estimate sleep patterns using non-contact audio technology. However, the current work only focuses on one disease, and future studies could expand to include different types of diseases.

Compared to conventional studies that utilize ECG signals, this research has distinct advantages. Specifically, the use of audio signals offers several benefits that make the approach more convenient and accessible. Audio signals can be obtained using standard consumer-grade devices like smartphones or microphones, unlike ECG signals that require specialized equipment and training to acquire and process. Moreover, analysing respiratory sounds through audio signals provides a more direct and intuitive way to gain additional insights into the nature and severity of the respiratory disorder. By taking advantage of these benefits, this study makes a valuable contribution to the diagnosis and treatment of respiratory disorders.

### REFERENCES

- [1] E. Dafna, A. Tarasiuk, and Y. Zigel, "Sleep-quality assessment from full night audio recordings of sleep apnea patients," *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 2012, pp. 3660–3663, 2012.
- [2] L. Almazaydeh, K. Elleithy, and M. Faezipour, "Obstructive sleep apnea detection using SVM-based classification of ECG signal

- features,” in 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2012.
- [3] X. Wang et al., “Obstructive sleep apnea detection using ecg-sensor with convolutional neural networks,” *Multimed. Tools Appl.*, vol. 79, no. 23–24, pp. 15813–15827, 2020.
  - [4] S. Kiranyaz, T. Ince, and M. Gabbouj, “Real-time patient-specific ECG classification by 1-D convolutional neural networks,” *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 664–675, 2016.
  - [5] L. Chen, X. Zhang, and C. Song, “An automatic screening approach for obstructive sleep apnea diagnosis based on single-lead electrocardiogram,” *IEEE Trans. Autom. Sci. Eng.*, vol. 12, no. 1, pp. 106–115, 2015.
  - [6] A. Jafari, “Sleep apnoea detection from ECG using features extracted from reconstructed phase space and frequency domain,” *Biomed. Signal Process. Control*, vol. 8, no. 6, pp. 551–558, 2013.
  - [7] B. L. Koley and D. Dey, “Automatic detection of sleep apnea and hypopnea events from single channel measurement of respiration signal employing ensemble binary SVM classifiers,” *Measurement (Lond.)*, vol. 46, no. 7, pp. 2082–2092, 2013.
  - [8] S. Saha, A. Bhattacharjee, M. A. A. Ansary, and S. A. Fattah, “An approach for automatic sleep apnea detection based on entropy of multi-band EEG signal,” in 2016 IEEE Region 10 Conference (TENCON), 2016.
  - [9] F. Ahmed, P. Paromita, A. Bhattacharjee, S. Saha, S. Azad, and S. A. Fattah, “Detection of sleep apnea using sub-frame based temporal variation of energy in beta band in EEG,” in 2016 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), 2016.
  - [10] S. Taran, V. Bajaj, and D. Sharma, “Robust Hermite decomposition algorithm for classification of sleep apnea EEG signals,” *Electron. Lett.*, vol. 53, no. 17, pp. 1182–1184, 2017.
  - [11] B. L. Koley and D. Dey, “On-line detection of apnea/hypopnea events using SpO<sub>2</sub> signal: a rule-based approach employing binary classifier models,” *IEEE J. Biomed. Health Inform.*, vol. 18, no. 1, pp. 231–239, 2014.
  - [12] P. Dehkordi, A. Garde, W. Karlen, D. Wensley, J. M. Ansermino, and G. A. Dumont, “Pulse rate variability compared with Heart Rate Variability in children with and without sleep disordered breathing,” *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 2013, pp. 6563–6566, 2013.
  - [13] M. F. Aksahin, A. Erdamar, A. Isik, and A. Karaduman, “Sleep apnea detection using with EEG, ECG and respiratory signals,” in 2017 25th Signal Processing and Communications Applications Conference (SIU), 2017.
  - [14] F. Jurysta et al., “A study of the dynamic interactions between sleep EEG and heart rate variability in healthy young men,” *Clin. Neurophysiol.*, vol. 114, no. 11, pp. 2146–2155, 2003.
  - [15] A. H. Khandoker, C. K. Karmakar, and M. Palaniswami, “Interaction between sleep EEG and ECG signals during and after obstructive sleep apnea events with or without arousals,” in 2008 Computers in Cardiology, 2008.
  - [16] Emin Tagluk, M., M. Akin, and N. Sezgin, “Classification of sleep apnea by using wavelet transform and artificial neural networks. Expert Systems with Applications. 2010.
  - [17] K. N. V. P. S. Rajesh, R. Dhuli, and T. S. Kumar, “Obstructive sleep apnea detection using discrete wavelet transform-based statistical features,” *Comput. Biol. Med.*, vol. 130, no. 104199, p. 104199, 2021.
  - [18] R. Lin, R.-G. Lee, C.-L. Tseng, H.-K. Zhou, C.-F. Chao, and J.-A. Jiang, “A new approach for identifying Sleep Apnea Syndrome using wavelet transform and neural networks,” *Biomed. Eng. (Singapore)*, vol. 18, no. 03, pp. 138–143, 2006.
  - [19] X. Zhao et al., “Classification of sleep apnea based on EEG sub-band signal characteristics,” *Sci. Rep.*, vol. 11, no. 1, p. 5824, 2021.
  - [20] M. K. Moridani, M. Heydar, and S. S. Jabbari Behnam, “A reliable algorithm based on combination of EMG, ECG and EEG signals for sleep apnea detection: (A reliable algorithm for sleep apnea detection),” in 2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI), 2019.
  - [21] S. Chattopadhyay, G. Sarkar, and A. Das, “Sleep apnea diagnosis by DWT-based kurtosis, radar and histogram analysis of electrocardiogram,” *IETE J. Res.*, vol. 66, no. 4, pp. 518–526, 2020.
  - [22] S. J. Orfanidis, *Introduction to signal processing: International edition*. Upper Saddle River, NJ: Pearson, 1995.
  - [23] S. Karmakar, S. Chattopadhyay, M. Mitra, and S. Sengupta, *Induction motor fault diagnosis: Approach through current signature analysis*. Singapore, Singapore: Springer, 2018.
  - [24] R. R. Majhi, A. Ghosh, and S. Chattopadhyay, “Analysis of electrocardiogram by radar and DWT based kurtosis comparison,” in Michael Faraday IET International Summit 2015, 2015.
  - [25] R. R. Majhi, S. Chattopadhyay, and A. Ghosh, “Radar Assessment of Wavelet decomposition based Skewness of ECG Signals,” in Michael Faraday IET International Summit 2015, 2015.
  - [26] K. P. Soman, N. G. Resmi, and K. I. Ramachandran, *Insight into wavelets: From theory to practice*. Delhi, India: PHI Learning, 2010.
  - [27] S. G. Mallat, “A theory for multiresolution signal decomposition: The wavelet representation,” in *Fundamental Papers in Wavelet Theory*, Princeton: Princeton University Press, 2009, pp. 494–513.

## Determinants of sustainable frugal innovation in higher education: a massive open online courses perspective.

## Determinants of sustainable frugal innovation in higher education: a massive open online courses perspective.

Dr. Shikha N. Khera<sup>1</sup>, Himanshu Pawar<sup>2</sup>

<sup>1</sup> Designation: Assistant Professor, Affiliation Address: Delhi Technological University, Delhi School of Management, Bawana Road, Shahbad Daulatpur Village, Rohini, New Delhi: 110042, India. Email Id: shikhankhera@yahoo.co.in Alternate Email: shikhankhera18@gmail.com Mobile No.: +91 9810930801

<sup>2</sup> Designation: PhD. Scholar, Affiliation Address: Delhi Technological University, Delhi School of Management Email Id: [himanshuihm1@gmail.com](mailto:himanshuihm1@gmail.com). Mobile No.: +91 7838969754 (author for correspondence)

### ABSTRACT

Massive Open Online Courses (MOOCs) have evolved from open educational resources during the last decade at the right pinnacle of technological advancements and thus, online learning exponentially evolved and spread with the expansion of MOOCs across various streams. We aim to explore the conceptual foundations of sustainable frugal innovation in higher education using MOOCs as a form of frugal product that might help bridge the gap between underprivileged sections of the society and their higher education systems from a developing country perspective. Using a systematic review approach we have analysed definitions pertaining to both the concepts published in peer-reviewed journal articles (n=71) and cross-validated our findings from grass-root frugal innovators and higher education academicians via group interviews. Accessibility, affordability and resource scarcity were found to be the most crucial determinants of sustainable frugal innovation that MOOCs have successfully embraced over the years. Strengthening our case from a developing country perspective our results signify the importance of instituting a frugal approach towards proliferating MOOCs in such systems that either lack quality education or are devoid of resources and leadership necessary to bank upon the underlying power of e-learning.

Keywords: massive open online courses, MOOCs, frugal innovation, higher education, technology, e-learning

### RESUMEN

Los cursos masivos abiertos en línea (MOOC) han evolucionado a partir de recursos educativos abiertos durante la última década en el pináculo de los avances tecnológicos y, por lo tanto, el aprendizaje en línea evolucionó y se extendió exponencialmente con la expansión de los MOOC en diversas corrientes. Nuestro objetivo es explorar los fundamentos conceptuales de la innovación frugal sostenible en la educación superior utilizando los MOOC como una forma de producto frugal que podría ayudar a cerrar la brecha entre los sectores desfavorecidos de la sociedad y sus sistemas de educación superior desde la perspectiva de un país en desarrollo. Utilizando un enfoque de revisión sistemática, analizamos definiciones relacionadas con los

conceptos publicados en artículos de revistas revisadas por pares ( $n = 71$ ) y validamos de forma cruzada nuestros hallazgos de innovadores frugales de base y académicos de educación superior a través de entrevistas grupales. Se descubrió que la accesibilidad, la asequibilidad y la escasez de recursos son los determinantes más cruciales de la innovación frugal sostenible que los MOOC han adoptado con éxito a lo largo de los años. Reforzando nuestro caso desde la perspectiva de un país en desarrollo, nuestros resultados significan la importancia de instituir un enfoque frugal hacia la proliferación de MOOC en sistemas que carecen de educación de calidad o carecen de los recursos y el liderazgo necesarios para aprovechar el poder subyacente del aprendizaje electrónico.

Palabras clave: cursos masivos abiertos en línea, MOOC, innovación frugal, educación superior, tecnología, e-learning

## INTRODUCTION

Innovation in education per se is a holistic concept which can be viewed from multiple perspectives for explaining the radical reforms over the years. What has evolved is not only a shift towards an engaged pedagogy i.e. one which has extensive institutional implications and not confined to changes in classroom dynamics (Saltmarsh et al. 2011), but also the inseparable role of technology in aiding such changes (Garcia et al. 2015). Al-Huneidi and Schreurs (2012) highlighted the prominence of flexible learning environments and collaborative online systems in refurbishing traditional educational ecosystem. These environments would not have existed if it were not for technological innovations to reach their existing forms; tabletPCs, classroom clickers, instant messaging and WebCT etc. (Blasco-Arcas et al. 2013). From the first use of computers in classrooms and universities towards the era of the internet, cloud computing and industry 4.0 technological innovation in education have seen expeditious growth. It wouldn't be wrong to presume that the nature of technological innovation inherits the essence of Kranzberg's second law of technology i.e. 'invention is the mother of necessity' (Kranzberg 1986). There exists a saturation point of every type or form of technology and its use; once reached it acts as a solid foundation for new technology to prosper and grow (Lawton, 2013). The evolution of open educational resources (OERs) to massive open online courses (MOOCs) can be considered as an apt example of how innovation in technology is changing the fabric of higher education. As soon as the availability and accessibility of internet became easy and cheap the OERs automatically evolved, which until the last decade were primarily meant for pre-recorded distance education purposes (Alario-Hoyos et al. 2017). OERs took ample time to advance from their static form to a much more dynamic MOOCs form but, with the pace at which machine learning and artificial intelligence are progressing we might soon be leapfrogging into future classroom transactions with augmented and virtual reality experiences (Leahy et al. 2019). But, before that materializes we must clear the air around the ongoing technological revolution in higher education and understand how we can leverage the underlying power of MOOCs as a type of frugal innovation for such higher education systems which are deprived of quality education. For instance: an innovation might be product innovation, process innovation or disruptive innovation etc. but, only the presence of certain attributes and distinctive features about their nature will help individuals to distinguish between them. We

believe the judicious apprehension of these features is one of the most effective ways to bank upon the underlying power of any type of innovation, which is also the underpinning theme of this article.

Over the last decade, researchers have focused on a new form of innovation (frugal) which we believe might be in sync with certain characteristics of MOOCs. Thus at first, we seek to examine the characteristic features of frugal innovation; which acts as an extended and improvised arm of innovation and has gained exponential momentum in the research domain over the last decade (Pisoni et al. 2018). Traces of this concept in actual practice dates back to ages in ancient civilizations and their philosophies (Tiwari et al. 2017) such as, the 'Greek Epicurean' ethics on fundamentals of living life with frugality and the movement of 'Neo Confucianism' in ancient China which appreciated simplicity and detachment to material self by one of its key proponents Lao-Tzu (Tiwari et al. 2017). However, academic research has just recently started to focus on the intricacies involved in defining frugal innovation and common grounds are being set up to hedge the unpredictable nature of the concept. According to Sharma and Iyer (2012), frugal innovation is a concept that "stems from resource scarcity: utilizing limited resources to meet the needs of low-income customers". Literature is replete with similar definitions which have faced barriers of subjective interpretations of the concept for example in India, the term 'Jugaad innovation' or in China as "Zizhu chuangxin (copycat)" or "jua kali" in Kenya (Radjou 2014) is constantly used in reference to frugal innovation but, the understanding, implementation and execution of the concept might vary across different countries and cultures (Tiwari et al. 2017). Nevertheless, it has been sincerely approached by authors such as Ray and Kanta Ray (2011), Zeschky, Widenmayer and Gassmann (2014) and Prabhu and Jain (2015) etc. in trying to define the boundaries and essential characteristics of such innovation thereby, providing our research with a concrete reference point for studying and understanding the determinants of frugal innovation in higher education sector from a MOOCs perspective.

Secondly, MOOCs on the sidelines of innovation in education technology have emerged as one of the most successful, widespread and sustainable models for the dissipation of knowledge and learning through the use of e-learning platforms (Jordan 2015). It is observed that during the initial years of exploratory research on MOOCs, majority of the researchers divulged more into apprehending the impact, paradox, learning, feasibility, performance evaluation and effectiveness etc. of the concept. Major emphases on learning theories and new conceptual foundations (Gasevic et al. 2014) have rigorously been researched leading to the culmination of key traits and characteristic of MOOCs. Since its inception the concept has been a part of academic dialogue amongst scholars who view it as a form of 'disruptive innovation' (Flynn, 2013, Yuan & Powell 2013). Presumably, the authors believe that MOOCs wield the power for disturbing the make-up of our current educational system by changing the roles of student-teacher interaction and technology (Flynn 2013), which is true if we understand how one complements the other in presence of rapid technological advancements. On the other hand, scepticism looms over the same as few authors believe that the evolution of MOOCs from OERs is nothing more than a technological shift and it doesn't suffice the characteristics of disruptive innovation as mentioned in the literature (Al-Imarah & Shields 2018, Kursun 2016). Perelman (2014) viewed MOOCs as a symptom of disruption, not a major cause since according to him the academic bureaucracy believes that broadcasting online lectures can only put on a masquerade threat to the existing

institutional norms and state of affairs in education; nothing substantial. Thus, due to these differences in opinions it is still an ongoing debate and we leave it to the scholarly minds out in the field to figure out if MOOCs are actually disruptive in nature or not. We however, would like to examine if MOOCs adhere to the principles of frugality (thriftiness/skimping) since, frugal innovation in the education sector is not even remotely studied. We wish to accumulate key characteristics of MOOCs and frugal innovation under one common umbrella and propose to superimpose similarities of both the concepts to form a common ground for mutual co-existence.

Furthermore, we will be drawing and driving our discussions and conclusions from a developing countries' perspective for examining the potential for frugal innovation in MOOCs and vetting them with expert comments over our analysis of key definitions.

## MATERIAL AND METHODS

In order to identify research papers with key definitions on both frugal innovation and MOOCs we have applied a systematic literature review (SLR) approach for determining an inclusion and exclusion criteria for article selection. SLR uses a through methodology to narrow down the scope of research for optimum use, re-use and feasibility (Liyanagunawardena et al. 2013). Textual analysis of the selected definitions for both the concepts was ideally performed by the authors and cross validated by experts in the fields of frugal innovation (primary education) and higher education research to minimize any form of bias arising due to subjective interpretations. We refrained from using automated text mining tools such as R-Studio, Python etc. since we are not working on the identification of key themes or word associations. We wish to make an educated guess at some highly likely explanations of the said text (McKee 2003) for which the feasibility of human interpretation is indispensable. We adhered to the same inclusion and exclusion criteria for searching research articles on frugal innovation and MOOCs.

**Inclusion criteria:** In order to funnel down and select high quality journals, we used the online database SCOPUS® for selection of research articles. Setting the publication language criteria to 'English' and using the keywords "MOOC" OR "Massive Open Online Courses" in 'title' OR 'abstract' OR 'keywords' and selecting only 'research articles' which are 'final published' we extracted a list of top (n=50) cited research papers on massive open online courses. Similar procedure was followed for research articles in frugal innovation (n=50) using the keywords "Frugal innovation" in 'title' OR 'abstract' OR 'keywords'. Therefore a consolidated list of (n=100) articles was prepared for analysis.

However, the initial number for our sample might seem to be arbitrary since, going for a fixed number does not guarantee the results one might require for qualitative analysis (Gergen et al. 2015). Therefore, we took this opportunity to also verify the concept of 'data saturation' (Fusch & Ness 2015) in qualitative analysis for our own research. In a major study conducted by Vasileiou et al. (2018) identified the key reasons behind sample size determination for qualitative research in health sciences and found that 55% of the studies reported determination of 'saturation level' of information as the benchmark for sample size justification. It was only appropriate to look for similarities in definitions up-to a certain point of information redundancy after which no new data could prove to be useful.



Exclusion criteria: Our study has focused only on published peer-reviewed research articles since we aim for high inclusion of quality not quantity. We are currently working only on the definitions of both the concepts thus; other published materials such as case studies, reports, conference proceedings, book review etc. did not fall in the scope for this research. We also withheld ourselves from selecting articles in press.

After reviewing the articles it was found that not all of them focused upon explicit definitions or characteristics of MOOCs thus, we removed those research articles from our list. Data from the sample further started to saturate at the final list of 30 articles for MOOCs and 41 articles for frugal innovation (see Appendix A for sample definitions and references) (n=71).

Elucidating definitions of MOOCs and Frugal Innovation: The definitions (see Appendix A) instantly gave away a quick and general understanding of researchers understanding of both the concepts. As, they expounded about MOOCs, it was found that the majority of them had a notion of MOOC as a free course whilst being open in nature. They are accessible from any part of the world to anyone who wishes to enrol and learn. The definitions consistently feature two major technological pre-requisites i.e. the presence of a digital device such as a laptop, computer or a mobile phone which can support MOOCs platforms and a good enough internet connection. Massive influxes of students attracted by top tier universities have laid the foundations for these courses thus, for an online course to be called a MOOC huge number of student enrolment is an important factor. Some of the authors believe, that in the right philanthropic mindset MOOCs have been efficient in removing the financial barriers for students coming from both developing and under-developed countries; allowing them to access high quality learning resources which otherwise would have been limited for them. Hence, they are also considered as a gateway for unlimited learning opportunities for students across all spectrums of socio-economic structures.

But, for MOOCs to be scrutinized as a form of frugal innovation they must hold true to the primary determinants of sustainable frugality such as affordability, accessibility and resource scarcity which, are the three most crucial aspects of frugal innovation as highlighted in the definitions. Since, the name speaks for itself resource constrained environments are at the core of defining frugal innovation thus, minimal use of resources is not a choice but a matter of human ingenuity and adaptability in problem solving using given resources at hand. Use of technology in such a way that minimizes not only the manufacturing cost but also the accessibility costs for destitute sections of the society is of major significance for every innovation to be considered frugal. It is interesting to note that majority of the frugal innovations are product centric i.e. the authors have till date focused only on those innovations which possess physical characteristics and their services as frugal innovation, for example Wonderbag (South Africa), ChotuKool (India), Aakash Tablet (India), BYD Lithium-ion batteries (China) etc. (Nevejan, 2016). Other definitions highlight the importance of a frugal mindset which in layman terms could be interpreted as, 'an ability to work out of line with creativity at its behest'.

Understanding sustainable frugality in the context of its determinants (resource scarcity, affordability and accessibility) from a developing country (Indian) perspective: In a research conducted by Shah and Santandreu Calonge (2017) an attempt has been made to address the frugal power of MOOCs and how they could be utilized to access millions of displaced refugees in war torn countries of the middle east. They

developed a 'frugal MOOCs' model for school going Syrian refugees which addresses real issues of learners' needs, local stakeholders and technological challenges. It is quite evident from their model that without an established infrastructure (mobile and internet) and the help of local stakeholders in customizing learners' education needs the frugal MOOCs model will not suffice the end goal of making quality education accessible to underprivileged sections of the society. This is the power of frugality in innovation when resource scarce educational environments could access free and high quality learning materials from some of the top universities in the world. It is also worthwhile to note that majority of the frugal innovations have come from developing or under-developed nations since they share similar technological, economic and leadership challenges. Therefore, we now look at frugal innovation from an Indian perspective because the country has always been at the forefront of frugal or 'jugaad innovation' in the world (Radjou 2014) and dive deeper into the conceptual confluence of MOOCs as a form of frugal innovation.

Resource scarcity in frugal innovation is generally debated in the context of people living under the bottom of pyramid (BoP) (Pansera et al. 2016) thus, acting as the driving force for some form of frugal innovation to happen but, when we discuss about resource scarcity in the context of higher education and particularly MOOCs, it could not be denied that there is a huge chunk of students in developing countries who still do not have ways to leapfrog technological barriers (Davison et al. 2000). Further, educational institutions and teachers in these countries might not have an idea about the power of MOOCs in enhancing their academic acumen. Lack of awareness and alignment of learners' digital literacy, background and culture with content and medium of instruction is a major hindrance for effective dissemination of MOOCs in these areas. Thus, at first the environment needs to be conducive enough to support MOOCs as a form of frugal innovation in higher education. For example in India, the government is on a mission to expand the reach of internet services to the marginalized sections of the society which will not only help in financial inclusion for government schemes but could also be used to connect with educational institutions on both national and international levels. The same platform could be used by higher education institutions (HEIs) lacking quality education to aid their curriculum with learning materials and instructional teaching available at both government and private funded MOOCs platforms such as SWAYAM®, IITBx®, mooKIT® etc. But, the integration will only work if the educational institutions have the right intent to embed MOOCs into their educational ecosystem. It might be true to say that online educational resources (OERs) re-invented themselves in the face of MOOCs over technological advancements and MOOCs now have the power to serve as a frugal solution to resource scarce educational environments where students don't have access to quality educational systems.

Affordability constraint is another factor that hinders the access of quality higher education services to students living in rural areas or tier 2 and 3 cities. Even in the current age of digitization and internet 4.0 it is absolutely not necessary in the developing and under-developed countries that underprivileged students could even afford the basic fee for MOOCs available on for-profit online platforms such as Udemy®, Edx® or Coursera®. As lucrative as they might sound but paying for such courses might not resonate with their actual needs. Thus, MOOCs which are entirely free of cost have massive potential to bridge this gap in the same way Indian car manufacturer TATA® with their Nano® car did as one of the most successful frugal innovations in the

automobile sector by bringing the luxury of owning a car to the common man (Rao 2013). To put the matter into an Indian perspective the exponential growth of a government run platform SWAYAM® (self-induced) is currently being used by students absolutely free of cost. All the courses are freely accessible which has led to the huge number of student enrolments across multiple domains from all parts of the country. Since the courses are made by lecturers from the top institutions in the country the appeal is much stronger and encouraging. Thus, the second step for banking upon the power of frugality is to make sure affordability is not a constraint on any level for any student.

Accessibility constraint is one that is based under the context of resource scarcity. According to (Horn et al. 2013) for an innovation to be considered as frugal, it has to be accessible by the masses and not restricted to a particular niche of the society. A student with a good enough internet connection can access MOOCs from any part of the world. But, good enough is a relative term and could be probed under the aegis of resource scarce environments where access to high speed internet might not be that easy. The video sessions for MOOCs which are broadcasted live require high speed internet connection and according to Roser, Ritchie and Ortiz-Ospina (2020) access to internet services in the world is still skewed on one side of the scale. In the Indian scenario the government is spending huge amounts of money via local schemes such as 'BharatNet' to make sure internet access is easy for villages and remote areas ("Vikaspedia Domains" 2016). It is an attempt to connect local villages with government schemes aimed at improving e-governance, e-banking, e-health and e-education. Therefore, improving accessibility to internet services is equally important for frugal innovations to thrive in the backdrop of establishing frugal MOOCs.

Finally, the relationship between innovation, frugal innovation and MOOCs could be summed up in the form of similarities depicted in the three concepts (see Figure 1).

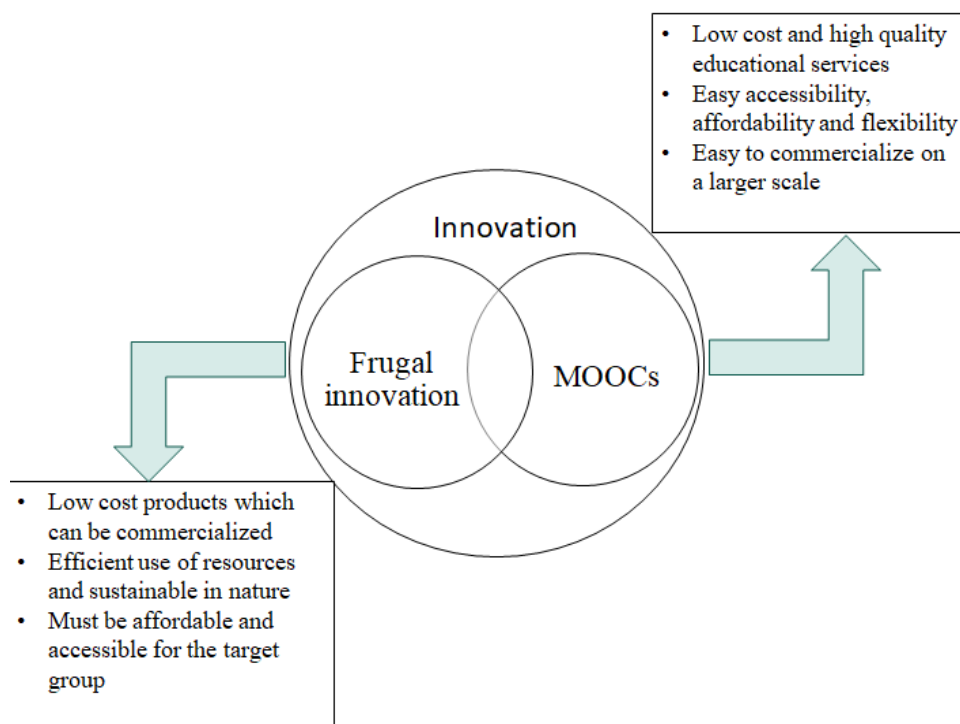


Figure 1. Relationship between innovation, MOOCs and frugal innovation

## RESULTS AND DISCUSSION

Once the review of definitions was done, it was necessary to vet our findings from users/experts in both the fields and understand how one can optimally make use of the frugal characteristics of MOOCs. We conducted semi-structured interviews with a group of lecturers from a technical university where a choice based credit system for MOOCs is embedded in undergraduate (UG) degree programme curriculum. A focused group interview was also conducted with grass root frugal innovators (teachers) in primary education in rural areas why? Because, these teachers work in some of the most remote areas of the country and face numerous challenges due to lack of resources (capital, labour and technological) pushing them to perform acts of frugality in their everyday lives. Some of the excerpts from the commentaries of various experts in both fields are mentioned below:

Opinion of Primary school teachers: “We have to perform ‘jugaad’ in our day to day lives due to resource constraints from the government. To run an institution various resources are needed but we have to eventually manage with little at hand (see figure 2). For example, we are teaching students to become self-sufficient and embed values of sustainability in them. These children come from the marginalized sections of the society and we teach them how to improvise for daily challenges. We don’t have text books to be distributed to all the students for primary education so; we have devised a method of teaching basic numeracy skills by changing our teaching pedagogy and all inclusive participatory learning etc.”

In developing countries such as Guatemala, Philippines, Bolivia etc. access to internet is still considered as a luxury in rural areas (Istance et al. 2019). In rural India there is only (21.2%) access to computers in primary and secondary (ASER Centre 2018) schools. Reports have shown that almost half of primary school (5th grade) students can’t read or write properly in India and one of the worlds’ biggest educational systems is facing a learning crisis (ASER Centre 2018). During our visit, it was not startling to see that there was a lack of resources in rural areas, but what were more important to note from our experience was teachers’ willingness and a pro-active approach to sensitize the stalled education system with change. Our interviewees considered themselves duty bound ethically and morally to teach and up-skill the students by simple acts of frugality. Therefore, for such education systems where the government leadership and policy making is consuming much more time than needed teachers could use OERs and MOOCs to suffice the immediate needs of the students. Further, we also introduced the teachers to online learning platforms (see figure 3) Khan Academy® (US) and Byjus® (India) and asked them if they can supplement their teaching with high quality learning material from these platforms for faster, efficient and up-to-date growth of the children. However, at the current stage, it is hard to measure how efficient these interventions would prove to be in the long run since lack of awareness and support from competent authorities might dilute their motivation to appreciate the power of frugality underlying OERs and MOOCs.

Opinions of academicians on frugal nature of MOOCs in higher education: “[...] I do acknowledge the presence of an inherent power of frugality in MOOCs but there is major lack of awareness amongst academicians in higher education regarding its feasibility and effectiveness in our country which is clouded by the rudimentary ideologies of higher authorities”



Figure 2. Students engaged via participatory learning and role play as street vendors during authors' visit at a primary school (Nhu district, Haryana, India, February 10, 2022).



Figure 3. Teacher acquainting students with online learning platforms for grades 1-5 via Khan Academy® and Byjus® apps on mobile phone during authors' visit at a primary school (Nhu district, Haryana, India, February 10, 2022).

"[...] By virtue of definition MOOCs might be called as frugal innovation since there are institutions which lack resources in our country. MOOCs can aid these institutions in providing quality higher education anytime at their disposal"

“[...] As, understood from the definitions, frugality is a mindset which means doing something more efficiently with limited resources and constraints. There is a major issue of skill-gap in our country and quality education is lacking in major domains thus, the students who cannot access or afford quality higher education can make use of the MOOCs model to up-skill themselves but proper guidance is a must”

It was evident from our discussions that the same awareness gap and lack of leadership that drags the growth of students in primary education lingers on in higher education as well. The power had predominantly been dormant in nature due to lack of attention and trust in MOOCs models for disseminating quality education. Hence, a niche of students' accessing MOOCs is rapidly evolving majorly in tier-1 institutions and cities and not across the rest of the country. In words of Stephen Downes one of the co-founders of MOOCs, sharp criticism of the rapidly evolving MOOCs system as a for-profit business model could be heard in an interview (Downes, 2012) where he explicitly said:

“ [...] I don't see how you can call something open and charge money for it, I am sorry those two concepts to me just don't down go together in the same sentence”

It is in nature of every system to evolve and mould itself according to the decisions taken by its key players. In the case of MOOCs the platforms such as Coursera®, Edx® and Udemy® etc. are charging a fee for earning certification of a course but, what value are these if a particular student segment can't afford them? Should we not debate about the acceptability of these certificates in various job markets? Why only the students from top institutions and tier-1 cities in the country are accessing MOOCs rapidly? The majority of the courses on these platforms have options for a paid certificate and the misconception around the word 'open' in MOOCs is now beginning to clear. Thus, the marketing and selling of 'education' as an online product is beginning to penetrate the upper layers of MOOCs. In coming years it would not be dramatic to view these online courses and certifications “on happy hour” sales or “1+1” offers. We are not counter arguing the business models of these platforms and MOOCs are definitely accessible to anyone with an internet connection but, we argue that the real value will not trickle down on its own until and unless students are guided by teachers and their institutions are financially aided by the government in such countries. Thus, Institutional and governmental interventions are a must for MOOCs to co-exist between all divisions of a society in an unbiased manner.

HEIs are not devoid of resources needed to exploit the potential of MOOCs but in order to maximize efficiency they must play an active role in developing networks with partnering institutions, prospective employers and the government. It is necessary for institutions to develop policies that communicate the benefits of MOOCs in a way, that doesn't disrupt or undermine the current educational systems in place. It is important to understand the needs of the market not only on a national level but also on a global level for appropriate student guidance and support. It is the right time to bank upon the frugal power of MOOCs i.e. easy accessibility and affordability for supplementing educational environments with high quality e-learning certifications and courses via a connectivist mode of learning. It will empower institutions to get connected with the national and global education systems which have progressed substantially in proliferating MOOCs on various platforms. Thus, it is altogether more crucial for underdeveloped higher education systems to embrace the MOOCs model with a frugal mindset. For example, the higher education systems across the globe have



recently shown exemplary behaviour in the darkness of the ongoing pandemic COVID-19. Around the globe multiple HEIs have moved towards the use of MOOCs and online education platforms to aid their stalled educational systems (Mineo 2020). Since free e-content knows no boundaries, voices from all education systems are being heard across top global universities which have opened access to free learning for students across the globe. During these crucial times live online learning has emerged as a potent tool to tackle problems of disseminating knowledge and learning activities (Burgess et al. 2020). Thus, the situation has been a blessing in disguise for all educational systems that were not globally connected and lacked sophisticated tools and technologies by pushing them to become more frugal in using e-resources for accessing quality higher education.

### CONCLUSION

On a very primary level of elucidation after thorough textual analysis we have observed that ‘any form of innovation be it new or induced after changes in the existing structure of products or services for the better good of masses be them poor or rich can be defined as frugal innovation’. MOOCs don’t fail to identify themselves as a form of frugal innovation on tracks of low cost educational services targeted at students who either have marginal access to study resources or limited affordability to quality higher education. The only promising way of realising the hidden potential of MOOCs is by unleashing the power of frugality which prerequisites a certain degree of philanthropic and visionary mindset on part of partnering HEIs, MOOCs offering platforms and the government. In context of developing countries primary issues such as, lack of awareness amongst academia, over-reliance on orthodox teaching pedagogies and stagnant curriculum across majority of HEIs needs to be revamped first by corroborative efforts of top institutions and the government. Only with efficient policy interventions, the issues of access to quality higher education and reduction of skill gaps arising due to lack of knowledge could be addressed with the help of MOOCs. Since, MOOCs are low cost educational services we vouch for government and HEIs support in aiding students deprived of quality education primarily due to financial constraints. Whilst looking at the higher education systems at large, the need for private players offering online platforms for the culmination of e-resources should not be sidelined. Our study is not against the commercialization of education; that has already happened long ago and will continue to flourish with changes and advancements in technologies. But, we aim to spread a message for an integrated approach which reduces the burden on HEIs for churning out individuals who are highly skilled, self sufficient and job ready for a disruptive global context.

Limitations and future research: Even though with all of our best knowledge and experience put to test, we believe our study might be limited by the subjective interpretations of select definitions. Additionally, we have selected limited research articles for the review and we might have had missed out on a few good papers. Techniques such as text mining via tools such as R-Studio® or Python® could also be used to analyse a greater number of papers depending upon the feasibility of the study. Since, frugal innovation is not appropriately researched in the education sector we leave it up to the research community to corroborate our findings and look for patterns of frugality in the education sector.

As of now we have highlighted the implicit determinants of frugal innovation and their relation with MOOCs but, we would also like to propose a conceptual model which is in its testing phase. Explanation of the model is not within the scope of the present study rather a brief overview is provided (see figure 4). We are primarily concerned about the efficient integration and acceptance of MOOCs into the HE systems of developing and underdeveloped countries which are plagued with several institutional and human-induced biases. Literature apprises that the three independent variables represent the basic nature of frugal innovation and we believe they might have a direct and substantial bearing on effective integration of MOOCs into such higher education systems. To actually benefit from the frugal power of MOOCs one must delve deeper into the reasons and degree of affect the aforementioned constraints have on such form of unification such that, effective policy making and guided decisions could be made.

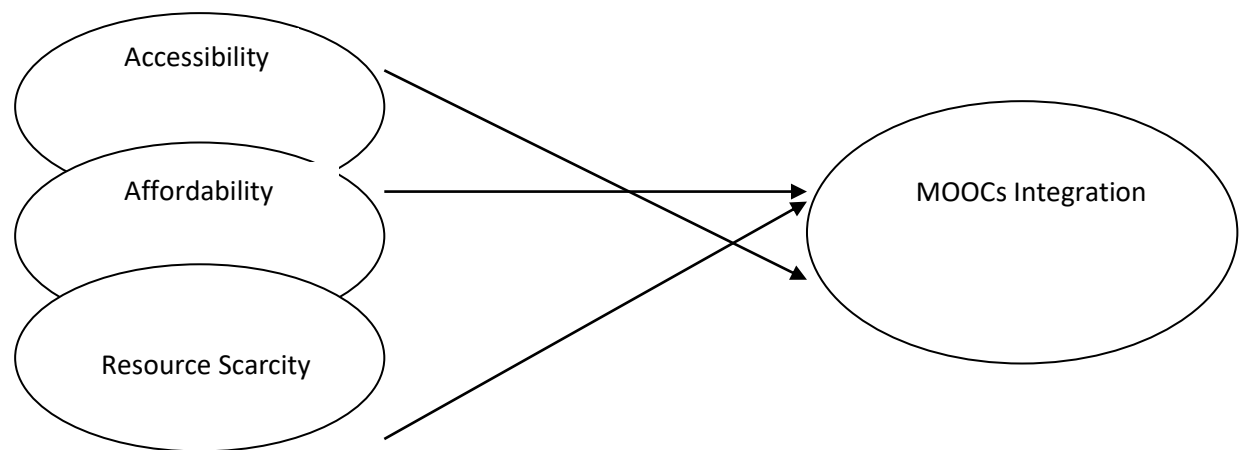


Figure 4: Conceptual framework for frugal MOOCs

## REFERENCES

- Agnihotri, A. 2014. Low-cost innovation in emerging markets. *Journal Of Strategic Marketing*, 23(5), 399-411. doi: 10.1080/0965254x.2014.970215
- Alario-Hoyos, C., Estévez-Ayres, I., Pérez-Sanagustín, M., Delgado Kloos, C., and Fernández-Panadero, C. 2017. Understanding Learners' Motivation and Learning Strategies in MOOCs. *The International Review of Research In Open And Distributed Learning*, 18(3). doi: 10.19173/irrodl.v18i3.2996
- Al-Huneidi, A., and Schreurs, J. 2012. Constructivism Based Blended Learning in Higher Education. *International Journal Of Emerging Technologies In Learning (Ijet)*, 7(1).
- Al-Imarah, A., and Shields, R. 2018. MOOCs, disruptive innovation and the future of higher education: A conceptual analysis. *Innovations In Education And Teaching International*, 56(3), 258-269. doi: 10.1080/14703297.2018.1443828
- Alraimi, K., Zo, H., and Ciganek, A. 2015. Understanding the MOOCs continuance: The role of openness and reputation. *Computers & Education*, 80, 28-38.

- ASER Centre. 2018. Annual Status of Education Report (Rural) 2018 (Provisional). ASER Centre. Retrieved from <http://img.asercentre.org/docs/ASER%202018/Release%20Material/aserreport2018.pdf>
- Baggaley, J. 2013. MOOC rampant. *Distance Education*, 34(3), 368-378.
- Barak, M., Watted, A., and Haick, H. 2016. Motivation to learn in massive open online courses: Examining aspects of language and social engagement. *Computers & Education*, 94, 49-60. doi: 10.1016/j.compedu.2015.11.010
- Blasco-Arcas, L., Buil, I., Hernández-Ortega, B., and Sese, F. 2013. Using clickers in class. The role of interactivity, active collaborative learning and engagement in learning performance. *Computers & Education*, 62, 102-110.
- Brem, A., and Wolfram, P. 2014. Research and development from the bottom up - introduction of terminologies for new product development in emerging markets. *Journal Of Innovation And Entrepreneurship*, 3(1), 9.
- Burgess, S., and Sievertsen, H. 2020. The impact of COVID-19 on education | VOX, CEPR Policy Portal. Retrieved 15 April 2020, from <https://voxeu.org/article/impact-covid-19-education>.
- Davison, R., Vogel, D., Harris, R., and Jones, N. 2000. Technology Leapfrogging in Developing Countries - An Inevitable Luxury?. *The Electronic Journal of Information Systems In Developing Countries*, 1(1), 1-10.
- DeBoer, J., Ho, A., Stump, G., and Breslow, L. 2014. Changing "Course": Reconceptualizing Educational Variables for Massive Open Online Courses. *Educational Researcher*, 43(2), 74-84.
- DeWaard, I., Abajian, S., Gallagher, M., Hogue, R., Keskin, N., Koutropoulos, A., and Rodriguez, O. 2011. Using mLearning and MOOCs to understand chaos, emergence, and complexity in education. *The International Review of Research In Open And Distributed Learning*, 12(7), 94. doi: 10.19173/irrodl.v12i7.1046
- Downes, S. 2012. #FUSION12 - Discussion about MOOCs with Stephen Downes [Hangout]. San Diego.
- Ebben, M., and Murphy, J. 2014. Unpacking MOOC scholarly discourse: a review of nascent MOOC scholarship. *Learning, Media And Technology*, 39(3), 328-345.
- Flynn, J. 2013. Moocs: Disruptive Innovation and the Future of Higher Education. *Christian Education Journal: Research on Educational Ministry*, 10(1), 149-162.
- Fusch, P. I., and Ness, L. R. 2015. Are We There Yet? Data Saturation in Qualitative Research. *The Qualitative Report*, 20(9), 1408-1416.
- Garcia, E., Elbeltagi, I., Brown, M., and Dungay, K. 2015. The implications of a connectivist learning blog model and the changing role of teaching and learning. *British Journal of Educational Technology*, 46(4), 877-894. doi: 10.1111/bjet.12184
- Gasevic, D., Kovanovic, V., Joksimovic, S., and Siemens, G. 2014. Where Is Research on Massive Open Online Courses Headed? A Data Analysis of the MOOC Research Initiative. *The International Review of Research in Open and Distributed Learning*, 15, 134-176.
- Gergen, K., Josselson, R., and Freeman, M. 2015. The promises of qualitative inquiry. *American Psychologist*, 70(1), 1-9.

- Gillani, N., and Eynon, R. 2014. Communication patterns in massively open online courses. *The Internet And Higher Education*, 23, 18-26.
- Glance, D., Forsey, M., and Riley, M. 2013. The pedagogical foundations of massive open online courses. *First Monday*, 18(5). doi: 10.5210/fm.v18i5.4350
- Gupta, A., Dey, A., Shinde, C., Mahanta, H., Patel, C., Patel, R. Sahay, N., Sahu, B., Vivekanandan, P., Verma, S., Ganesham, P., Kumar, V., Kumar, V., Patel, M., and Tole, P. 2016. Theory of open inclusive innovation for reciprocal, responsive and respectful outcomes: coping creatively with climatic and institutional risks. *Journal Of Open Innovation: Technology, Market, And Complexity*, 2(1).
- Hone, K., and El Said, G. 2016. Exploring the factors affecting MOOC retention: A survey study. *Computers & Education*, 98, 157-168.
- Horn, C., and Brem, A. 2013. Strategic directions on innovation management – a conceptual framework. *Management Research Review*, 36(10), 939-954.
- Hossain, M., Simula, H., and Halme, M. 2016. Can Frugal Go Global? Diffusion Patterns of Frugal Innovations. *Technology In Society*, 46, 132-139.
- Istance, D., and Paniagua, A. 2019. Learning to Leapfrog: Innovative Pedagogies to Transform Education. Brookings. Retrieved from [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=EDU/WKP\(2018\)8&docLanguage=En](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=EDU/WKP(2018)8&docLanguage=En)
- Jordan, K. 2014. Initial trends in enrolment and completion of massive open online courses. *The International Review Of Research In Open And Distributed Learning*, 15(1).
- Jordan, K. 2015. Massive open online course completion rates revisited: Assessment, length and attrition. *International Review Of Research In Open And Distributed Learning*, 16(3). doi: 10.19173/irrodl.v16i3.2112
- Kaplan, A., and Haenlein, M. 2016. Higher education and the digital revolution: About MOOCs, SPOCs, social media, and the Cookie Monster. *Business Horizons*, 59(4), 441-450. doi: 10.1016/j.bushor.2016.03.008
- Kay, J., Reimann, P., Diebold, E., and Kummerfeld, B. 2013. MOOCs: So Many Learners, So Much Potential. *IEEE Intelligent Systems*, 28(3), 70-77. doi: 10.1109/mis.2013.66
- Khan, R. 2016. How Frugal Innovation Promotes Social Sustainability. *Sustainability*, 8(10), 1034. doi: 10.3390/su8101034
- Kranzberg, M. 1986. Technology and History: "Kranzberg's Laws". *Technology and Culture*, 27(3), 544-560. doi:10.2307/3105385
- Kursun, E. 2016. Does Formal Credit Work for MOOC-Like Learning Environments?. *The International Review of Research In Open And Distributed Learning*, 17(3).
- Lawton, B. 2013. The Characteristics of Technology. *The International Journal for the History of Engineering & Technology*, 79(1), 91-112.
- Leahy, S., Holland, C., & Ward, F. 2019. The digital frontier: Envisioning future technologies impact on the classroom. *Futures*, 113, 102422. doi: 10.1016/j.futures.2019.04.009

- Levvnen, J., Hossain, M., Lyytinen, T., Hyvvrinen, A., Numminen, S., and Halme, M. 2016. Implications of Frugal Innovations on Sustainable Development: Evaluating Water and Energy Innovations. *Sustainability*, 8(4). doi: <https://doi.org/10.3390/su8010004>
- Lim, C., Han, S., and Ito, H. 2013. Capability building through innovation for unserved lower end mega markets. *Technovation*, 33(12), 391-404.
- Littlejohn, A., Hood, N., Milligan, C., and Mustain, P. 2016. Learning in MOOCs: Motivations and self-regulated learning in MOOCs. *The Internet And Higher Education*, 29, 40-48. doi: 10.1016/j.iheduc.2015.12.003
- Liyanagunawardena, T., Adams, A., and Williams, S. 2013. MOOCs: A systematic study of the published literature 2008-2012. *The International Review of Research In Open And Distributed Learning*, 14(3), 202. doi: 10.19173/irrodl.v14i3.1455
- Martin, F. 2012. Will massive open online courses change how we teach?. *Communications Of The ACM*, 55(8), 26-28. doi: 10.1145/2240236.2240246
- McKee, A. 2003. *Textual Analysis: A Beginner's Guide* (1st ed.). London: Sage Publications.
- Mineo, L. 2020. The pandemic's impact on education. Retrieved 15 April 2020, from <https://news.harvard.edu/gazette/story/2020/04/the-pandemics-impact-on-education/>
- Nevejan, C. 2016. Frugal Innovations Around the World. Retrieved 20 April 2020, from <https://tudelft.openresearch.net/page/15976/frugal-innovations-around-the-world>
- Pansera, M., and Sarkar, S. 2016. Crafting Sustainable Development Solutions: Frugal Innovations of Grassroots Entrepreneurs. *Sustainability*, 8(1), 51.
- Perelman, L. 2014. MOOCs: Symptom, not cause of disruption: MOOCs and technology to advance learning and learning research (ubiquity symposium). *Ubiquity*, 1-15.
- Pisoni, A., Michelini, L., and Martignoni, G. 2018. Frugal Approach to Innovation: State of the Art and Future Perspectives. *Journal of Cleaner Production*, 171(10): 107–126.
- Prabhu, J., and Jain, S. 2015. Innovation and entrepreneurship in India: Understanding jugaad. *Asia Pacific Journal of Management*, 32(4), 843-868.
- Radjou, N. 2014. Frugal innovation: a pioneering strategy from the South. Retrieved 24 December 2019, from <http://regardssurlaterre.com/en/frugal-innovation-pioneering-strategy-south>
- Rao, B. 2013. How disruptive is frugal?. *Technology In Society*, 35(1), 65-73.
- Ray, S., and Kanta Ray, P. 2011. Product innovation for the people's car in an emerging economy. *Technovation*, 31(5-6), 216-227. doi: 10.1016/j.technovation.2011.01.004
- Rosca, E., Arnold, M., and Bendul, J. 2017. Business models for sustainable innovation – an empirical analysis of frugal products and services. *Journal Of Cleaner Production*, 162, S133-S145. doi: 10.1016/j.jclepro.2016.02.050
- Roser, M., Ritchie, H., Ortiz-Ospina, E. 2020. Internet. Retrieved 20 April 2020, from <https://ourworldindata.org/internet>
- Saltmarsh, J., and Zlotkowski, E. 2011. *Higher education and democracy*. Philadelphia: Temple University Press.
- Shah, M., and Santandreu Calonge, D. 2017. Frugal MOOCs. *The International Review Of Research In Open And Distributed Learning*, 20(5). doi: 10.19173/irrodl.v20i4.3350

- Sharma, A., and Iyer, G.R., 2012. Resource-constrained product development: implications for green marketing and green supply chains. *Industrial Marketing Management*. 41 (4), 599-608.
- Soni, P., T. Krishnan, R. 2014. Frugal innovation: aligning theory, practice, and public policy. *Journal Of Indian Business Research*, 6(1), 29-47.
- Tiwari, R., Fischer, L., and Kalogerakis, K. 2017. Frugal Innovation: An Assessment of Scholarly Discourse, Trends and Potential Societal Implications. In C. Herstatt & R. Tiwari, *Lead Market India: Key Elements and Corporate Perspectives for Frugal Innovations* (1st ed.). Springer International Publishing.
- Tiwari, R., and Herstatt, C. 2012. Assessing India's lead market potential for cost-effective innovations. *Journal Of Indian Business Research*, 4(2), 97-115.
- Urlich, N. 2017. The four most significant shifts in modern pedagogy. Retrieved 28 December 2019, from <http://blog.core-ed.org/blog/2017/12/the-four-most-significant-shifts-in-modern-pedagogy.html>
- Vasileiou, K., Barnett, J., Thorpe, S., and Young, T. 2018. Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. *BMC Medical Research Methodology*, 18(1).
- Vikaspedia Domains. 2016. Retrieved 20 April 2020, from <https://vikaspedia.in/e-governance/digital-india/national-optical-fibre-network-nofn>
- Winterhalter, S., Zeschky, M., Neumann, L., and Gassmann, O. 2017. Business Models for Frugal Innovation in Emerging Markets: The Case of the Medical Device and Laboratory Equipment Industry. *Technovation*, 66-67, 3-13.
- Wu, B., and Chen, X. 2017. Continuance intention to use MOOCs: Integrating the technology acceptance model (TAM) and task technology fit (TTF) model. *Computers In Human Behavior*, 67, 221-232. doi: 10.1016/j.chb.2016.10.028
- Yuan, L., and Powell, S. 2013. MOOCs and disruptive innovation: Implications for higher education. *Elearning Papers*. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.422.5536&rep=rep1&type=pdf>
- Zeschky, M., Widenmayer, B., and Gassmann, O. 2014. Organising for reverse innovation in Western MNCs: the role of frugal product innovation capabilities. *International Journal of Technology Management*, 64(2/3/4), 255. doi: 10.1504/ijtm.2014.059948

Received: 02<sup>th</sup> April 2023; Accepted: 20<sup>th</sup> May 2023; First reception: 19<sup>th</sup> September, 2023.



## Appendix A

Sample definitions from papers reviewed on Frugal Innovation

Authors	Definition or features (in-text)	Article Title	Year	Cited by*
(Zeschky, Widenmayer & Gassmann, 2014)	....In contrast to good-enough innovations, frugal innovations are not re-engineered solutions but products or services developed for very specific applications in resource constrained environments.	From cost to frugal and reverse innovation: Mapping the field and implications for global competitiveness	2014	87
(Ray & Kanta Ray, 2011)	....a deliberate and singular focus on frugal use of technology and resources is required for crafting a disruptive technology that provides basic functionalities at a very low price.	Product innovation for the peoples car in an emerging economy	2011	83
(Sharma & Iyer, 2012)	... frugal-innovations possessing a no frills structure have been developed for the thrifty consumer under constraints of developing countries.  ... The adoption of frugality entails design principles that advocate minimal use of resources for realizing efficient functioning of products.	Resource-constrained product development: Implications for green marketing and green supply chains	2012	78
(Prabhu & Jain, 2015)	....Frugality refers to the ingenious use of limited resources at hand.... Flexibility alludes to the ability to rapidly adapt and improvise to changing circumstances. And finally inclusivity involves developing goods and services for individuals and communities who are significantly constrained in their capacity to pay and are often marginal participants in the market-based economy.	Innovation and entrepreneurship in India: Understanding jugaad	2015	57
(Brem & Wolfram, 2014)	....For frugal innovation, the BoP is seen as a potential market where sales might be gained and new competition arises. The frugal innovation approach is predominantly product-oriented to cut costs for materials or processes through frugality and simplicity that includes, partially, an ecological idea	Research and development from the bottom up - introduction of terminologies for new product development in emerging markets	2014	54
(Rosca, Arnold & Bendul, 2017)	....As such, frugal innovations do not only involve new technologies, but also innovative ways of altering traditional value creation and capture mechanisms through value chain elements reconfiguration, business models reshaping, re-engineered products and services, inclusion of poor into the economic markets and extreme focus on affordability constraints.	Business models for sustainable innovation – an empirical analysis of frugal products and services	2017	44
(Tiwari & Herstatt, 2012)	....frugal product innovations, as shown by the examples above, may require complex and concerted research & development (R&D) efforts to design an easy-to-use, low-cost solution to a complex problem.	Assessing India's lead market potential for cost-effective innovations	2012	44
(Agnihotri, 2014)	....Frugal innovation refers to those innovative products and services which are developed under conditions of resource constraints.	Low-cost innovation in emerging markets	2015	38
(Pansera & Sarkar, 2016)	.... “frugal innovation”, i.e., the search for simple but effective solutions to deliver affordable products/services.	Crafting sustainable development solutions: Frugal innovations of grassroots entrepreneurs	2016	36
(Soni & T. Krishnan, 2014)	....The process through which this is done is often referred to as “frugal engineering”, and the outcome, which are generally low-cost, good-enough products or services, are known as “frugal innovations”.	Frugal innovation: Aligning theory, practice, and public policy	2014	36

(Lim, Han & Ito, 2013)	....Here, the product innovation for the ULM can be considered as 'frugal' or 'Ghandian innovation', in that the product has to bear resource-saving product for low income consumers.	Capability building through innovation for unserved lower end mega markets	2013	31
(Hossain, Simula & Halme, 2016)	....that frugal innovation refers to products, services or combination of them that are affordable, sustainable, easy-to-use, and have been innovated under the resource scarcity.	Can frugal go global? Diffusion patterns of frugal innovations	2016	27
(Horn & Brem, 2013)	....The concept of frugal innovation aims at modifying and adopting products to foreign, emerging markets on the one hand, and the establishment of R&D capacity and product development centers on the other hand.....  ....Frugality postulates a concept of products being easier to produce and be more adapted to the use of consumers in emerging economies.	Strategic directions on innovation management - a conceptual framework	2013	27
(Levvnen et al., 2016)	....It refers to solutions created under the circumstances of resource constraints.	Implications of frugal innovations on sustainable development: Evaluating water and energy innovations	2016	23
(Khan, 2016)	.....Frugal innovation is developed in severe resource constraints; it involves good quality and reasonably priced products or services even for the customers with modest lifestyles.  ....Generally, frugal innovation is viewed as low cost innovation but it is much more than that. Frugal innovation uses the concept of simplification and strives for less instead of more by using clever technology.	How frugal innovation promotes social sustainability	2016	22
(Winterhalter, Zeschky, Neumann & Gassmann, 2017)	....frugal mindset represents the creation of very high customer value at very low costs for resource-constrained people in emerging markets.	Business Models for Frugal Innovation in Emerging Markets: The Case of the Medical Device and Laboratory Equipment Industry	2017	19
(Gupta et al., 2016)	...The frugality (or low-cost, affordable nature of innovations) emerged as an inalienable feature of grassroots innovations.  ...frugality must blend affordability with circularity (the ability of waste being repurposed, recycled or incorporated in different value chains without affecting the environment adversely).	Theory of open inclusive innovation for reciprocal, responsive and respectful outcomes: Coping creatively with climatic and institutional risks	2016	17

\*At the time of extraction from SCOPUS® database

Sample definitions from papers reviewed on MOOCs

Authors	Definition or features (in-text)	Article Title	Year	Cited by*
(Liyanagunawardena, Adams & Williams, 2012)	....Connectivity through freely accessible online resources	MOOCs: A systematic study of the published literature 2008-2012	2013	485
(Jordan, 2014)	....Free courses from a range of elite universities	Initial trends in enrolment and completion of massive open online courses	2014	370
(Martin, 2012)	....No fees for courses, large scale applicability	Education will massive open online courses change how we teach	2012	188
(Alraimi, Zo & Ciganeck, 2015)	....Free online classes open to all	Understanding the MOOCs continuance: The role of openness and reputation	2015	165
(Littlejohn, Hood, Milligan & Mustain, 2016)	....MOOCs emphasise their openness and scale, which allow learners, regardless of location or previous experience and qualification, to engage at no (or minimal) cost in learning opportunities, which often are curated by leading universities.	Learning in MOOCs: Motivations and self-regulated learning in MOOCs	2016	147
(Hone & El Said, 2016)	....Massive Open Online Courses (MOOCs) are a rapidly growing mode of educational provision, holding the potential to open up access to world class teaching and educational resources beyond geographical and social boundaries.	Exploring the factors affecting MOOC retention: A survey study	2016	146
(DeBoer, Ho, Stump & Breslow, 2014)	....MOOCs are online learning environments that feature course like experiences—for example, lectures, labs, discussions, and assessments—for little to no cost	Changing "Course": Reconceptualizing Educational Variables for Massive Open Online Courses	2014	139
(Kaplan & Haenlein, 2016)	....A MOOC is an open-access online course (i.e., without specific participation restrictions) that allows for unlimited (massive) participation.	Higher education and the digital revolution: About MOOCs, SPOCs, social media, and the Cookie Monster	2016	132
(Kay, Reimann, Diebold & Kummerfeld, 2013)	....They're <i>open</i> , meaning that anyone can use them to learn. This also logically implies that they're free, removing any financial barrier for even the poorest student.	MOOCs: So many learners, so much potential.	2013	128
(Wu & Chen, 2017)	....The advantages of MOOCs are large scale, openness and self-organization. MOOCs enable students to access free and open education provided by the most reputable universities, which attract substantially larger audiences than traditional online education.	Continuance intention to use MOOCs: Integrating the technology acceptance model (TAM) and task technology fit (TTF) model	2017	127
(DeWaard et al., 2011)	....It is our belief that the MOOC format allows massive participation leading to the creation of possible educational futures.	Using mLearning and MOOCs to understand chaos, emergence, and complexity in education	2011	117
(Glace, Forsey & Riley, 2013)	....What is new is the numbers of participants, and the fact that the format concentrates on short form videos, automated or peer/self-assessment, forums and ultimately open content from a representation of the world's leading higher educational institutions.	The pedagogical foundations of massive open online courses	2013	108

(Ebben & Murphy, 2014)	....Today's MOOCs are internet-provided courses, open to anyone with web access, typically free of charge with no penalties for non-participation.	Unpacking MOOC scholarly discourse: A review of nascent MOOC scholarship	2014	101
(Baggaley, 2013)	....MOOCs tend to be simpler and more impersonal than previous forms of online education: no teachers; no supervision; no fees nor entry requirements; the only equipment required being the computers purchased by the students; thousands of students in a single course; students teaching each other; students grading each others' work.	MOOC rampant	2013	100
(Barak, Watted & Haick, 2016)	....Massive open online courses (MOOCs) provide people from all over the world the opportunity to expand their education for free without any commitment or prior requirements.	Motivation to learn in massive open online courses: Examining aspects of language and social engagement	2016	99
(Gillani & Eynon, 2014)	....MOOCs are hybrids of previous attempts at online distance education: they bring together early approaches to online learning and the scale and potential reach of open courseware efforts.	Communication patterns in massively open online courses	2014	98

\*At the time of data extraction from SCOPUS® database



# Device modelling of lead free $(\text{CH}_3\text{NH}_3)_2\text{CuX}_4$ based perovskite solar cells using SCAPS simulation

Rahul Kundara<sup>1</sup> · Sarita Baghel<sup>1</sup>

Received: 21 August 2022 / Accepted: 28 July 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

## Abstract

The Copper (Cu)-based perovskite materials,  $(\text{CH}_3\text{NH}_3)_2\text{CuX}_4$  or  $(\text{MA})_2\text{CuX}_4$  with  $[\text{X} = \text{Cl}_4, \text{Cl}_2\text{I}_2, \text{and } \text{Cl}_2\text{Br}_2]$  are explored for use in perovskite solar cells (PSCs). The foremost objectives of this investigation are the optimization and finding the combination of Electron Transport Layer [ETL], Perovskite Absorber Layer (PAL) and the different organic and inorganic Hole Transport Layers [HTLs] for better device performance. The impact of other important functional parameters on the performance of PSCs are also studied. These parameters are, thicknesses of PAL, operating temperature (T), series resistance ( $R_s$ ), and radiative recombination rate under the illuminance of AM1.5. This SCAPS-1D simulation study deduced the optimized value of the thickness for  $(\text{MA})_2\text{CuCl}_4$ ,  $(\text{MA})_2\text{CuCl}_2\text{I}_2$  and  $(\text{MA})_2\text{CuCl}_2\text{Br}_2$  based absorber layer to be 400 nm, 500 nm and 600 nm, respectively at defect density ( $N_t$ ) of  $1 \times 10^{13} \text{ cm}^{-3}$  and 300 K operating temperature. The optimum value of operating temperature is 300 K for all PSCs but for  $\text{C}_{60}/(\text{MA})_2\text{CuCl}_4/\text{Cu}_2\text{O}$  PSC, optimum value is 320 K at 400 nm of absorber layer. With considerations of all these optimum values, the highest power conversion efficiency of 28.31% has been obtained for the PCBM/ $(\text{MA})_2\text{CuCl}_2\text{Br}_2/\text{CuI}$  configuration at operating temperature of 300 K. Thus, the study reveals that PCBM as ETL, while CuI and  $\text{Cu}_2\text{O}$  as HTLs are most suitable for the Cu-based PSC. Based upon the comparison with experimental results, our findings are indicative of the fact that traditional charge transport materials like  $\text{TiO}_2$  and spiro-OMeTAD may not be the best choices for new lead-free Cu-based PSCs.

**Keywords** SCAPS-1D · Cu-based perovskite solar cell · Hole transport layer · Electron transport layer

## 1 Introduction

In the field of solar photovoltaic research, organic–inorganic perovskite materials are gaining considerable interest. These materials have a lot of potential as they are inexpensive, abundant and have easy processing techniques. Currently, in

✉ Rahul Kundara  
rahul\_2k20phdap10@dtu.ac.in

<sup>1</sup> Department of Applied Physics, Delhi Technological University, Bawana Road, Delhi 110042, India

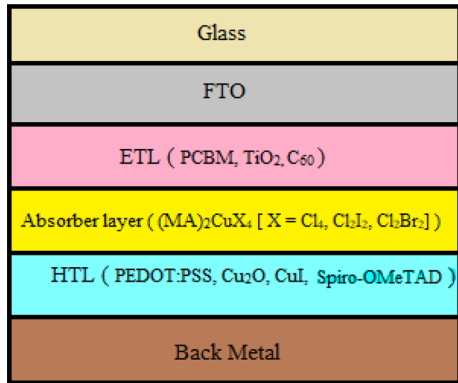


comparison with other perovskite materials, lead perovskites of methyl ammonium halide ( $\text{CH}_3\text{NH}_3\text{PbX}_3$ ) are achieving higher power conversion efficiencies (Mahajan et al. 2021; Sahli et al. 2018). Some of the factors associated with their superior photovoltaic performance are suitable band gap, long electron and hole diffusion lengths, high absorption coefficients, low defect density ( $N_i$ ) (Momblona et al. 2016; Park 2015; Cao et al. 2021; Kumar et al. 2021). However, despite the several advantages, lead based perovskites cannot be regarded as a sustainable class of photovoltaics as lead is highly toxic, not only to living beings but to the entire earth ecosystem (Giustino and Snaith 2016). In addition to the environmental concern, lead based perovskite solar cells also face stability issues due to rapid oxidation of Pb cation (Ke et al. 2019).

There is a need to replace lead-based perovskites with some other perovskite materials which are not only less toxic and environment friendly, but also, have potential to deliver superior photovoltaic performance. In order to find a suitable substitute for lead, researchers have investigated many potential candidates. Lead has been successfully replaced by Sn (Noel et al. 2014), Ge (Ju et al. 2018), Bi (Zhang et al. 2017), Sb (Wang et al. 2018), Ag (Zong et al. 2018) and subsequent experiments have demonstrated device power conversion efficiency up to 9% (Kour et al. 2019). However, there remain many challenges that need to be addressed.  $\text{Sn}^{2+}$  and  $\text{Sb}^{2+}$  exhibit low open circuit voltage whereas  $\text{Ge}^{2+}$  has stability issues due to oxidation.

Therefore, exploration of transition metals has now gathered momentum. Many transition metals (e.g.  $\text{Fe}^{2+}$ ,  $\text{Cu}^{2+}$ ,  $\text{Zn}^{2+}$ ) are attractive alternatives as they are cost effective, earth abundant and low on toxicity. In this direction, Cu metal has also been investigated as a substitute for lead. To the best of our knowledge there are only two studies that explored the application of copper as a lead substitute (Cortecchia et al. 2016; Elseman et al. 2018). Recently, PCE efficiencies up to 2.41% has been experimentally reported by Elseman et al. (2018) by using Cu substituted lead perovskite materials ( $(\text{MA})_2\text{CuX}_4$ ) for solar cell fabrication. This Cu based hybrid perovskite material also exhibits band gap tunability with varying content of halides ( $\text{X} = \text{Cl}_4, \text{Cl}_2\text{I}_2, \text{Cl}_2\text{Br}_2$ ). However, despite optimized Low efficiency was attributed to factors such as recombination rate, absorption coefficient, surface roughness and thickness of perovskite films. Also, mismatch between the band levels of electron transport layer (ETL) and hole transfer layer (HTL) with respect to perovskite absorber layer (PAL) was identified as an important functional parameter for low efficiency. A general perovskite solar cell structure contains a PAL sandwiched between ETL and HTL. A photocurrent is generated when photoelectrons ( $e^-$ ) are injected into ETL and holes ( $h^+$ ) are transferred to HTL from the perovskite layer. Hence, proper alignment of energy band levels of all three layers is extremely crucial for efficient device performance. so there exists a need to carefully select and optimize ETL and HTL according to the properties of the perovskite absorber layer.  $(\text{MA})_2\text{CuX}_4$ ,  $\text{TiO}_2$  and spiro-OMeTAD were used as ETL and HTL materials respectively, which are the traditional choices. However, other suitable alternatives are also available for this role. In this work, we have examined effect of different organic and inorganic ETL and HTL materials on the performance of  $(\text{MA})_2\text{CuCl}_4$ ,  $(\text{MA})_2\text{CuCl}_2\text{I}_2$ ,  $(\text{MA})_2\text{CuCl}_2\text{Br}_2$  based PSCs using SCAPS-1D simulation software. Simulation work also includes analysis and optimization of different device parameters (perovskite layer thickness, recombination rate, device temperature, series resistance) for enhanced photovoltaic performance.

**Fig. 1** Architecture of the simulated model of  $(\text{MA})_2\text{CuCl}_4$ ,  $(\text{MA})_2\text{CuCl}_2\text{I}_2$  and  $(\text{MA})_2\text{CuCl}_2\text{Br}_2$  PSC



**Table 1** Various input parameters of perovskite layers employed in simulation

Parameter	$(\text{MA})_2\text{CuCl}_4$ (Elseman et al. 2018)	$(\text{MA})_2\text{CuCl}_2\text{I}_2$ (Elseman et al. 2018)	$(\text{MA})_2\text{CuCl}_2\text{Br}_2$ (Elseman et al. 2018)
Thickness ( $\mu\text{m}$ )	0.400	0.400	0.400
$E_g$ (eV)	2.36	1.99	1.04
$\chi$ (eV)	2.92	3.9	3.8
$\epsilon_r$	25	20	15
$N_C$ ( $\text{cm}^{-3}$ )	$3.5 \times 10^{20}$	$2.5 \times 10^{20}$	$3.0 \times 10^{18}$
$N_V$ ( $\text{cm}^{-3}$ )	$3.5 \times 10^{20}$	$2.5 \times 10^{20}$	$4.0 \times 10^{18}$
$\mu_n$ ( $\text{cm}^2/\text{Vs}$ )	10	14	15
$\mu_p$ ( $\text{cm}^2/\text{Vs}$ )	10	14	15
$N_D$ ( $\text{cm}^{-3}$ )	$7.0 \times 10^{14}$	—	$1.0 \times 10^{10}$
$N_A$ ( $\text{cm}^{-3}$ )	$7.0 \times 10^{14}$	$6.0 \times 10^{14}$	$1.0 \times 10^{10}$
$N_t$ ( $\text{cm}^{-3}$ )	$1 \times 10^{13}$	$1 \times 10^{13}$	$1 \times 10^{13}$

## 2 Material and methods

### 2.1 Device architecture

The device architecture for the simulation of PSCs is shown below in Fig. 1. The primary architecture of PSC basically consists of following different layers, a PAL is placed in the middle of the ETL and HTL. In this structure we have  $(\text{MA})_2\text{CuX}_4$  PAL with different halides groups such as  $\text{Cl}_4$ ,  $\text{Cl}_2\text{I}_2$ , and  $\text{Cl}_2\text{Br}_2$  are utilized in PSC with the thickness of 400 nm shown below in Table 1. In modelling of Cu-based PSC, the various ETLs and HTLs are employed for obtaining the maximum PCE with optimum value of absorber layer thickness. The thickness of three different ETLs such as PCBM,  $\text{TiO}_2$  and  $\text{C}_{60}$  with values of 500 nm, 40 nm and 50 nm respectively employed shown below in Table 2. The various HTLs are employed for efficient PSC such as PEDOT: PSS,  $\text{Cu}_2\text{O}$ , CuI, and Spiro-OMeTAD. The thickness of the HTLs varies from 80 to 250 nm. In p-i-n devices, electrons are collected at the fluorine doped tin oxide (FTO) and holes at the metal back contact (Bhattacharai and Das 2021).

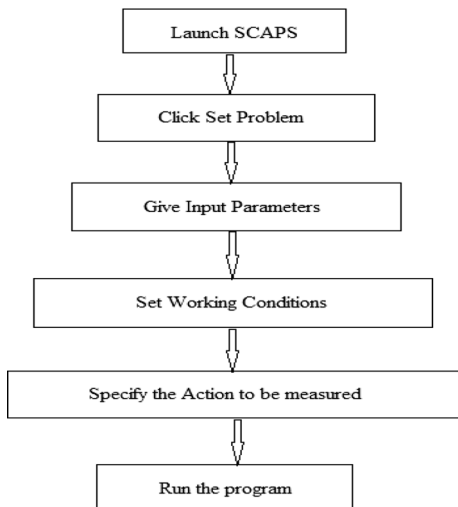
**Table 2** Various input parameters employed in simulation of different ETLs

Parameter	TiO <sub>2</sub> (Rai et al. 2020; Lakhdar and Hima 2020)	PCBM (Mandadapu et al. 2017; Azri et al. 2019)	C <sub>60</sub> (Jayan and Sebastian 2021)
Thickness (μm)	0.040	0.500	0.050
E <sub>g</sub> (eV)	3.2	2.1	1.7
χ (eV)	3.9	3.9	3.9
ε <sub>r</sub>	9	3.9	4.2
N <sub>C</sub> (cm <sup>-3</sup> )	1 × 10 <sup>21</sup>	2.2 × 10 <sup>19</sup>	8.0 × 10 <sup>19</sup>
N <sub>V</sub> (cm <sup>-3</sup> )	2 × 10 <sup>20</sup>	2.2 × 10 <sup>19</sup>	8.0 × 10 <sup>19</sup>
μ <sub>n</sub> (cm <sup>2</sup> /Vs)	20	0.001	8.0 × 10 <sup>-2</sup>
μ <sub>p</sub> (cm <sup>2</sup> /Vs)	10	0.002	3.5 × 10 <sup>-3</sup>
N <sub>D</sub> (cm <sup>-3</sup> )	1 × 10 <sup>19</sup>	1 × 10 <sup>19</sup>	2.6 × 10 <sup>17</sup>
N <sub>A</sub> (cm <sup>-3</sup> )	–	–	–
N <sub>t</sub> (cm <sup>-3</sup> )	1 × 10 <sup>15</sup>	1 × 10 <sup>9*</sup>	1 × 10 <sup>14</sup>

\*In this work

## 2.2 Simulation parameters

In this work, the SCAPS-1D software is employed in simulation. The Gent University of Belgium developed SCAPS simulation software, which allows users to simulate a maximum of seven semiconducting layers in both light and dark conditions. The stepwise procedure of SCAPS-1D simulation is demonstrated in Fig. 2. The Photovoltaic (PV) parameters such as open circuit voltage ( $V_{OC}$ ), short circuit current density ( $J_{SC}$ ), fill factor (FF) and PCE can be calculated at different temperatures and illuminations. The software is numerically based on solving semiconductor, Poisson's, and continuity equations for both electrons and holes under steady-state conditions given below.

**Fig. 2** SCAPS-1D software simulation procedure

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{q}{\epsilon} [p(x) - n(x) + N_D - N_A + \rho_p - \rho_n] = 0 \quad (1)$$

$$\frac{1}{q} \frac{dJ_p}{dx} = G_{op}(x) - R(x) \quad (2)$$

$$\frac{1}{q} \frac{dJ_n}{dx} = -G_{op}(x) + R(x) \quad (3)$$

The basic structure of PSC consists of three different layers such as  $\text{TiO}_2$  (n-type ETL), PAL and p-type HTL demonstrated in Fig. 1. The primarily input data employed in simulation of various layers are enclosed in Tables 1, 2 and 3 including thickness, band gap energy ( $E_g$ ), electron affinity ( $\chi$ ), relative dielectric permittivity ( $\epsilon_r$ ), mobility of electron ( $\mu_n$ ), mobility of hole ( $\mu_p$ ) and  $N_t$ . This simulation is based on the change in thickness of absorber layer, operating temperature and  $N_t$  of different PSCs (Rai et al. 2020; Haider et al. 2019).

### 3 Result and discussion

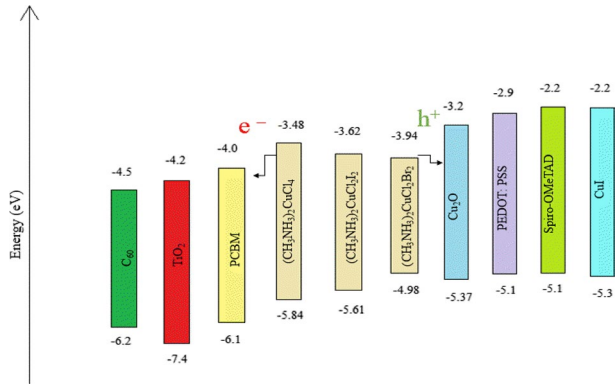
#### 3.1 Modulation of various ETLs and HTLs to impact device performance

In this study we have simulated different ETLs and HTLs, the ETLs are  $\text{TiO}_2$ , PCBM and  $\text{C}_{60}$  with the  $E_g$  of 3.2 eV, 2.1 eV and 1.7 eV respectively. Four different HTLs employed in simulation are CuI,  $\text{Cu}_2\text{O}$ , inorganic materials and Spiro-OMeTAD and PEDOT: PSS are organic materials. The bandgap energy for CuI is 3.1 eV,  $\text{Cu}_2\text{O}$  is 2.17 eV, Spiro-OMeTAD is 3 eV and for PEDOT: PSS is 2.2 eV shown in Table 3. The energy level diagram of ETL, PAL and HTL demonstrated in Fig. 3. The optimum thickness of each ETL was first found to be of 500 nm, 40 nm and 50 nm respectively by multiple simulations as shown below in Table 2. These values were then kept constant. For a particular PAL, all possible combinations of 3 ETLs and 4 HTLs are studied. Hence for each PAL, there are 12 combinations. Hence, the resulting PCEs are shown in the Table 4. Best two configurations for each perovskite are chosen for further evaluation through J–V characteristics. J–V characteristics for the best six configurations have been shown in Fig. 4. The variation of quantum efficiency with wavelength for various optimized Cu-based PSCs shown in Fig. 5. These Simulation outcomes demonstrated that  $(\text{MA})_2\text{CuCl}_4$  based PSC gives the maximum PCE is 18.41% for PCBM as ETL and PEDOT: PSS as HTL. This can be attributed to high electrical conductivity of PCBM and the highest occupied molecular orbital (HOMO) level properly matching with PAL with their  $E_g$  difference of 0.74 eV. Similarly, by using  $(\text{MA})_2\text{CuCl}_2\text{I}_2$  as absorber layer PSC gives PCE of 17.38% for  $\text{Cu}_2\text{O}$  with PCBM as ETL due to high hole mobility responsible for high performance of PSC and exact band alignment between HOMO level of PAL and  $\text{Cu}_2\text{O}$  (Anwar et al. 2017; Hossain and Alharbi 2013; Minami et al. 2014). However, the  $(\text{MA})_2\text{CuCl}_2\text{Br}_2$  based PSC gives the maximum PCE of 28.31% which is highest among all of three Cu-based PSCs by employing PCBM as ETL and CuI as HTL (Yamada et al. 2016). The  $(\text{MA})_2\text{CuCl}_2\text{Br}_2$  PSC has a high absorption coefficient which is responsible for high value of current density ( $J_{sc}$ ) resulting in high efficiency shown in Fig. 6 (Elseman et al. 2018). Material with a low absorption coefficient,

**Table 3** Various input parameters employed in simulation of different HTLs

Parameter	FTO (Rai et al. 2020; Haider et al. 2019)	CuI (Kanoun et al. 2019)	Spiro-OMeTAD (Bhattarai and Das 2021; Los Santos et al. 2020)	Cu <sub>2</sub> O (Rai et al. 2020; Kanoun et al. 2019)	PEDOT: PSS (Mandadapu et al. 2017; Jayan and Sebastian 2021)
Thickness (μm)	0.400	0.100	0.213	0.250	0.080
E <sub>g</sub> (eV)	3.5	3.1	3	2.17	2.2
χ (eV)	4.0	2.1	2.2	3.2	2.9
ε <sub>r</sub>	9.0	6.5	3	7.11	3.0
N <sub>C</sub> (cm <sup>-3</sup> )	2.02 × 10 <sup>18</sup>	2.8 × 10 <sup>19</sup>	1 × 10 <sup>19</sup>	2.02 × 10 <sup>17</sup>	2.2 × 10 <sup>15</sup>
N <sub>V</sub> (cm <sup>-3</sup> )	1.8 × 10 <sup>19</sup>	1 × 10 <sup>19</sup>	1 × 10 <sup>19</sup>	1.1 × 10 <sup>19</sup>	1.8 × 10 <sup>18</sup>
μ <sub>n</sub> (cm <sup>2</sup> /Vs)	20	100	10 <sup>-4</sup>	200	1 × 10 <sup>-2</sup>
μ <sub>p</sub> (cm <sup>2</sup> /Vs)	10	43.9	10 <sup>-4</sup>	80	2 × 10 <sup>-4</sup>
N <sub>D</sub> (cm <sup>-3</sup> )	2 × 10 <sup>19</sup>	—	—	—	—
N <sub>A</sub> (cm <sup>-3</sup> )	0	1 × 10 <sup>18</sup>	2 × 10 <sup>19</sup>	1 × 10 <sup>18</sup>	10 <sup>19</sup>
N <sub>I</sub> (cm <sup>-3</sup> )	1 × 10 <sup>15</sup>	1 × 10 <sup>14</sup>	1 × 10 <sup>14</sup>	1 × 10 <sup>14</sup>	1 × 10 <sup>14</sup>

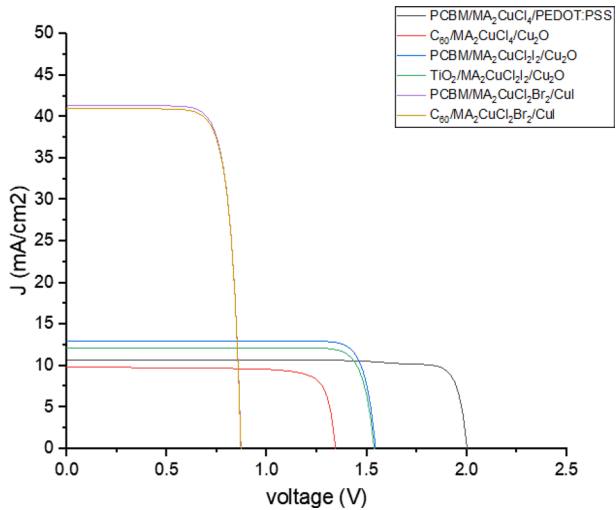
**Fig. 3** The energy level alignment between ETL, PAL and HTL materials



**Table 4** PCE (%) of various combinations of Cu-based PSC with different ETL and HTL

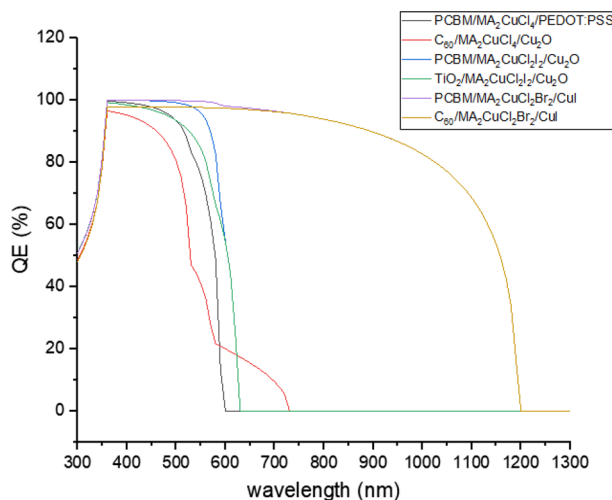
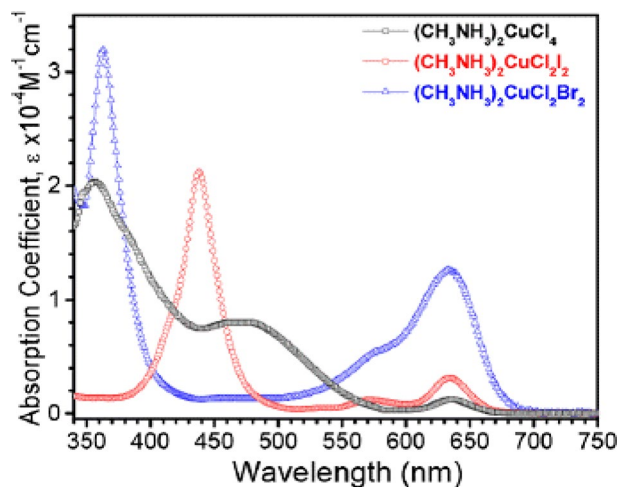
HTL	(MA) <sub>2</sub> CuCl <sub>4</sub>			(MA) <sub>2</sub> CuCl <sub>2</sub> I <sub>2</sub>			(MA) <sub>2</sub> CuCl <sub>2</sub> Br <sub>2</sub>		
	ETL			ETL			ETL		
	TiO <sub>2</sub>	C <sub>60</sub>	PCBM	TiO <sub>2</sub>	C <sub>60</sub>	PCBM	TiO <sub>2</sub>	C <sub>60</sub>	PCBM
PEDOT: PSS	8.42	9.87	18.41	13.44	11.67	14.94	12.20	22.45	22.53
Spiro-OMeTAD	7.81	9.47	18.40	14.69	12.63	16.51	12.49	6.68	6.68
CuI	7.76	9.46	18.40	14.46	11.79	16.26	12.33	28.09	28.31
Cu <sub>2</sub> O	8.96	10.68	18.39	16.13	14.28	17.38	14.41	13.79	13.78

**Fig. 4** J-V characteristics of Cu-based PSCs



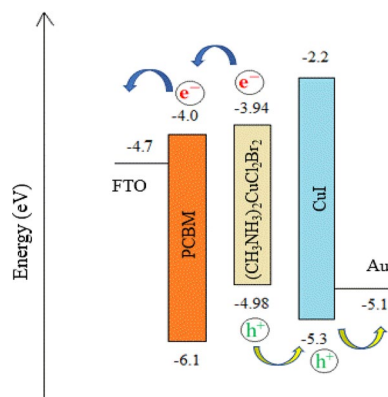
light is imperfectly absorbed, it should be large for high performance of PSC (<https://www.pveducation.org/pvcdrom/pn-junctions/absorption-coefficient/>; Kalaiselvi et al. 2018). The absorption coefficient of (MA)<sub>2</sub>CuCl<sub>2</sub>Br<sub>2</sub> based PSC is greater than other two Cu-based PSCs hence the PCE of (MA)<sub>2</sub>CuCl<sub>2</sub>Br<sub>2</sub> based PSC is higher than the (MA)<sub>2</sub>CuCl<sub>2</sub>I<sub>2</sub> based



**Fig. 5** Quantum efficiency (QE) curve for Cu-based PSCs**Fig. 6** Absorption coefficient of Cu-based PSCs (Elseman et al. 2018)

PSC (Elseman et al. 2018). In addition,  $(\text{MA})_2\text{CuCl}_2\text{Br}_2$  absorber layer has an appropriate band gap (1.0–1.6 eV) that would maximize efficiency (Qiu et al. 2017).

The band alignment is also a crucial factor for superior performance of  $(\text{MA})_2\text{CuCl}_2\text{Br}_2$  PSC as shown in Fig. 7. The conduction band of the ETL is always less than the lowest unoccupied molecular orbital (LUMO) of the PAL. The difference between the valence band of ETL and HOMO level of PAL should be always high, which prevents the recombination in the PAL. In this simulation work we have employed various HTLs among which CuI is most suitable for obtaining high performance of Cu-based PSCs. The CuI layer as HTL is more appropriate than spiro-OMeTAD layer because of the high hole mobility and low hysteresis. The HOMO level of CuI and PEDOT: PSS has good band alignment with absorber layer of  $(\text{MA})_2\text{CuCl}_2\text{Br}_2$  based PSC therefore with CuI it has achieved best PCE but not for PEDOT: PSS because lower value of holes mobility ( $\mu_p$ ) than CuI. Hence, CuI is a more suitable HTL for achieving high efficiency for  $(\text{MA})_2\text{CuCl}_2\text{Br}_2$  based PSC. Furthermore, the CuI based PSCs exhibit good long-term stability in the ambient atmosphere

**Fig. 7** Band alignment diagram of optimized device

because of its hydrophobic property (Shi et al. 2021). The CuI has outstanding properties such as band matching with the PAL; a wide  $E_g$  of 3.1 eV; a high  $\mu_p$  of  $43.9 \text{ cm}^2/\text{Vs}$ , inexpensive; high chemical stability; and non-toxicity make CuI a good choice for HTL in PSCs. This numerical modelling helped in the investigation of the best ETL and HTL combination with  $(\text{MA})_2\text{CuCl}_2\text{Br}_2$  based PSC for the high PCE. The optimized result of various ETLs and HTLs with Cu-based PSCs is shown below in Table 5.

### 3.1.1 Comparison with experimental results

The experimentally obtained efficiency of three PAL:  $(\text{MA})_2\text{CuCl}_4$ ,  $(\text{MA})_2\text{CuCl}_2\text{Br}_2$  and  $(\text{MA})_2\text{CuCl}_2\text{I}_2$  PSCs were 2.41%, 1.75% and 0.99% respectively shown below in Table 6 using thin film structure of glass/FTO/ $\text{TiO}_2$ / $(\text{MA})_2\text{CuX}_4$ /spiro-OMeTAD/Au. The reason behind the generally low PCE achieved in all reported cells may come from the recombination occurring in perovskite layers. Apart from perovskite absorber layer properties, lower

**Table 5** Optimized PCE of Cu-based PSC with various ETLs and HTLs

Perovskite	$V_{OC}$ (V)	$J_{SC}$ ( $\text{mA}/\text{cm}^2$ )	FF (%)	PCE (%)
PCBM/ $(\text{MA})_2\text{CuCl}_4$ /PEDOT: PSS	2.00	10.63	86.48	18.41
$\text{C}_{60}/(\text{MA})_2\text{CuCl}_4/\text{Cu}_2\text{O}$	1.34	9.83	80.85	10.68
PCBM/ $(\text{MA})_2\text{CuCl}_2\text{I}_2/\text{Cu}_2\text{O}$	1.54	12.97	86.82	17.38
$\text{TiO}_2/(\text{MA})_2\text{CuCl}_2\text{I}_2/\text{Cu}_2\text{O}$	1.53	12.14	86.46	16.13
PCBM/ $(\text{MA})_2\text{CuCl}_2\text{Br}_2/\text{CuI}$	0.87	41.30	78.69	28.31
$\text{C}_{60}/(\text{MA})_2\text{CuCl}_2\text{Br}_2/\text{CuI}$	0.87	41.00	78.70	28.09

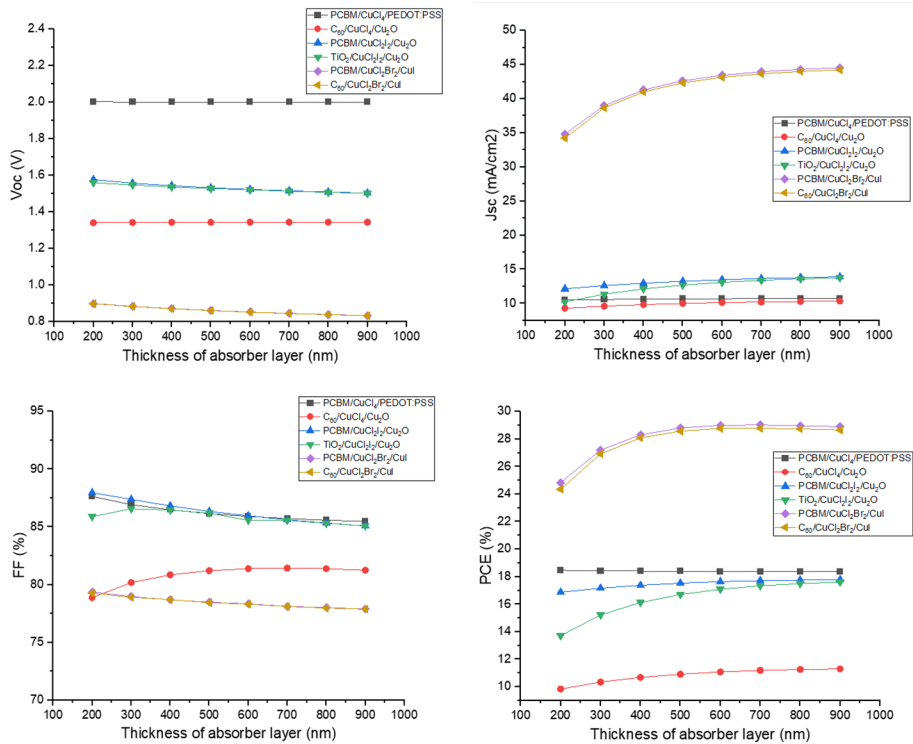
**Table 6** The previously reported result of different Cu-based PSCs

Device	$V_{OC}$ (V)	$J_{SC}$ ( $\text{mA}/\text{cm}^2$ )	FF (%)	PCE (%)
$\text{TiO}_2/(\text{MA})_2\text{CuCl}_4$ /Spiro-OMeTAD	0.560	8.12	52	2.41 (Elseman et al. 2018)
$\text{TiO}_2/(\text{MA})_2\text{CuCl}_2\text{I}_2$ /Spiro-OMeTAD	0.545	6.78	47	1.75 (Elseman et al. 2018)
$\text{TiO}_2/(\text{MA})_2\text{CuCl}_2\text{Br}_2$ /Spiro-OMeTAD	0.581	3.35	50	0.99 (Elseman et al. 2018)

efficiency can be explained by mismatch in the energy bands of ETL, HTL and PAL. For all three perovskites,  $\text{TiO}_2$  and Spiro-OMeTAD are employed as ETL and HTL respectively. These can be suitable choices for  $(\text{MA})_2\text{CuCl}_4$  but not for  $(\text{MA})_2\text{CuCl}_2\text{Br}_2$  and  $(\text{MA})_2\text{CuCl}_2\text{I}_2$  PSCs. This is evident from our simulation results. By using more appropriate HTL such as CuI and  $\text{Cu}_2\text{O}$   $(\text{MA})_2\text{CuCl}_2\text{Br}_2$  and  $(\text{MA})_2\text{CuCl}_2\text{I}_2$  PSCs can give much better results.  $\text{Cu}_2\text{O}$  as HTL is suitable for high efficiency because of its high carrier mobility and a long carrier diffusion length and high charge extraction ability leads to the high  $J_{\text{SC}}$  (Chatterjee and Pal 2016; Nejand et al. 2016). It shows high acceptor density because of a suitable band gap of 2.17 eV which may give high PV performance (Kale et al. 2021). Also, for CuI, Higher values of  $J_{\text{SC}}$  are obtained because of a much higher  $\mu_p$  of CuI relative to spiro-OMeTAD. Optimization of the perovskite/HTL interface is very crucial to reduce recombination as well as for achieving higher  $V_{\text{OC}}$  and FF for PSCs (Gharibzadeh et al. 2016). It has been found that there is an enhancement in  $V_{\text{OC}}$  due to two factors: (i) increase in charge carrier mobilities ratio ( $\mu_n/\mu_h$ ) of HTL (ii) decrease in the energy gap between HOMO of the HTL and conduction band of PAL (Omping and Singh 2018). The electron and hole mobility ratio for CuI and  $\text{Cu}_2\text{O}$  is above 2 while for spiro-OMeTAD, this ratio is 1, hence simulation results for CuI and  $\text{Cu}_2\text{O}$  as HTL give high  $V_{\text{OC}}$  leading to higher efficiencies, more so for  $(\text{MA})_2\text{CuCl}_2\text{Br}_2$  PSC than to experimental results (obtained with spiro-OMeTAD as HTL). Our findings are indicative of the fact that traditional charge transport materials like  $\text{TiO}_2$  and spiro-OMeTAD may not be the best choices for new lead-free Cu-based PSCs.

### 3.2 Impact of thickness of absorber layer on device performance

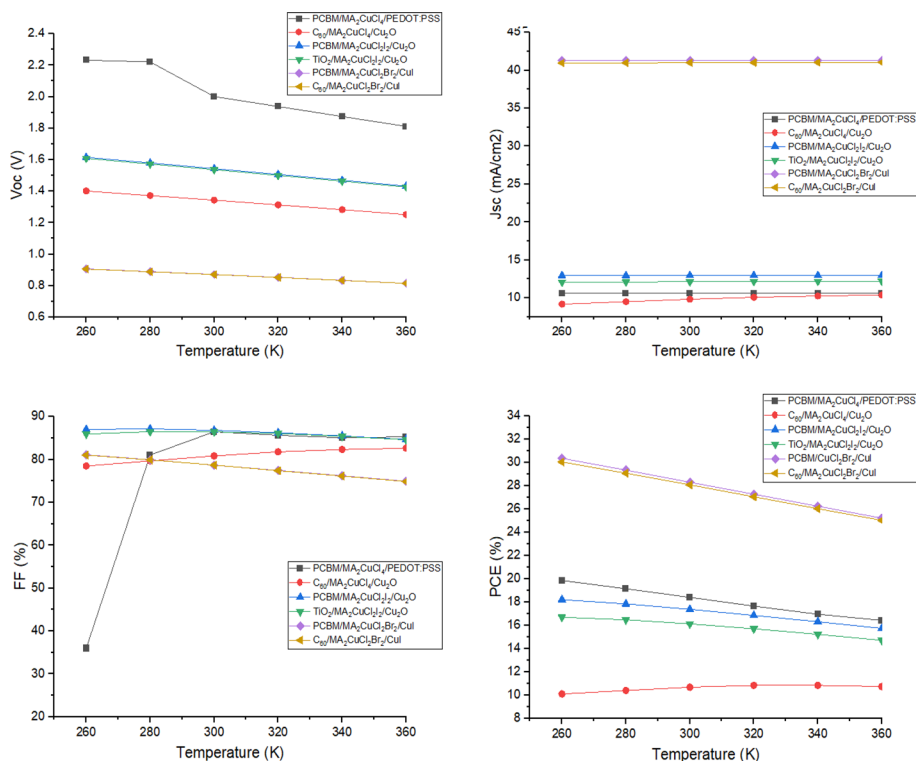
The thickness of the  $(\text{MA})_2\text{CuX}_4$  absorber layer has exhibited an apparent impact on the diffusion length of charge carriers. The thickness of PAL varied in the range of 200–900 nm at defect density ( $N_t$ ) of  $1 \times 10^{13} \text{ cm}^{-3}$  and operating temperature of 300 K. The absorption rate is low for the thin PAL which results in low photocurrent, hence efficiency decreases. As the thickness increases, the charge carriers may not have suitable diffusion length to travel up to the charge collecting layers resulting in a high value of  $R_s$  which limits the device performance. Beyond the optimum value of thickness, the device performance stagnates due to the higher recombination of electron–hole pairs. Therefore, there is an increment in the dark saturation current which allows  $V_{\text{OC}}$  and  $J_{\text{SC}}$  to fluctuate extremely slowly after a definite value of thickness. This results in flattening of the PCE curve (Mushtaq et al. 2023; Hao et al. 2021). The n-layer thickness should be optimized for high PCE of a PSC, when the thickness of ETL is near to the surface of PAL, conductance of the device increases towards absorption of radiation (Nejand et al. 2016). The obtained optimized value of thickness, for PCBM/ $(\text{MA})_2\text{CuCl}_4$ /PEDOT: PSS is 400 nm,  $\text{C}_{60}/(\text{MA})_2\text{CuCl}_4/\text{Cu}_2\text{O}$  is 600 nm, PCBM/ $(\text{MA})_2\text{CuCl}_2\text{I}_2/\text{Cu}_2\text{O}$  is 500 nm,  $\text{TiO}_2/(\text{MA})_2\text{CuCl}_2\text{I}_2/\text{Cu}_2\text{O}$  is 600 nm, PCBM/ $(\text{MA})_2\text{CuCl}_2\text{Br}_2/\text{CuI}$  is 600 nm and  $\text{C}_{60}/(\text{MA})_2\text{CuCl}_2\text{Br}_2/\text{CuI}$  is 500 nm shown below in Fig. 8 for better performance of PSC. The maximum efficiency of PCBM/ $(\text{MA})_2\text{CuCl}_2\text{Br}_2/\text{CuI}$  PSC goes up to 29.01% together with  $V_{\text{OC}}$  is 0.85 V,  $J_{\text{SC}}$  is 43.44 mA/cm<sup>2</sup> and FF is 78.34% at 600 nm. The change in other PV parameters is also shown in Fig. 8. The PCBM/ $(\text{MA})_2\text{CuCl}_4$ /PEDOT: PSS PSC is most stable towards variation in thickness of PAL because change in the PCE is only 0.43% for the range of thickness of PAL from 200 to 900 nm. For the  $\text{TiO}_2/(\text{MA})_2\text{CuCl}_2\text{I}_2/\text{Cu}_2\text{O}$  PSC change in PSC is 28.35% with thickness which varies from 200 to 900 nm, hence it is the most unstable PSC with the variation of thickness of PAL.



**Fig. 8** Change in photovoltaic parameters with the thickness of PAL for different ETL and HTL of Cu-based PSC

### 3.3 Effect of operating temperature on absorber layer

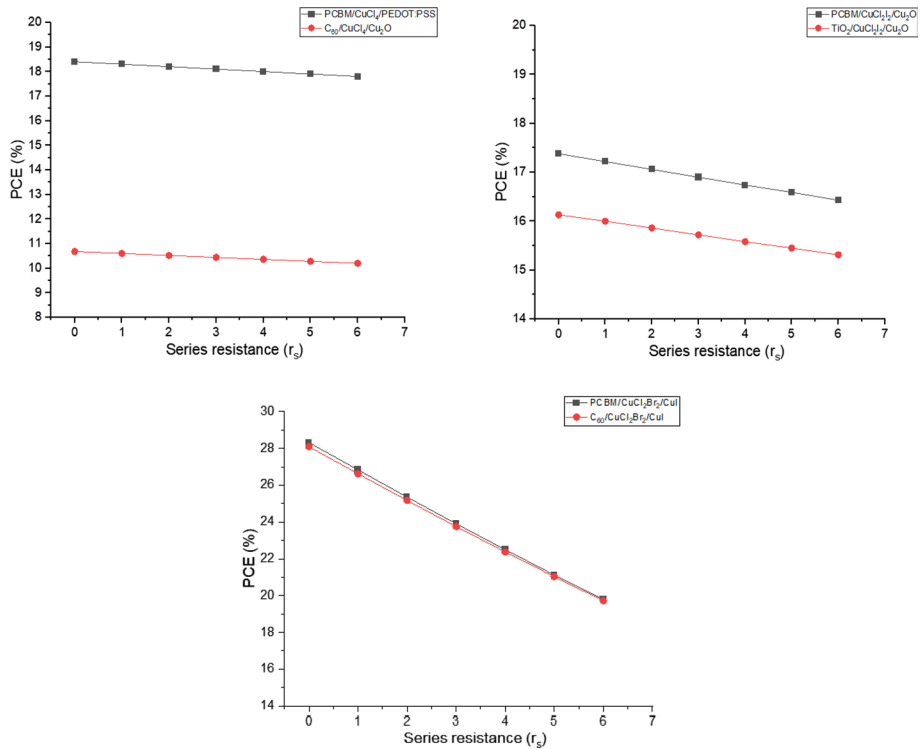
To analyze the impact of temperature on the performance of solar cells, the working temperature has been varied in the range of 260–360 K. As a rise in temperature, the PCE drops drastically. Fall in efficiency with temperature is related with decrease in diffusion length of the charge carriers. If the deformation stress on the layers is high, it results in interfacial defects and low interconnectivity among layers. This low interconnectivity alludes to increment of recombination rate in PAL, hence resulting in reducing the diffusion length and increasing  $R_s$ , fall in the performance of solar cells (Nejand et al. 2016). In PCBM/(MA)<sub>2</sub>CuCl<sub>4</sub>/PEDOT: PSS fall in efficiency with temperature is 17.31%, for C<sub>60</sub>/(MA)<sub>2</sub>CuCl<sub>4</sub>/Cu<sub>2</sub>O fall is 0.92%, for PCBM/(MA)<sub>2</sub>CuCl<sub>2</sub>I<sub>2</sub>/Cu<sub>2</sub>O fall is 13.56%, TiO<sub>2</sub>/(MA)<sub>2</sub>CuCl<sub>2</sub>I<sub>2</sub>/Cu<sub>2</sub>O is 11.85%, PCBM/(MA)<sub>2</sub>CuCl<sub>2</sub>Br<sub>2</sub>/CuI is 16.95% and for C<sub>60</sub>/(MA)<sub>2</sub>CuCl<sub>2</sub>Br<sub>2</sub>/CuI is 16.73%. This shows that the fall in efficiency of C<sub>60</sub>/(MA)<sub>2</sub>CuCl<sub>4</sub>/Cu<sub>2</sub>O PSC is most stable with increase in operating and for PCBM/(MA)<sub>2</sub>CuCl<sub>4</sub>/PEDOT: PSS PSC is most unstable. The trend of PV parameters of different Cu-based PSCs as function of operating temperature shown in Fig. 9.



**Fig. 9** Change in photovoltaic parameters with the operating temperature for various Cu-based PSCs

### 3.4 Effect of change in series resistance on performance

The impact of variation in the series resistance ( $R_s$ ) on the performance of Cu-PSC has been examined. It has a significant effect, especially on the FF and  $J_{SC}$ . As  $R_{SC}$  increases, resulting in a decrease in FF,  $J_{SC}$  also starts decreasing for a high value of  $R_s$ . Consequently, high values of  $R_s$  in a solar device result in poor PCE (Chakraborty et al. 2019). In PSC,  $R_s$  mainly exists in the interfaces: resistance at the HTL/perovskite interface, ETL/perovskite interface, and at the metal contacts. When solar cells come in contact with the environment, thermomechanical fatigue or cracks evolve in the solder bonds depending on weather circumstances. These cracks result in increment in the value of  $R_s$  hence PCE drops (Islam et al. 2021a; Poorkazem et al. 2015). The numerical modelling outcomes on the change in  $R_s$  on the introduced (MA)<sub>2</sub>CuX<sub>4</sub> based PSCs structure is shown below in Fig. 10. The PCE of three Cu-based PSCs, (MA)<sub>2</sub>CuCl<sub>4</sub>, (MA)<sub>2</sub>CuCl<sub>2</sub>I<sub>2</sub>, and (MA)<sub>2</sub>CuCl<sub>2</sub>Br<sub>2</sub>, varies with the value of series resistance. In this study, it is analyzed that for less value of  $R_s$ , the solar cell device is performance high with large FF, resulting in high PCE. As the  $R_s$  increases, the performance of active material falls significantly (Jeon et al. 2015). It has been studied that when the  $R_s$  varied from 0 to 6 ( $\Omega \text{ cm}^2$ ), PCE fall nearly 3.25%, 4.49%, 5.46%, 5.08%, 30.06% and 29.79% respectively for PCBM/(MA)<sub>2</sub>CuCl<sub>4</sub>/PEDOT:PSS, C<sub>60</sub>/(MA)<sub>2</sub>CuCl<sub>4</sub>/Cu<sub>2</sub>O, PCBM/(MA)<sub>2</sub>CuCl<sub>2</sub>I<sub>2</sub>/Cu<sub>2</sub>O, TiO<sub>2</sub>/(MA)<sub>2</sub>CuCl<sub>2</sub>I<sub>2</sub>/Cu<sub>2</sub>O, PCBM/(MA)<sub>2</sub>CuCl<sub>2</sub>Br<sub>2</sub>/CuI and C<sub>60</sub>/(MA)<sub>2</sub>CuCl<sub>2</sub>Br<sub>2</sub>/CuI at thickness of 400 nm, defect density is  $1 \times 10^{13} \text{ cm}^{-3}$



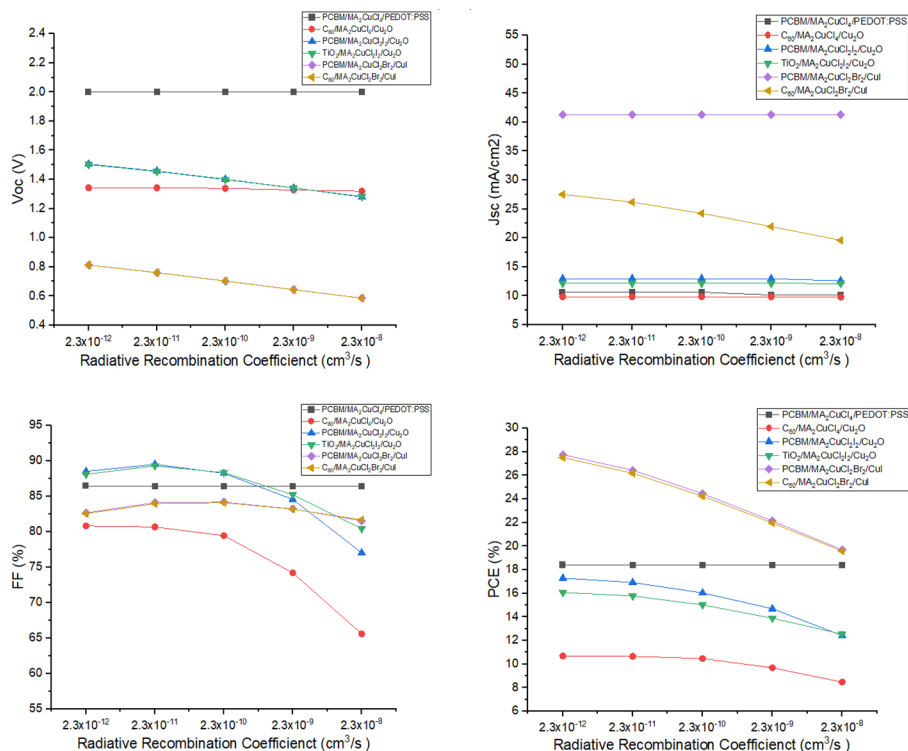
**Fig. 10** Change in PCE of different PSCs with  $R_s$

and operating temperature is 300 K shown below in Fig. 10. This shows that the PCBM/ $(\text{MA})_2\text{CuCl}_4$ /PEDOT: PSS is the most stable and PCBM/ $(\text{MA})_2\text{CuCl}_2\text{Br}_2$ /CuI is the most unstable PSC with increase in series resistance.

### 3.5 Radiative recombination rate effect on PCE

The effect on performance of solar devices by changing the radiative recombination rate. In this simulation study we varied the radiative recombination rate from  $2.3 \times 10^{-8}$  to  $2.3 \times 10^{-12}$   $\text{cm}^3/\text{sec}$  and examined PV parameters trend of PSC devices for different rates with the thickness of PAL which is between 200 and 900 nm. Usually, increment in radiative recombination, carrier lifetime reduces, and PV parameters of device is influenced as shown in Fig. 11 for Cu-based PSCs respectively. The Gaussian shape of the curve can be explained by the fact that with an increment in thickness of PAL, large electron-hole pairs are produced (Islam et al. 2021b). The PCE of  $(\text{MA})_2\text{CuCl}_4$  based PSC with PCBM as ETL and PEDOT: PSS as HTL, the maximum PCE has achieved is 18.43% at PAL thickness of 200 nm and recombination rate of  $2.3 \times 10^{-12}$   $\text{cm}^3/\text{sec}$  whereas when  $\text{C}_{60}$  as ETL and  $\text{Cu}_2\text{O}$  as HTL the maximum PCE is 11.29% at thickness of 900 nm with recombination rate of  $2.3 \times 10^{-12}$   $\text{cm}^3/\text{sec}$ . The obtained PCE is 17.74% for  $(\text{MA})_2\text{CuCl}_2\text{I}_2$  based PSC with PCBM as ETL and  $\text{Cu}_2\text{O}$  as HTL and when  $\text{TiO}_2$  as ETL and  $\text{Cu}_2\text{O}$  as HTL maximum PCE is 17.55% at PAL thickness of 900 nm with the value of recombination rate is  $2.3 \times 10^{-12}$   $\text{cm}^3/\text{sec}$ . For the  $(\text{MA})_2\text{CuCl}_2\text{Br}_2$  based PSC with PCBM as ETL and CuI as HTL, the PCE is 28.61% whereas  $\text{C}_{60}$  as ETL and





**Fig. 11** The change in PV parameters with recombination coefficient for Cu-based PSCs

CuI as HTL obtained maximum PCE is 28.41% at PAL thickness of 700 nm for the recombination rate of  $2.3 \times 10^{-12} \text{ cm}^3/\text{sec}$  shown below in Fig. 11.

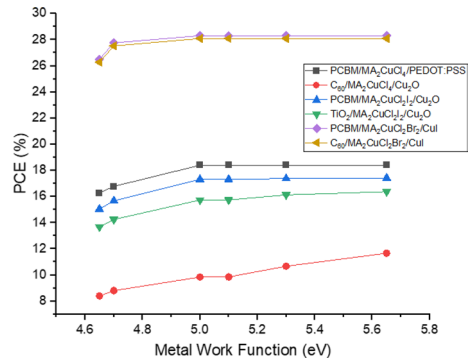
### 3.6 Effect of metal back contact work function on PCE

The power conversion efficiencies listed in Table 5 have been achieved by choosing Gold (Au) as a metal electrode. Gold was taken after carefully studying the impact of metal electrode work function on the PCE as shown in Fig. 12. The work function was varied from 4.6 to 5.8 eV. This range covers the work function of most of the available metals which are employed as back contact in PSCs. Figure 12 reveals that the photovoltaic performance decreases gradually as the work function of the metal back contact is reduced below 5.1 eV due to formation of Schottky junction at a smaller value of work function (Jannat et al. 2021). We observed that the maximum PCE of 28.31% achieved at work function of 5.1 eV (Au) (Kanoun et al. 2019).

## 4 Conclusions

We have simulated environment-friendly Cu-based PSCs with optimized ETLs and HTLs for achieving maximum PCE. The PV parameters for first optimized cell FTO/PCBM/(MA)<sub>2</sub>CuCl<sub>4</sub>/PEDOT: PSS PSC are PCE of 18.41%,  $V_{oc}$  is 2.00 V,  $J_{sc}$  is

**Fig. 12** Change in PCE of different PSCs with metal work function (eV)



10.63 mA/cm<sup>2</sup> and FF is 86.48%. The second optimized cell structure is FTO/PCBM/(MA)<sub>2</sub>CuCl<sub>2</sub>I<sub>2</sub>/Cu<sub>2</sub>O with the maximum PCE of 17.38%, V<sub>OC</sub> is 1.54 V and FF is 86.82%. Third optimized cell structure is FTO/PCBM/(MA)<sub>2</sub>CuCl<sub>2</sub>Br<sub>2</sub>/CuI has obtained maximum PCE of 28.31%, V<sub>OC</sub> is 0.87 V, J<sub>SC</sub> is 41.30 mA/cm<sup>2</sup> and FF is 78.69% at 400 nm thickness of PAL, operating temperature is 300 K and N<sub>t</sub> is 1 × 10<sup>13</sup> cm<sup>-3</sup>. The maximum PCE of 18.41%, 17.53% and 29.01% have been obtained for all three optimized PSCs with the thickness of PAL as 400 nm, 500 nm and 600 nm respectively for first, second and third cell by using SCAPS-1D. This work also comprises the operating temperature optimization for high performance of PSCs. The optimized value of temperature is 300 K for the PSCs structure. The maximum PCE has obtained of 18.41%, 17.38% and 28.31% for the PSC structure such as FTO/PCBM/(MA)<sub>2</sub>CuCl<sub>4</sub>/PEDOT: PSS, FTO/PCBM/(MA)<sub>2</sub>CuCl<sub>2</sub>I<sub>2</sub>/Cu<sub>2</sub>O and FTO/PCBM/(MA)<sub>2</sub>CuCl<sub>2</sub>Br<sub>2</sub>/CuI respectively at 300 K. Experimentally published results are also compared with simulation results and findings were analyzed which highlighted poor ETL and HTL selection as one of the reasons for lower performance of (MA)<sub>2</sub>CuX<sub>4</sub> PSCs. The series resistance (R<sub>s</sub>) effect on PCE is also studied which indicates that (MA)<sub>2</sub>CuCl<sub>4</sub> PSC is most stable and (MA)<sub>2</sub>CuCl<sub>2</sub>Br<sub>2</sub> based PSC is highly unstable with increase in value of R<sub>s</sub>. The optimized radiative recombination rate is 2.3 × 10<sup>-12</sup> cm<sup>3</sup>/sec with function of PAL thickness for high PCE of Cu-based PSC with different ETLs and HTLs. Copper based PSC overcome stability and toxicity issue in PSCs. This work helpful in deeply understanding of design, operation mechanism, and in optimization of high efficiency Cu-based PSC in near future.

**Acknowledgements** The author, Rahul Kundara is thankful to Delhi Technological University, New Delhi for providing junior research fellowship. We are also very grateful to Professor Marc Burgelman (University of Gent, Belgium) for providing us with SCAPS-1D software.

**Author contributions** RK—Conceptualization, Formal analysis, Data curation, Investigation, Methodology, Writing—original draft, Visualization and Software. SB—Supervision, Project administration, critical review and commentary or revision.

**Funding** This declaration is “not applicable”.

**Data availability** This declaration is “not applicable”.

## Declarations

**Conflict of interest** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Ethical approval** This declaration is “not applicable”.

## References

- Anwar, F., Mahbub, R., Satter, S.S., Ullah, S.M.: Effect of different HTM layers and electrical parameters on ZnO nanorod-based lead-free perovskite solar cell for high-efficiency performance. *Int. J. Photoenergy* 2017, 9846310 (2017). <https://doi.org/10.1155/2017/9846310>
- Azri, F., Meftah, A., Sengouga, N., Meftah, A.: Electron and hole transport layers optimization by numerical simulation of a perovskite solar cell. *Sol. Energy* 181, 372–378 (2019). <https://doi.org/10.1016/j.solener.2019.02.017>
- Bhattacharjee, S., Das, T.D.: Optimization of carrier transport materials for the performance enhancement of the  $\text{MAGeI}_3$  based perovskite solar cell. *Sol. Energy* 217, 200–207 (2021). <https://doi.org/10.1016/j.solener.2021.02.002>
- Cao, Y., Liu, Z., Li, W., Zhao, Z., Xiao, Z., Lei, B., Zi, W., Cheng, N., Liu, J., Tu, Y.: Efficient and stable  $\text{MAPbI}_3$  perovskite solar cells achieved via chlorobenzene/perylene mixed anti-solvent. *Sol. Energy* 220, 51–257 (2021). <https://doi.org/10.1016/j.solener.2021.03.055>
- Chakraborty, K., Choudhury, M.G., Paul, S.: Numerical study of  $\text{Cs}_2\text{TiX}_6$  ( $\text{X} = \text{Br}^-$ ,  $\text{I}^-$ ,  $\text{F}^-$  and  $\text{Cl}^-$ ) based perovskite solar cell using SCAPS-1D device simulation. *Sol. Energy* 194, 886–892 (2019). <https://doi.org/10.1016/j.solener.2019.11.005>
- Chatterjee, S., Pal, A.J.: Introducing  $\text{Cu}_2\text{O}$  thin films as a hole-transport layer in efficient planar perovskite solar cell structures. *J. Phys. Chem. C* 120(3), 1428–1437 (2016). <https://doi.org/10.1021/acs.jpcc.5b11540>
- Cortecchia, D., Dewi, H.A., Yin, J., Bruno, A., Chen, S., Baikie, T., Boix, P.P., Grätzel, M., Mhaisalkar, S., Soci, C., Mathews, N.: Lead-free  $\text{MA}_2\text{CuCl}_x\text{Br}_{4-x}$  hybrid perovskites. *Inorg. Chem.* 55(3), 1044–1052 (2016). <https://doi.org/10.1021/acs.inorgchem.5b01896>
- De Los Santos, I.M., Cortina-Marrero, H.J., Ruíz-Sánchez, M.A., Hechavarría-Difur, L., Sánchez-Rodríguez, F.J., Courel, M., Hu, H.: Optimization of  $\text{CH}_3\text{NH}_3\text{PbI}_3$  perovskite solar cells: a theoretical and experimental study. *Sol. Energy* 199, 198–205 (2020). <https://doi.org/10.1016/j.solener.2020.02.026>
- Elseman, A.M., Shalan, A.E., Sajid, S., Rashad, M.M., Hassan, A.M., Li, M.: Copper-substituted lead perovskite materials constructed with different halides for working  $(\text{CH}_3\text{NH}_3)_2\text{CuX}_4$ -based perovskite solar cells from experimental and theoretical view. *ACS Appl. Mater. Interfaces* 10(14), 11699–11707 (2018). <https://doi.org/10.1021/acsami.8b00495>
- Gharibzadeh, S., Nejand, B.A., Moshaii, A., Mohammadian, N., Alizadeh, A.H., Mohammadpour, R., Ahmadi, V., Alizadeh, A.: Two-step physical deposition of a compact CuI Hole-Transport layer and the formation of an interfacial species in perovskite solar cells. *Chemosuschem* 9(15), 1929–1937 (2016). <https://doi.org/10.1002/cssc.201600132>
- Giustino, F., Snaith, H.J.: Toward lead-free perovskite solar cells. *ACS Energy Lett.* 1(6), 1233–1240 (2016). <https://doi.org/10.1021/acsenenergylett.6b00499>
- Haider, S.Z., Anwar, H., Wang, M.: Theoretical device engineering for high-performance perovskite solar cells using  $\text{CuSCN}$  as hole transport material boost the efficiency above 25%. *Phys. Status Solidi (a)* 216(11), 1900102 (2019). <https://doi.org/10.1002/pssa.201900102>
- Hao, L., Li, T., Ma, X., Wu, J., Qiao, L., Wu, X., Hou, G., Pei, H., Wang, X., Zhang, X.: A tin-based perovskite solar cell with an inverted hole-free transport layer to achieve high energy conversion efficiency by SCAPS device simulation. *Opt. Quantum Electron.* 53, 1–17 (2021). <https://doi.org/10.1007/s11082-021-03175-5>
- Hossain, M.I., Alharbi, F.H.: Recent advances in alternative material photovoltaics. *Mater. Technol.* 8(1–2), 88–97 (2013) <https://www.pveducation.org/pvcdrom/pn-junctions/absorption-coefficient/> (2022). Accessed 20 February 2022
- Islam, S., Sobayel, K., Al-Kahtani, A., Islam, M.A., Muhammad, G., Amin, N., Shahiduzzaman, M., Akhtaruzzaman, M.: Defect study and modelling of  $\text{SnX}_3$ -based perovskite solar cells with SCAPS-1D. *Nanomaterials* 11(5), 1218 (2021). <https://doi.org/10.3390/nano11051218>
- Islam, M.T., Jani, M.R., Shorowordi, K.M., Hoque, Z., Gokcek, A.M., Vattipally, V., Nishat, S.S., Ahmed, S.: Numerical simulation studies of  $\text{Cs}_3\text{Bi}_2\text{I}_9$  perovskite solar device with optimal selection of electron and hole transport layers. *Optik* 231, 166417 (2021). <https://doi.org/10.1016/j.ijleo.2021.166417>

- Jannat, F., Ahmed, S., Alim, M.A.: Performance analysis of cesium formamidinium lead mixed halide based perovskite solar cell with MoOx as hole transport material via SCAPS-1D. *Optik* 228, 166202 (2021). <https://doi.org/10.1016/j.ijleo.2020.166202>
- Jayan, K.D., Sebastian, V.: Comprehensive device modelling and performance analysis of  $\text{MASnI}_3$  based perovskite solar cells with diverse ETM, HTM and back metal contacts. *Sol. Energy* 217, 40–48 (2021). <https://doi.org/10.1016/j.solener.2021.01.058>
- Jeon, N.J., Noh, J.H., Yang, W.S., Kim, Y.C., Ryu, S., Seo, J., Seok, S.I.: Compositional engineering of perovskite materials for high-performance solar cells. *Nature* 517, 476–480 (2015). <https://doi.org/10.1038/nature14133>
- Ju, M.G., Chen, M., Zhou, Y., Garces, H.F., Dai, J., Ma, L., Padture, N.P., Zeng, X.C.: Earth-abundant nontoxic titanium (IV)-based vacancy-ordered double perovskite halides with tunable 1.0 to 1.8 eV bandgaps for photovoltaic applications. *ACS Energy Lett.* 3(2), 297–304 (2018). <https://doi.org/10.1021/acsenergylett.7b01167>
- Kalaiselvi, C.R., Muthukumarasamy, N., Velauthapillai, D., Kang, M., Senthil, T.S.: Importance of halide perovskites for next generation solar cells: a review. *Mater. Lett.* 219, 198–200 (2018). <https://doi.org/10.1016/j.matlet.2018.02.089>
- Kale, A.J., Chaurasiya, R., Dixit, A.: Inorganic lead-free  $\text{Cs}_2\text{AuBiCl}_6$  perovskite absorber and  $\text{Cu}_2\text{O}$  hole transport material based single-junction solar cells with 22.18% power conversion efficiency. *Adv. Theory Simul.* 4(3), 2000224 (2021). <https://doi.org/10.1002/adts.202000224>
- Kanoun, A.A., Kanoun, M.B., Merad, A.E., Goumri-Said, S.: Toward development of high-performance perovskite solar cells based on  $\text{CH}_3\text{NH}_3\text{GeI}_3$  using computational approach. *Sol. Energy* 182, 237–244 (2019). <https://doi.org/10.1016/j.solener.2019.02.041>
- Ke, W., Stoumpos, C.C., Kanatzidis, M.G.: “Unleaded” perovskites: status quo and future prospects of tin-based perovskite solar cells. *Adv. Mater.* 31(47), 1803230 (2019). <https://doi.org/10.1002/adma.201803230>
- Kour, R., Arya, S., Verma, S., Gupta, J., Bandhoria, P., Bharti, V., Datt, R., Gupta, V.: Potential substitutes for replacement of lead in perovskite solar cells: a review. *Glob. Chall.* 3(11), 1900050 (2019). <https://doi.org/10.1002/gch2.201900050>
- Kumar, N., Rani, J., Kurchania, R.: Advancement in  $\text{CsPbBr}_3$  inorganic perovskite solar cells: fabrication, efficiency and stability. *Sol. Energy* 221, 197–205 (2021). <https://doi.org/10.1016/j.solener.2021.04.042>
- Lakhdar, N., Hima, A.: Electron transport material effect on performance of perovskite solar cells based on  $\text{CH}_3\text{NH}_3\text{GeI}_3$ . *Opt. Mater.* 99, 109517 (2020). <https://doi.org/10.1016/j.optmat.2019.109517>
- Mahajan, P., Datt, R., Tsoi, W.C., Gupta, V., Tomar, A., Arya, S.: Recent progress, fabrication challenges and stability issues of lead-free tin-based perovskite thin films in the field of photovoltaics. *Coord. Chem. Rev.* 429, 213633 (2021). <https://doi.org/10.1016/j.ccr.2020.213633>
- Mandadapu, U., Vedanayakam, S.V., Thyagarajan, K.: Simulation and analysis of lead based perovskite solar cell using SCAPS-1D. *Indian J. Sci. Technol.* 10(11), 65–72 (2017). <https://doi.org/10.17485/ijst/2017/v11i10/110721>
- Minami, T., Miyata, T., Nishi, Y.:  $\text{Cu}_2\text{O}$ -based heterojunction solar cells with an Al-doped ZnO/oxide semiconductor/thermally oxidized  $\text{Cu}_2\text{O}$  sheet structure. *Sol. Energy* 105, 206–217 (2014). <https://doi.org/10.1016/j.solener.2014.03.036>
- Momblona, C., Gil-Escrig, L., Bandiello, E., Hutter, E.M., Sessolo, M., Lederer, K., Blochwitz-Nimoth, J., Bolink, H.J.: Efficient vacuum deposited pin and nip perovskite solar cells employing doped charge transport layers. *Energy Environ. Sci.* 9(11), 3456–3463 (2016). <https://doi.org/10.1039/C6EE02100J>
- Mushtaq, S., Tahir, S., Ashfaq, A., Bonilla, R.S., Haneef, M., Saeed, R., Ahmad, W., Amin, N.: Performance optimization of lead-free  $\text{MASnBr}_3$  based perovskite solar cells by SCAPS-1D device simulation. *Sol. Energy* 249, 401–413 (2023). <https://doi.org/10.1016/j.solener.2022.11.050>
- Nejand, B.A., Ahmadi, V., Gharibzadeh, S., Shahverdi, H.R.: Cuprous oxide as a potential low-cost hole-transport material for stable perovskite solar cells. *Chemsuschem* 9(3), 302–313 (2016). <https://doi.org/10.1002/cssc.201501273>
- Noel, N.K., Stranks, S.D., Abate, A., Wehrenfennig, C., Guarnera, S., Haghighirad, A.A., Sadhanala, A., Eperon, G.E., Pathak, S.K., Johnston, M.B., Petrozza, A.: Lead-free organic–inorganic tin halide perovskites for photovoltaic applications. *Energy Environ. Sci.* 7(9), 3061–3068 (2014). <https://doi.org/10.1039/C4EE01076K>
- Ompeng, D., Singh, J.: High open-circuit voltage in perovskite solar cells: the role of hole transport layer. *Org. Electron.* 63, 104–108 (2018). <https://doi.org/10.1016/j.orgel.2018.09.006>
- Park, N.G.: Perovskite solar cells: an emerging photovoltaic technology. *Mater. Today* 18(2), 65–72 (2015). <https://doi.org/10.1016/j.mattod.2014.07.007>

- Poorkazem, K., Liu, D., Kelly, T.L.: Fatigue resistance of a flexible, efficient, and metal oxide-free perovskite solar cell. *J. Mater. Chem. A* 3(17), 9241–9248 (2015). <https://doi.org/10.1039/C5TA00084J>
- Qiu, X., Cao, B., Yuan, S., Chen, X., Qiu, Z., Jiang, Y., Ye, Q., Wang, H., Zeng, H., Liu, J., Kanatzidis, M.G.: From unstable  $\text{CsSnI}_3$  to air-stable  $\text{Cs}_2\text{SnI}_6$ : a lead-free perovskite solar cell light absorber with bandgap of 1.48 eV and high absorption coefficient. *Sol. Energy Mater. Sol. Cells* 159, 227–234 (2017). <https://doi.org/10.1016/j.solmat.2016.09.022>
- Rai, S., Pandey, B.K., Dwivedi, D.K.: Modeling of highly efficient and low cost  $\text{CH}_3\text{NH}_3\text{Pb}(\text{I}_{1-x}\text{Cl}_x)_3$  based perovskite solar cell by numerical simulation. *Opt. Mater.* 100, 109631 (2020). <https://doi.org/10.1016/j.optmat.2019.109631>
- Sahli, F., Werner, J., Kamino, B.A., Bräuninger, M., Monnard, R., Paviet-Salomon, B., Barraud, L., Ding, L., Diaz Leon, J.J., Sacchetto, D., Cattaneo, G.: Fully textured monolithic perovskite/silicon tandem solar cells with 25.2% power conversion efficiency. *Nat. Mater.* 17(9), 820–826 (2018). <https://doi.org/10.1038/s41563-018-0115-4>
- Shi, B., Jia, J., Feng, X., Ma, G., Wu, Y., Cao, B.: Thermal evaporated CuI film thickness-dependent performance of perovskite solar cells. *Vacuum* 187, 110076 (2021). <https://doi.org/10.1016/j.vacuum.2021.110076>
- Wang, M., Zeng, P., Bai, S., Gu, J., Li, F., Yang, Z., Liu, M.: High-quality sequential-vapor-deposited  $\text{Cs}_2\text{AgBiBr}_6$  thin films for lead-free perovskite solar cells. *Sol. Rrl* 2(12), 1800217 (2018). <https://doi.org/10.1002/solr.201800217>
- Yamada, N., Ino, R., Ninomiya, Y.: Truly transparent p-type  $\gamma$ -CuI thin films with high hole mobility. *Chem. Mater.* 28(14), 4971–4981 (2016). <https://doi.org/10.1021/acs.chemmater.6b01358>
- Zhang, Z., Li, X., Xia, X., Wang, Z., Huang, Z., Lei, B., Gao, Y.: High-quality  $(\text{CH}_3\text{NH}_3)_3\text{Bi}_2\text{I}_9$  film-based solar cells: pushing efficiency up to 1.64%. *J. Phys. Chem. Lett.* 8(17), 4300–4307 (2017). <https://doi.org/10.1021/acs.jpclett.7b01952>
- Zong, Y., Zhou, Y., Zhang, Y., Li, Z., Zhang, L., Ju, M.G., Chen, M., Pang, S., Zeng, X.C., Padture, N.P.: Continuous grain-boundary functionalization for high-efficiency perovskite solar cells with exceptional stability. *Chem* 4(6), 1404–1415 (2018). <https://doi.org/10.1016/j.chempr.2018.03.005>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



# DVRGNet: an efficient network for extracting obscenity from multimedia content

Kamakshi Rautela<sup>1</sup> · Dhruv Sharma<sup>1</sup> · Vijay Kumar<sup>2</sup> · Dinesh Kumar<sup>1</sup>

Received: 3 August 2022 / Revised: 19 May 2023 / Accepted: 21 August 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

## Abstract

The availability of adult content on the internet through images or videos is easily accessible to minors and adults as well. In addition, this type of content may lead to poor mental health, sex-ism and objectification, and sexual violence. Therefore, It is extremely important to detect and classify pornographic content. In this paper, DVRGNet, a hierarchical CNN framework for the detection and classification of obscene content from videos is proposed. The proposed framework incorporates motion data and the capture of motion movement to deal with the problem of mapping skin exposure to pornographic content. DVRGNet is a network that leverages DenseNet, VGGNet, ResNet, and GoogLeNet for feature extraction. This network includes different fusions of various sub-networks, which can be seen as diverse tiers of neurons in human brains. The framework also incorporates a 5-layer Bi-LSTM-based classification of obscenity from videos. The proposed framework makes better use of automated pornography detection through computational intelligence architectures. Furthermore, the fusion of these four networks strengthens feature propagation by reducing the vanishing gradient problem. Extensive experiments are conducted on Pornography-2K and Pornography-800 datasets to validate the effectiveness of the proposed framework. The proposed framework achieves an accuracy of 99.42% on the Pornography-2K and 99.04% on the Pornography-800 datasets. An ablation study is also conducted to demonstrate the performance of proposed framework.

**Keywords** Convolution neural networks · DenseNet · GoogLeNet · Motion vectors · Pornography classification · ResNet · VGG

## 1 Introduction

We witness a lot of content being placed, and/or accessed on social media and the Internet everyday. This content may or may not contain useful information. The presence of illicit content

---

✉ Vijay Kumar  
vijaykumarchahar@gmail.com

<sup>1</sup> Electronics & Communication Engineering Department, Delhi Technological University, Delhi, India

<sup>2</sup> Information Technology Department, Dr B R Ambedkar National Institute of Technology, Jalandhar, Punjab, India

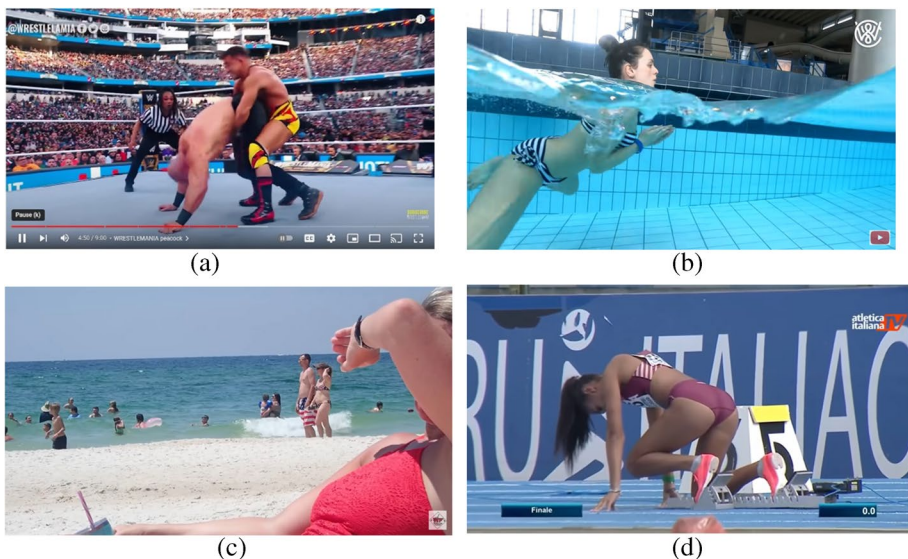


cannot be denied. This is of more concern to society at large when one notices the image of an infant's adult material. This may become a contentious problem because this market has the potential to incite more abuse and is an obvious assault on children's dignity by exposing them as sex products [1, 2]. The illicit access to pornographic content by the teenagers is caused by more than just curiosity. However, this may be an inappropriate form of self-expression. Teenagers may take it as a fashion statement and keep it going. The worst part could be that children may emulate this phenomenon as well. If no timely efforts are made to impose restriction on access to such content, this may jeopardize moral quality and character in the long run. In concrete terms, the number of criminal cases involving immoral acts will also increase [3].

Pornographic content can be spread through pornographic websites, social networking, e-mails, and live streaming platforms. This will undoubtedly have an impact on the video streaming platform's image. Because pornographic content is so easily accessible, service providers frequently keep on blocking the accounts that are suspected of revealing such action. The issue here is that the platform is unable to automatically filter such content, thereby making us think of an approach that may help block such content. So, the main aim is to prohibit certain demographics or surroundings from submitting or downloading inappropriate content [4].

A natural method for detecting pornography is to automatically detect nudity first, and then develop reasonable thresholds to further filter the data. Human skin traits, including color and texture, as well as human geometry, are widely used in such detection techniques [5–8]. These methods typically use the data to model the pixel values and spatial distribution of a nude person. But, the problem with these methods is that people may expose a lot of skin while participating in activities like wrestling, sunbathing, running, swimming, and similar ones, leading to numerous false positives. On the other hand, some sexual activities expose very little skin, resulting in undesirable false negatives [4]. So, pornography detection cannot be based solely on skin exposure. The examples of some cases are depicted in Fig. 1.

Many researchers have explored alternatives to low-level skin-based techniques for sexually explicit content filtering solutions in recent years. Some of them include word



**Fig. 1** Examples of video mistaken cases based solely on skin exposure (a) shows match between two wrestlers, (b) kids are swimming, (c) peoples are sunbathing at a beach, and (d) running

models from text classification, convolution neural network-based classification videos, and third-party solutions. Due to the complexity of developing suitable thresholds for skin-based detectors, numerous options are available for programming low and mid-level features. The absence of proper features related to motion downgrades the overall performance of deep learning systems. In addition to that, a complex neural network is suitable for classifying images that are hard to distinguish, but it could cause the model to overfit on simple images. Since humans process information by activating different levels of neurons depending on how difficult it is to recognize an object, we should use different models for different frames. We propose an integration of networks named DVRGNet to automatically gather static and motion-related deep representations directly from the data.

This paper make an effort to integrate previously trained models for image classification problems, including motion and information to support spatial and static data, in order to handle the aforementioned problems. The contributions of this paper are as follows:

- A literature review of various methods for combining motion and static data extracted from investigated videos.
- Our proposed image classification model, DVRGNet, works by utilizing sub-network modules with varying depths of network layers. These modules extract different levels of visual features from images and provide classification results for the corresponding images.
- DVRGNet can leverage the fusion of features extracted from DenseNet, VGGNet, ResNet, and GoogLeNet. This makes it simple and quick to classify the obscenity content in videos.
- Using an integrated network along with static and motion information, we aim to decrease the distance between the features of samples belonging to the same category and increase the distance between the features of samples from different categories.
- The proposed model also provides Bi-LSTM at the classification stage, which learns the image features from both backward and forward directions.
- Extensive experiments on the Pornography-800 and Pornography-2K datasets are performed. An ablation study is also carried out to justify the performance efficacy of the proposed DVRGNet.

The organization of this paper is as follows: The existing approaches for addressing issues associated with pornography detection are mentioned in Section 2. In Section 3, the proposed method for classifying pornography in videos using both static and motion data is briefly described. The computational complexity of the proposed approach is discussed in Section 4. Section 5 presents the experimental results and discussion. The concluding remarks are drawn in Section 6.

## 2 Related works

Skin exposure is at the epicenter of most pornographic content detection techniques. The use of CNN is a common method for classifying images [29]. It consists of the feature extraction and classification steps. The human skin is by far the most significant area of interest in detection. If the input includes an exceedingly massive area of skin, it is considered an important indicator of nudity. However, the skin may not be the most obvious factor that impacts the output [9]. It is complex to develop a suitable threshold for skin-based

detectors. As a result, the challenge is to discover the best method for comprehending video context. Video classification is widely studied in both computer vision and machine learning as one of the fundamental concepts for video understanding [10].

Avila et al. [12] developed a new pooling strategy to enhance the performance of the Bag-of-Words model. They developed a new representation for content description named BossaNova. BossaNova preserved the information about the distribution of local descriptors in codewords. BoosNova attained a 2.4% improvement over BOSSA [11] for video classification. Wehrmann et al. [9] proposed an adult content detector using the Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models. The features were extracted from videos by using CNN and prepared a set of semantic descriptors. These descriptors were applied in LSTM for classification, evaluated on the NPDI dataset, and attained the accuracy of 95.3%. However, the parameter tuning greatly affected the performance of the developed detector. Perez et al. [4] used CNN to extract both static and dynamic features from pornographic videos. The classification accuracy obtained from this method was 96.4%. Wang et al. [13] proposed a multimodal deep learning framework for detecting inappropriate contents in live video. The audio and visual features were extracted from CNN. These features were applied in Bidirectional Gated Recurrent Unit (Bi-GRU) to determine temporal context. Their approach was evaluated on the BJUTSD\_V2 dataset and attained the accuracy of 69.24%. However, it suffered from high computational complexity. Cheng et al. [14] utilized a Deep CNN (DCNN) to classify the images into three different classes. The local and global contexts were utilized for classification. Both AIC and NPDI datasets were used to validate the performance of the proposed approach. The classification accuracies obtained from this approach were 96.6% and 92.7% over AIC and NPDI datasets, respectively. Silva et al. [15] explored the Convolution3D based CNN architectures to encapsulate Spatio-temporal information in a single stream network using Conv3D layer, which learns Spatio-temporal embedding on training from the videos. This technique was used to detect the pornographic content.

The attention mechanism is frequently employed in image classification, object identification, natural language processing, and other fields due to its various benefits [23]. However, the use of attention mechanisms in the classification of obscene images is slowly growing. An innovative dot-product-based attention technique with 92.72% accuracy was proposed for pornography detection [24]. A unique visual attention mechanism called CBAM and Scale Constraint Pooling (SCP) were used to create a moderate CNN called DOCAPorn, which had an accuracy of 98.41%. It also minimized same-class fluctuation and optimized the distance between classes [25].

Gautam et al. [16] proposed a Frame Sequence ConvNet Pipeline that used ResNet-18 for feature extraction and ConvNet to analyze N frame feature-maps for frame sequence classification. Their accuracy rates for Pornography-800 and Pornography-2k datasets were 98.25% and 97.17%, respectively. Yousaf et al. [17] proposed a deep learning model EfficientNet-B7 for detection and classification of pornographic content in videos. A dataset of 111,156 YouTube cartoon clips that had been manually annotated was used, and the accuracy obtained was 95.66%. Further, deep learning-based methods have demonstrated excellent performance in detecting the abundance of pornographic images and videos on social media. Samal et al. [26] utilized transfer learning and feature fusion to identify pornographic images. Samal et al. [27] also developed an attention mechanism and a suitable pooling strategy to classify and label the obscene portion. Live broadcasting has enriched people's lives and become an indispensable part of their entertainment because of the quick development of online live streaming. Yuan et al. [28] proposed a deep-learning framework-based detection of pornographic content from live broadcasting.

There is a vast literature on detecting inappropriate video content with handcrafted or techniques based on deep learning feature extraction. These studies used binary classification to categories the videos as safe or risky. It does not, however, include research into detecting or classifying various types of discomforting content using motion features. Secondly, no studies have investigated the use of a hybrid network for enhanced feature extraction in the detection of inappropriate video content. This paper proposes the integration of four most widely used learning models to obtain better results.

### 3 Proposed work

The proposed approach classifies pornographic content by integrating different CNN learning models, as shown in Fig. 2. The framework starts with dividing a video clip into two sources, i.e., motion information and static information. The static information uses raw frames from video clips as an input, then filters and detects the needed information for classification. Whereas the motion detector uses the motion from the video clip to filter and detect useful information. The information, so detected, from both (motion information and static information) is then added and fed to the DVRG-Net, which is an integration of four different CNN models, i.e., DenseNet, VGG, ResNet and GooGleNet, used for image feature extraction and classification. The integration of these four networks results in strengthening feature propagation and elevating vanishing gradient problem. Further, this integration increases the training speed without increasing the error percentage. Also, for the classification of obscenity-based content from videos, 5-layers of Bi-LSTM are incorporated that enhances abstract visual features and provides higher recognition accuracy and efficiency. The unified model is trained on the Pornography-800 and Pornography-2k datasets, individually. Finally, the proposed network is evaluated using the test images.

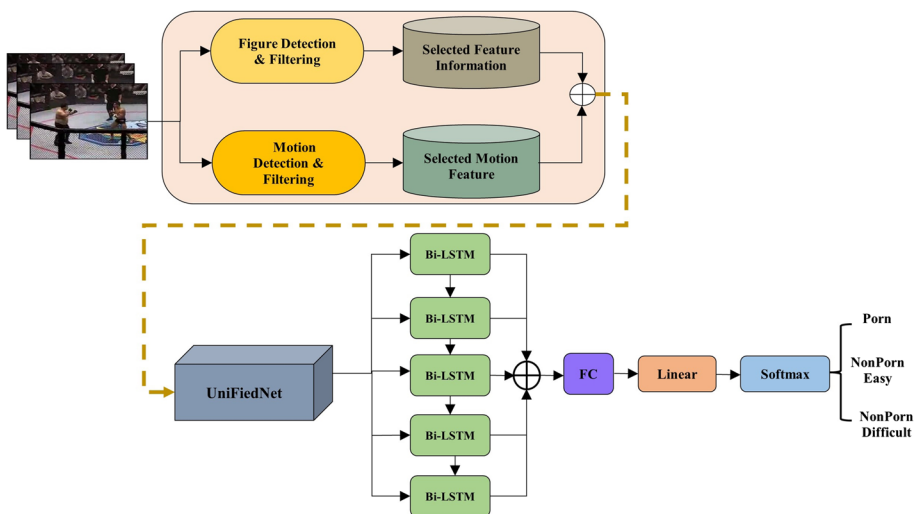


Fig. 2 Proposed obscenity detection framework

### 3.1 Dataset used

The dataset used in this work is developed by Avila et al. [12]. Previously, most of the experimentation was done on the Pornography-800 dataset. This dataset comprises both pornographic and non-pornographic videos, the latter of which is categorized into two classes namely, easy and difficult. It includes 77 h of video, 57 of which are pornographic, with the other hours being non-pornographic. This dataset presents a difficult classification problem because the videos involve people of various skin colors, ranging from dark skin to white complexions. Pornography-2k, on the other hand, is a larger version of the Pornography-800 dataset. There are 1000 pornographic and non-pornographic videos in its 140 h of content. The length of the video ranges from 6 s to 33 min. This dataset is more difficult to analyze because it includes a wide range of photo styles, including a variety of cartoons, genres, diverse behavior patterns, and ethnicity.

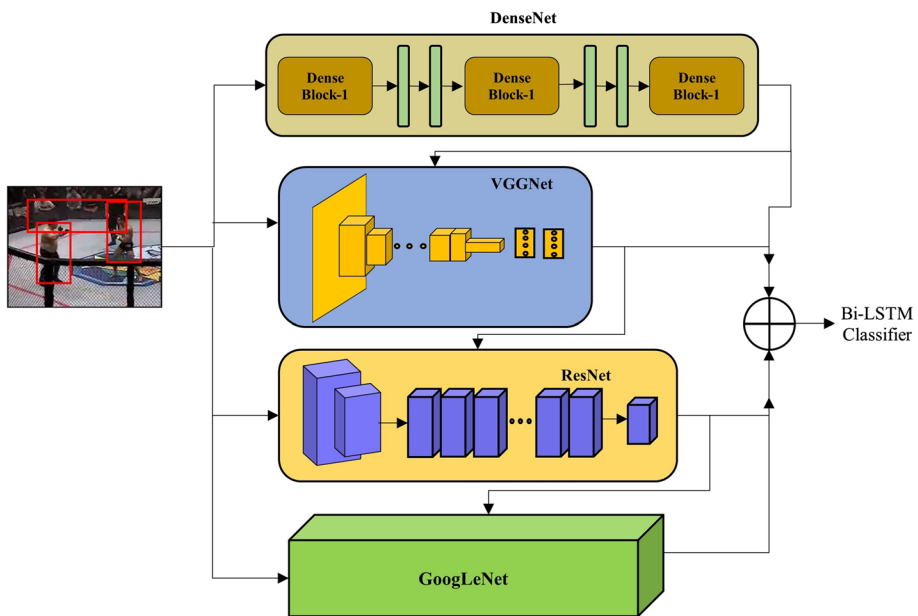
### 3.2 Data preprocessing

Prior to feeding into the proposed DVRGNet, the images underwent a number of preprocessing steps. The dataset contains videos that are divided into two components namely, video and frames. To accommodate the input structure of a DVRGNet, the initial video frames are used to detect the actions of the figures, and motions are detected with the help of motion vector data obtained from the videos. Each vector for a specific frame contains an encoded representation of the locations of a given macro-block of pixels in both the current frame and the reference frame.

Let  $I_m(a, b)$  be intensity representation of an image at  $(a, b)$  and  $I_T$  be a complete image at time  $T$ , which is split into micro-blocks  $B_m[n](a', b')$ . The velocity vector  $v = (v_1, v_2)^T$  is used to describe motion in continuous images. Here,  $v(z)$  is velocity at  $z$  position. Then, the distance in pixel coordinates between the current and reference positions is computed independently in both the vertical and horizontal directions to see how far the blocks have moved using each motion vector. Furthermore, these distances are proportional to the magnitude of movement in the micro-block region, resulting in two motion maps, one per direction. The information from both the blocks is then fed to proposed DVRGNet. In this proposed model, we merge multiple CNNs in accordance with the fundamentals of ensemble learning in order to produce a robust image classification model.

### 3.3 DVRGNet

DVRGNet is an aggregation of multiple CNNs for a strong image classification model. The proposed DVRGNet model, as shown in Fig. 3, is a network that leverages four well-known CNN models namely, DenseNet, VGG, ResNet and GoogLeNet, respectively. These models extract features from preprocessed images separately and then integrate them for classification using a fully connected layer. High-level functionality is represented by these integrated parameters. The proposed framework "DVRGNet" has 11,177,538 parameters after combining all of the features, which is nearly three times more than the independent DenseNet, VGG, ResNet and GoogLeNet network, respectively. Our approach, referred to as DVRGNet, is distinct from conventional ensemble learning methods, which frequently employ identical sub-networks that are



**Fig. 3** Internal architecture of proposed DVRGNet

trained and tested indiscriminately. Instead, DVRGNet is made up of numerous different CNN sub-networks that are trained sequentially and progressively. Also, a deeper network (like DVRGNet) is beneficial for the recognition of obscene scenes. The subsections that follow include details on the basic structure of each CNN network that has been adopted as well as the process of fine-tuning.

### 3.3.1 DenseNet

The DenseNet network was created by Huang et al. [18] in 2017 and has the best classification accuracy on the CIFAR100, ImageNet, and CIFAR-10 datasets. This network is based on the ResNet architecture, where each layer is connected to others using a feed forward. This connection allows the network to communicate vital information internally, improving network performance and enhancing network training [19].

### 3.3.2 VGGNet

In 2014 at ILSVRC competition, VGGNet [20] network was first designed for image localization before being used for classification. When compared to the AlexNet architecture, this network performed exceptionally well, with an error rate of only 8.1%. We use VGG as the feature extractor in this study. Here, the first two blocks consist of four convolution layers, while the remaining nine convolution layers are located in the next three blocks. Further, each block is followed by a max-pooling layer.



### 3.3.3 ResNet

Simonyan and Zisserman [21] introduced a skip connection architecture known as the Residual Network. The important thing to note is that each pair of filters now has a shortcut connection. For all shortcuts, ResNet uses identity mapping and for increasing dimensions, zero-padding.

### 3.3.4 GoogLeNet

The main goal of the Inception architecture is to determine whether the optimal local sparse structure of a convolutional vision network can be estimated and covered by commonly available dense components [22].

## 3.4 Fine tuning of DVRGNet

Here, CNN models for categorizing pornographic videos are integrated using different fully connected layers of all the four networks. After extracting the features one at a time, all of the networks use GlobalAveragePooling2D to flatten all of the layers into a vector by computing the mean value for each of the source channels at the same time. The concatenate layer is then used to combine all of the individual vectors into a single vector. Following that, six layers are used to fine-tune the integrated features for classification, which is followed by the activation function softmax. The descriptions of each layer are listed below.

In the proposed classification model, we use four batch normalization layers, each of which is crucial. All of the data is re-scaled by the batch normalization layer so that we can normalize it. During the training phase, images are assigned weights to reflect the level of difficulty in correctly classifying them by the model. If a sub-network cannot accurately classify an image, the weight assigned to that image will be increased. The rescaled data aids in the training phase and reduces network initialization sensitivity. The gradient descent loss and optimizer function are the two key hyperparameters for training a model. Adam optimizer is used that combines the features of RMSProp and AdaGrad, thereby, allowing it to handle sparse gradients on large amounts of data. Further, each previous and current layers' neurons is connected with the dense layer to process the data and produce a result. In this case, four dense layers are used, with the last dense layer performing the classification task, followed by the activation function. This layer will make a prediction based on the length of the prediction class. The activation function determines which features are most closely related to the predicted class based on the outcome probability. The outcome value for the softmax activation function ranges from 0 to 1. It is defined mathematically as:

$$\text{Soft max}(s)_i = \frac{\exp(s_i)}{\sum_{m=1}^n \exp(s_m)} \quad (1)$$

The detailed architecture of the proposed DVRGNet is given in the Table 1. This model is made up of eleven layers that can be learned, each of which has different

**Table 1** Architecture of DVRGNet

Type	Layer	Parameters	Output Size
Input	Layer 1	0	$64 \times 64 \times 112 \times 112$
DenseNet	Layer 2	3,147,256	$64 \times 64 \times 56 \times 56$
VGG	Layer 3	2,269,716	$64 \times 128 \times 28 \times 28$
ResNet	Layer 4	3,347,296	$64 \times 64 \times 56 \times 56$
GoogLeNet	Layer 5	1,179,648	$64 \times 256 \times 14 \times 14$
(AdaptiveAvg) Pooling	Layer 6	0	$64 \times 512 \times 1 \times 1$
Concatenate	Layer 7	0	$64 \times 512 \times 1 \times 1$
Flatten	Layer 8	0	$64 \times 2$
Dense (FC)	Layer 9	1,155,340	$64 \times 2$
Dropout	Layer 10	0	$64 \times 2$
Softmax	Layer 11	120,230	3

Total parameters: 11,177,538

Trainable parameters: 11,177,538

Non-trainable parameters: 0

parameters and different output sizes. During the backpropagation process, the DVRGNet is trained with approximately eleven million parameters.

During the training phase, images are assigned weights to reflect the level of difficulty in correctly classifying them by the model. A sub-network's weight will be increased if it is unable to correctly classify an image. The images are then sent into the next sub-network to extract more complex and effective visual features. This sub-network has been given new weights based on the difficulty of the classification task. Assume that DVRGNet consists of  $S$  distinct sub-networks that are successively trained. Each image sample has its weight for the specific sub-networks.  $Z_1, Z_2, \dots, Z_4$  represent the image weights for the sub-networks, respectively. Each sub-network combines the results of the preceding sub-networks to make decisions. Let  $W_i^s$  be the weight of  $i^{\text{th}}$  image for  $s^{\text{th}}$  sub-network, and  $Z^s = [W_1^s, W_2^s, W_3^s, \dots, W_t^s]$ . Here,  $i \in \{1, 2, \dots, t\}$ ,  $s \in \{1, 2, \dots, S\}$ , while  $t$  is the total number of images in the training dataset.

Initially, the weights for all the images are set and initialized. Then, these training images along with their respective initial weights are fed into the first sub-network, which is DenseNet with  $s = 1$ , to train the model.

$$Z^1 = [W_1^1, W_2^1, W_3^1, \dots, W_t^1] \quad (2)$$

The weights for the first sub-network, denoted as  $W_i^1$ , are set to  $1/n$  for  $i = \{1, 2, \dots, t\}$ . The first sub-network is then trained for several iterations using gradient descent to update its parameters. After the training is completed, the sub-network is capable of making predictions.

$$p_i^s = M^s(x_i), \quad (3)$$

In the equation,  $M^s(\cdot)$  refers to the  $m$ -th sub-network, and  $p_i^s$  is the predicted label for the  $i^{\text{th}}$  sample by  $M^s$ . Then, we identify the samples for which  $p_i^s$  is not equal to its true label  $n_i$ . Using the following equation, we calculate the weighted error rate ( $\epsilon^s$ ) of the  $s^{\text{th}}$  sub-network  $M^s(\cdot)$  for all the selected samples in the training set.

$$\varepsilon^s = \sum_{i_a}^{N_{i_a}} W_{i_a}^s, \quad (4)$$

In the equation,  $W_{i_a}^s$  represents the weight of the  $i_a^{th}$  selected sample for  $M^s(\cdot)$ , and  $N_{i_a}$  is the total number of selected samples. Then, we use  $\varepsilon^s$  to determine the weight coefficient  $\alpha^s$  of  $M^s$ , which represents the significance of  $M^s$  in DVRGNet.

$$\alpha^s = \frac{1}{2} \log \frac{1 - \varepsilon^s}{\varepsilon^s}, \quad (5)$$

Equation (5) indicates that  $\alpha^s$  is inversely related to  $\varepsilon^s$ , suggesting that the sub-network with the lower error rate  $\varepsilon^s$  will have a higher significance coefficient value in the entire DVRGNet. Moreover,  $\alpha^s$  is utilized to adjust the weights of the samples for training the subsequent sub-network. We have for the images that meet the condition  $p_i^s = n_i$ ,

$$W_i^{s+1} = W_i^s \exp(-\alpha^s), \quad (6)$$

Otherwise,

$$W_i^{s+1} = W_i^s \exp(\alpha^s), \quad (7)$$

Then,

$$Z^{s+1} = [W_1^{s+1}, W_2^{s+1}, W_3^{s+1}, \dots, W_t^{s+1}] \quad (8)$$

Consequently, the weights of the images for the subsequent sub-network are decreased when the predicted results  $p_i^s$  closely match the true labels  $n_i$  of the images. Conversely, the weights of the images are increased if there is a discrepancy between the predicted results and the true labels. The next sub-network is trained iteratively for multiple rounds using the updated weights of the image samples.

### 3.5 Obscenity classification network

This section discusses the Bi-LSTM-based classification. The input of Bi-LSTM is the feature vector extracted from DVRGNet module. The extracted features are learned via five layers of Bi-LSTM, and each layer's extracted learned features are concatenated. The input is divided into three categories namely, vPorn, VPorn Easy, and vPorn Difficult by passing the concatenated features via the FC, linear layer, and softmax layer.

## 4 Computational complexity

This section describes the computational complexity of the proposed DVRGNet. The space and time complexities are explained below:

### 4.1 Time complexity

- The initialization of proposed model needs  $O(V_F(g \times h)) + O(V_{VD}(g \times h))$  time. Where  $V_F$  represents the image frame captured from videos.  $V_{VD}$  represents the video for motion detection.  $g$  and  $h$  are the number of rows and columns, respectively.

**Input:** Training set ( $\alpha$ ), Test set ( $\chi$ ), Validation set ( $\gamma$ )

Learning rate ( $\mathfrak{R}$ ) = 0.001

Epochs ( $\varepsilon$ ) = 50

Batch size ( $\beta$ ) = 64

Images in 1 batch ( $\kappa$ )

**Output:** Model Accuracy ( $Ac_{epoch}$ ), Loss ( $L_{epoch}$ ), Precision ( $P_{epoch}$ ), Recall ( $R_{epoch}$ )

**Begin:** Each frame in the training set to be converted to 64 x 64

**Extract Features:**

```

for epoch 0→50
    for all  $\alpha$  ←feature extraction
    end for

```

**Train Classifier:**

```

X→ np.array( $\alpha$ )
Y→Transform
Model→ Train classifier (X,Y)
end for
for epoch 0→50
    for all  $\gamma$  ←feature extraction
    end for

```

**Validate:**

```

X→ np.array( $\gamma$ )
Y→ Validation
Model→ Test classifier (X,Y)
end for

```

**Algorithm 1.** Automated obscenity detection and classification

- The information from  $V_F$  and  $V_{VD}$  are merged together and requires  $O(V_{F+VD}(g \times h))$ .
- DVRGNet is an integrated network combining DenseNet, VGG, ResNet and GooGleNet, for achieving efficient detection and classification.

$$O(U_{F+VD}) = O[(D_{F+VD}(g \times h)) + (VGG_{F+VD}(g \times h)) + (R_{F+VD}(g \times h)) + (G_{F+VD}(g \times h))]$$

Here,  $D_{F+VD}$ ,  $VGG_{F+VD}$ ,  $R_{F+VD}$ , and  $G_{F+VD}$  are the outputs of DenseNet, VGG, ResNet and GooGleNet, respectively.

- Time taken by training and testing of the proposed model is

$$O_{TT} = O(V_{F+VD}(g \times h)) + O(V_{F+VD}(Train)) + O(V_{F+VD}(Test))$$

- Therefore, the overall time complexity of the proposed model is

$$O(U_{F+VD}) + O(V_{F+VD}(Train)) + O(V_{F+VD}(Test))$$

## 4.2 Space complexity

The proposed network space complexity is considered during the initialization. The overall space complexity of the proposed approach is as given below:

$$O(V_F(g \times h)) + O(V_{VD}(g \times h)) + O(V_{F+VD}(Train)) + O(V_{F+VD}(Test))$$

## 5 Results and discussions

This section presents and discusses the outputs from experimental evaluations of various CNN model approaches and the proposed DVRGNet for video obscenity detection and classification.

### 5.1 Evaluation metrics

The DVRGNet model's accuracy, precision, recall, and loss are calculated, as well as those of individual networks whose merger produced DVRGNet. Accuracy is calculated as the ratio of the percentage of positive predictions for every class to the total number of predictions for all classes [4]:

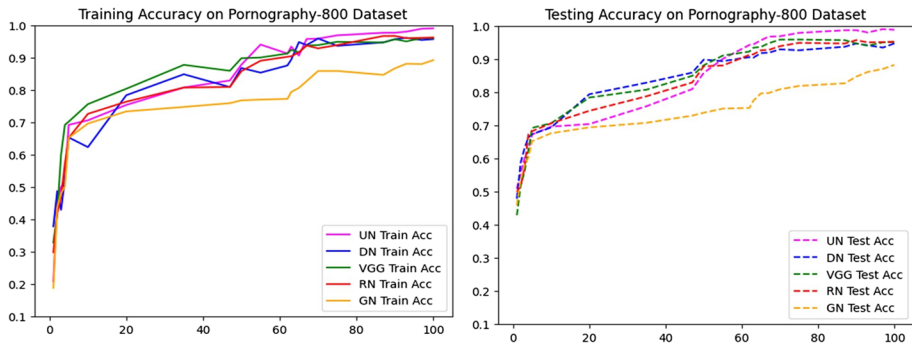
$$Ac = \frac{T_p + T_n}{T_p + F_p + F_n + T_n} \quad (9)$$

Here, Ac represents the accuracy.  $T_p$  and  $F_p$  depict the true and false positives, respectively.  $F_n$  and  $T_n$  denote the false and true negatives, respectively. The ratio of the total number of correct positive to the total number of positive predictions is known as precision. Sensitivity (Recall) is the ratio of the number of correct positive to the total number of predictions in actual class. The mathematical representations of precision (P) and recall (R) can be rewritten as [4, 5]:

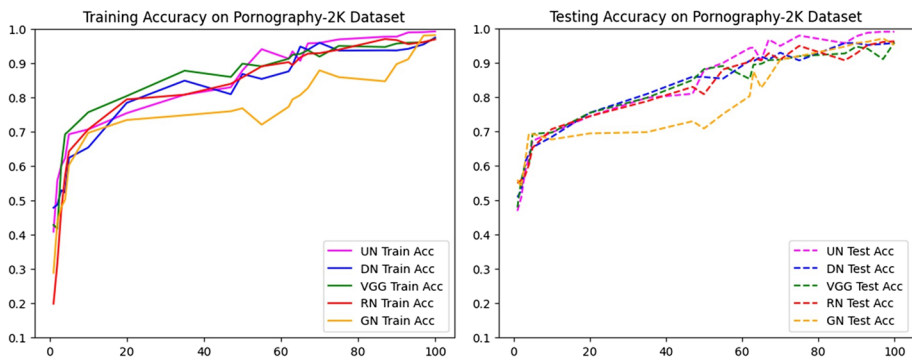
$$P = \frac{T_p}{T_p + F_p} \quad (10)$$

$$R = \frac{T_p}{T_p + F_n} \quad (11)$$

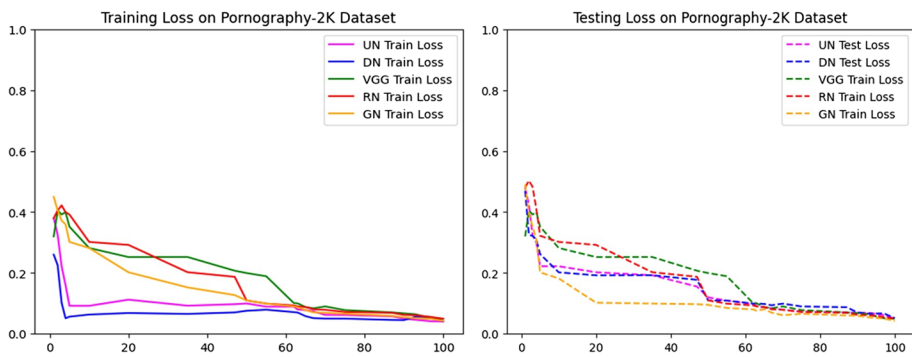
Figures 4 and 5 represent the training and testing accuracy curves for the proposed DVRGNet model on the Pornography-800 and Pornography-2K datasets. Additionally, the curves provide comparisons of the proposed model with DenseNet, VGGNet, GoogLeNet, and ResNet respectively. On the Pornography-2K dataset, training accuracies obtained from DVRGNet, DenseNet, VGG, ResNet, and GooGleNet are 99.04%, 95.7%, 96%, 96.2%, and 89.2%, respectively. On the Pornography-800 dataset, the training



**Fig. 4** Training and testing accuracy curves on pornography-800 dataset



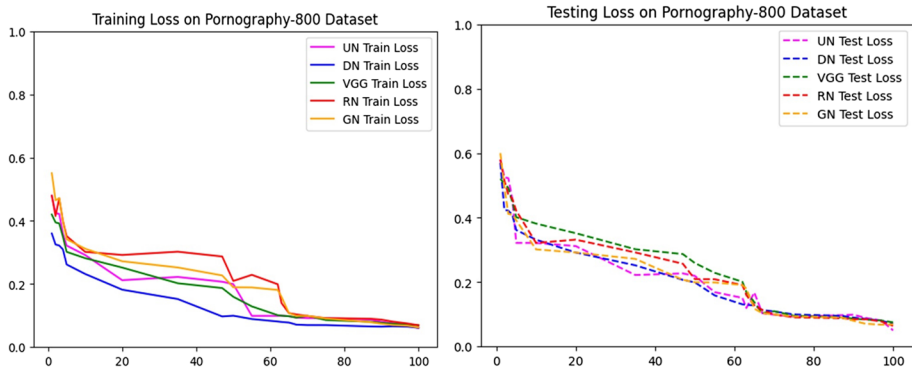
**Fig. 5** Training and testing accuracy curves on pornography-2K dataset



**Fig. 6** Training and testing loss curves on pornography-2K dataset

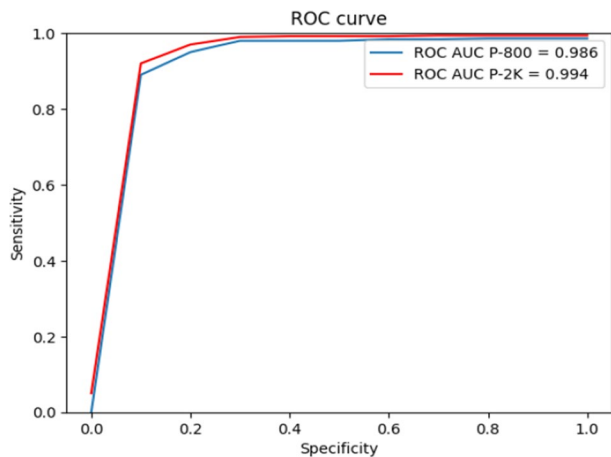
accuracies for the corresponding networks are 99.42%, 97.4%, 97%, 96.9% and 98.1%, respectively. Similarly, Figs. 6 and 7 show the losses corresponding to DVRGNet, DenseNet, VGG, ResNet and GoogLeNet for the Pornography-800 and Pornography-2k datasets, respectively. Also, Fig. 8 shows the ROC-AUC curve, which provides the sensitivity to specificity ratio at each threshold.





**Fig. 7** Training and testing loss curves on pornography-800 dataset

**Fig. 8** ROC-AUC curve for pornography-2K and pornography-800 dataset



## 5.2 Comparison with state-of-the-art

Table 2 shows the comparative analysis of the proposed obscenity detection model and other existing models. It is evident from Table 2 that the proposed obscenity detection model was able to provide the optimal results over Pornography-800 and Pornography-2K datasets. The proposed model aggregates four CNN models and Bi-LSTM layers to form a strong image classification model. In addition, the proposed model has been compared with six well-known techniques. Perez et al. [4] utilized CNN-based architecture with LSTM model to achieve 95.3% accuracy on the NDPI dataset. Avila et al. [12] used Bag-of-Words method and attained the accuracy of 88.6% on the NDPI dataset. Wehrmann et al. [9] and Yousaf and Nawaz [17] utilized CNN-based architectures and attained the accuracy of 96.4% and 98.25%, respectively. In addition, Gautam et al. [16] used ResNet-18 to extract image features and encapsulated the motion information, obtaining accuracy of 97.15% and 96% using 8-frame and 16-frame sequence, respectively, which is enhanced by the proposed framework.

**Table 2** Comparison Results of Proposed Framework with other State-of-the-art

Ref	Year	Model used	Dataset	Performance
[12]	2013	Bag-of-Words	NPDI	Acc = 88.6%
[4]	2017	CNN, LSTM	NPDI	Acc = 95.3%
[9]	2018	CNN	Pornography-2K, Pornography-800	Acc = 96.4%
[13]	2020	CNN, Bi-GRU	BJUTSD_V2	Acc = 69.2%
[15]	2019	DCNN	AIC, NPDI	Acc = 96.6% (AIC) Acc = 92.7% (NPDI)
[16]	2022	CNN, ConvNet	Pornography-2K, Pornography-800	Acc = 98.25% (P-800) Acc = 97.15% (P-2 K)
[17]	2022	EfficientNet-B7	YouTube cartoon	Acc = 95.66%
Proposed		DVRGNet	Pornography-2K, Pornography-800	Acc = 99.04% (P-800) Acc = 99.42% (P-2 K)

### 5.3 Ablation studies

An ablation study is carried out to demonstrate the influence of the proposed obscenity detection framework. Four CNN models are individually updated with a B-LSTM-based classification model, and DVRGNet is created by combining these models. Table 3 depicts the accuracy, precision and recall for the Pornography-800 and Pornography-2K datasets. We began the investigation by employing the individual networks, and analyzed the accuracies of DenseNet, ResNet, VGGNet, and GoogleNet. Further, the merging of two models results in improvements in the accuracy of the Pornography-800 and Pornography-2K datasets.

The Pornography-800 and Pornography-2K datasets showed an accuracy of 99.04% and 99.42%, respectively, after they were combined to improve detection and classification performance. Furthermore, it is clear from the findings listed in Table 3 that the proposed framework (DVRGNet) provided the best results in terms of accuracy, precision, and recall for both datasets. Therefore, the classification accuracy, precision and recall are improved by integrating four CNN models.

**Table 3** Ablation study conducted on Pornography datasets using different deep learning models

Model	Pornography-800			Pornography-2K		
	Acc	P	R	Acc	P	R
DVRGNet	99.04%	92.61%	93.22%	99.42%	93.54%	93.88%
DenseNet	95.97%	82.39%	90.19%	97.4%	85.59%	92.6%
VGG	96%	80.91%	89.98%	97%	81.36%	91.92%
ResNet	96.2%	82.36%	92%	96.9%	83.22%	93.67%
GooGleNet	89.2%	79.81%	89.1%	98.1%	81.61%	91.28%
DenseNet + VGG	95.21%	89.99%	91.19%	92.66%	87.62%	90.92%
ResNet + GooGleNet	94.08%	88.51%	90.00%	95.86%	89.55%	89.53%
VGG + ResNet + GooGleNet	96.44%	90.02%	91.14%	95.94%	91.21%	92.44%

## 6 Conclusions

In this paper, DVRGNet with a Bi-LSTM-based classification module is proposed for the detection and classification of obscenity in videos. This network employed motion detection with the help of motion vector data, which strengthened feature propagation and effectively solved the vanishing gradient problem. Also, the proposed DVRGNet incorporates four CNNs networks (i.e., DenseNet, VGGNet, ResNet, and GoogLeNet) as sub-networks. These sub-networks are trained sequentially in a progressive manner to enhance the performance of the model. Additionally, a Bi-LSTM-based classification module has been incorporated into the framework to improve its overall accuracy. The proposed framework is trained and tested on Pornography-800, and Pornography-2K datasets that offer accuracy of 99.04% and 99.42%, respectively. The performance of the proposed framework was superior to the other existing methods. Furthermore, the ablation study was conducted to validate the performance of the proposed framework. In future, more similar models may be concatenated in order to increase recognition accuracy and efficiency, depending on the precise visual task being performed and the complexity of the classification model.

**Data availability** Data sharing not applicable to this article as no datasets were generated during the current study.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

**Competing interest** The authors declare that they have no competing interests.

## References

1. Shojae Chaeikar S, Zamani M, Abdul Manaf AB, Zeki AM (2018) PSW statistical LSB image steganalysis. *Multimed Tools Appl* 77(1):805–835
2. Karamizadeh S, Shojae Chaeikar S, Jolfaei A (2023) Adult content image recognition by Boltzmann machine limited and deep learning. *Evol Intel* 16:1185–1194. <https://doi.org/10.1007/s12065-022-00729-8>
3. Dines G (2017) Growing up with porn: the developmental and societal impact of pornography on children. *Dignity* 2(3):3
4. Perez M, Avila S, Moreira D, Moraes D, Testoni V, Valle E, ... Rocha A (2017) Video pornography detection through deep learning techniques and motion information. *Neurocomputing* 230:279–293
5. Jones MJ, Rehg JM (2002) Statistical color models with application to skin detection. *Int J Comput Vision* 46(1):81–96
6. Rowley HA, Jing Y, Baluja S (2006) Large-scale image-based adult-content filtering. 1st International Conference on Computer Vision Theory
7. Lee S, Shim W, Kim S (2009) Hierarchical system for objectionable video detection. *IEEE Trans Consum Electron* 55(2):677–684
8. Bouirouga H, Elfkihi S, Jilbab A, Aboutajdine D, El Fkihi S (2012) Skin detection in pornographic videos using threshold technique. *J Theor Appl Inf Technol* 35(1):7–19
9. Wehrmann J, Simões GS, Barros RC, Cavalcante VF (2018) Adult content detection in videos with convolutional and recurrent neural networks. *Neurocomputing* 272:432–438
10. Jiang YG, Wu Z, Wang J, Xue X, Chang SF (2017) Exploiting feature and class relationships in video categorization with regularized deep neural networks. *IEEE Trans Pattern Anal Mach Intell* 40(2):352–364

11. Avila S, Thome N, Cord M, Valle E, de A. Araújo A (2011) BOSSA: Extended bow formalism for image classification. In: 2011 18th IEEE International Conference on Image Processing, Brussels, pp 2909–2912. <https://doi.org/10.1109/ICIP.2011.6116268>
12. Avila S, Thome N, Cord M, Valle E, Araújo ADA (2013) Pooling in image representation: the visual codeword point of view. *Comput Vis Image Underst* 117(5):453–465
13. Wang L, Zhang J, Wang M, Tian J, Zhuo L (2020) Multilevel fusion of multimodal deep features for porn streamer recognition in live video. *Pattern Recogn Lett* 140:150–157
14. Cheng F, Wang SL, Wang XZ, Liew AWC, Liu GS (2019) A global and local context integration DCNN for adult image classification. *Pattern Recogn* 96
15. da Silva MV, Marana AN (2018) Spatiotemporal CNNs for pornography detection in videos. In: Iberoamerican Congress on Pattern Recognition, pp 547–555. Springer International Publishing, Cham
16. Gautam N, Vishwakarma DK (2022) Obscenity detection in videos through a sequential convnet pipeline classifier. In: *IEEE Transactions on Cognitive and Developmental Systems* 15(1):310–318
17. Yousaf K, Nawaz T (2022) A deep learning-based approach for inappropriate content detection and classification of youtube videos. *IEEE Access* 10:16283–16298
18. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708
19. Khan SI, Shahrir A, Karim R, Hasan M, Rahman A (2022) MultiNet: a deep neural network approach for detecting breast cancer through multi-scale feature fusion. *Journal of King Saud University-Computer and Information Sciences* 34(8):6217–6228
20. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*
21. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 770–778
22. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 1–9
23. Guo MH, Xu TX, Liu JJ, Liu ZN, Jiang PT, Mu TJ, Zhang S-H, Martin RR, Cheng M-M, Hu S-M (2022) Attention mechanisms in computer vision: a survey. *Comput Vis Media* 8:1–38
24. Gangwar A, González-Castro V, Alegre E, Fidalgo E (2021) AttM-CNN: attention and metric learning based CNN for pornography, age and child sexual abuse (CSA) detection in images. *Neurocomputing* 445:81–104
25. Chen J, Liang G, He W, Xu C, Yang J, Liu R (2020) A pornographic images recognition model based on deep one-class classification with visual attention mechanism. *IEEE Access* 8:122709–122721
26. Samal S, Nayak R, Jena S et al (2023) Obscene image detection using transfer learning and feature fusion. *Multimed Tools Appl*. <https://doi.org/10.1007/s11042-023-14437-7>
27. Samal S, Zhang Y-D, Gadekallu TR, Nayak R, Balabantaray BK (2023) SBMYv3: improved MobYOLOv3 a BAM attention-based approach for obscene image and video detection. *Expert Systems* e13230
28. Huang C, Yuan C, Zhang J (2020) Violation detection of live video based on deep. Learning. <https://doi.org/10.1155/2020/1895341>
29. Aggarwal K, Mijwil MM, Al-Mistarehi AH, Alomari S, Gök M, Alaabdin AMZ, Abdulrhman SH (2022) Has the future started? The current growth of artificial intelligence, machine learning, and deep learning. *Iraqi J Comput Sci Math* 3(1):115–123

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



**Kamakshi Rautela** , received M.Tech. degree from Graphic Era Hill University, Bhimtal, India, in 2017. She is currently pursuing the Ph.D. degree with the Department of Electronics and Communication Engineering from Delhi Technological University, New Delhi, India. Her current research interest includes machine learning, deep learning, computer vision, and medical image processing. She is also a reviewer in Computers in Biology and Medicine.



**Dhruv Sharma** , received his M.Tech. degree from Ambedkar Institute of Advanced Communication Technology & Research, New Delhi, India, in 2017. He is currently pursuing the Ph.D. degree with the Department of Electronics and Communication Engineering from Delhi Technological University, New Delhi, India. His current research interest includes machine learning, deep learning, computer vision, natural language processing and image captioning.

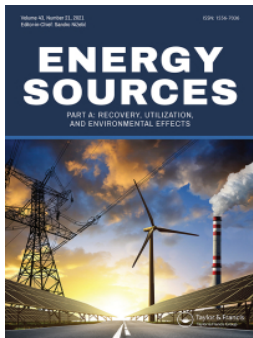


**Dr. Vijay Kumar** is Associate Professor in Information Technology Department, Dr B R Ambedkar NIT Jalandhar, Punjab. He received his Ph.D. degree from NIT Kurukshetra. Previously, he received M.Tech. and B.Tech. degrees from GJUS&T, Hisar and Kurukshetra University Kurukshetra, respectively. He has 16 years of teaching and research experience in various reputed institutes namely, NIT Hamirpur, TIET Patiala, and Manipal University Jaipur. He completed 2 DST SERB and 01 CSIR sponsored research projects. He has published more than 160 research papers in International Journals/Conferences. He has many book chapters in international reputed publishers. He has supervised many Ph.D. and M.Tech. thesis on Metaheuristics, Image Mining, and Data Clustering. He is the reviewer of several reputed SCI journals. He is member of ACM, CSI, International Association of Engineers, International Association of Computer Science and Information Technology, Singapore. His current research area is Soft Computing, Data Mining, Deep Learning, Steganography, and Pattern Recognition.



**Prof. Dinesh Kumar** received his B.Tech. (ECE) and M.Tech.(ECE) (Gold Medalist) from National Institute of Technology (NIT) Kurukshetra (erstwhile REC Kurukshetra) and received Ph.D. degree in the field of Biometrics (Face Recognition) from GGS Indraprastha University, Delhi. He has more than 30 years of teaching and research experience and currently working as Professor in the Electronics and Communication Engineering Department of Delhi Technological University, Delhi. He has research publications in National/International Journals and Conferences including IEEE, Elsevier, Springer, Wiley. His current research interests include image processing, biometrics, soft computing, clustering, biomedical imaging, machine learning, image captioning. He is on the panel of reviewers for IEEE, Elsevier and Springer journals/Transactions.





## Effect of partial shading on photovoltaic systems performance and its mitigation techniques-a review

Nikhil Kushwaha, Vinod Kumar Yadav & Radheshyam Saha

**To cite this article:** Nikhil Kushwaha, Vinod Kumar Yadav & Radheshyam Saha (2023) Effect of partial shading on photovoltaic systems performance and its mitigation techniques-a review, Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 45:4, 11155-11180, DOI: [10.1080/15567036.2023.2254731](https://doi.org/10.1080/15567036.2023.2254731)

**To link to this article:** <https://doi.org/10.1080/15567036.2023.2254731>



Published online: 06 Sep 2023.



Submit your article to this journal [↗](#)



Article views: 23




View related articles [↗](#)



View Crossmark data [↗](#)



# Effect of partial shading on photovoltaic systems performance and its mitigation techniques-a review

Nikhil Kushwaha , Vinod Kumar Yadav, and Radheshyam Saha

Department of Electrical Engineering, Delhi Technological University, Delhi, India

## ABSTRACT

The installed capacity of photovoltaic (PV) systems is increasing at an exponential rate around the world because it has the potential to meet the ever-increasing demand for energy and simultaneously mitigate the climate change crisis. Sustained investment in this energy sector over the last two decades has enabled researchers to introduce innovations in all related aspects, including maximizing cell efficiency, optimizing manufacturing processes, building public opinion, and project financing. These advancements have made PV technology the most affordable energy technology globally. However, PV technology faces some inherent technical challenges that diminish its effectiveness in providing green energy leading to a lower scale of decarbonization. One of these challenges is the premature failure of PV modules due to a phenomenon called a hot spot under partial shading. Research shows that PV cells may potentially undergo reverse breakdown under partial shading conditions, leading to temperatures of up to 400°C. Such high temperatures not only reduce PV performance but also cause irreversible damage and premature module failure, and even fire in extreme cases. The extent of power output reduction depends on the shading pattern on a PV system, irradiation, geographical location, and time of the day. For example, a single shaded cell in a module can cause a power loss of up to 50%, while multiple shaded cells can lead to a reduction of up to 90%. On average, partial shading can cause a power loss of 10–15% in a PV system. In this paper, a comprehensive review on the theoretical background of reverse breakdown mechanisms in PV cells/systems and various techniques to mitigate the effects of partial shading has been carried out with an exhaustive literature survey. As of the current date, researchers have suggested using module-level power electronics (MLPEs) to increase the energy yield of shaded PV systems by 5–25%, depending on the shading conditions and the type of MLPE technology. Nevertheless, the use of maximum power point tracking (MPPT) can enhance the efficiency of shaded PV systems is proposed to have augmented up to 30%.

## ARTICLE HISTORY

Received 7 February 2023  
Revised 28 August 2023  
Accepted 29 August 2023

## KEYWORDS

Hot spot; partial shading; photovoltaics; reverse breakdown; MPPT

## Introduction

The use of solar energy through photovoltaic (PV) systems is rapidly increasing worldwide due to its affordability, quick installation, and abundant solar resources. Currently, the dominant PV technology is crystalline-based PV cells, which make up about 90% of the market (Andreani et al. 2019). However, researchers are continuously working to develop new PV technologies that are more efficient, reliable, and cost-effective. One recent advancement is the PV-thermometric (PV-TE) hybrid device (Zhou et al. 2017), which utilizes the entire solar energy spectrum. Researchers have created a model of a concentrated PV-TE system and used advanced techniques to enhance the distribution of absorbed solar energy. This method improves the uniformity of energy absorption and increases the mean absorbed energy by 1.6 times compared to a flat surface.

Another promising technology is the use of Carbon Nanotube (CNT) Silicon solar cells (Yerkar, Bisane, and Waghchore 2017). These cells have several advantages over traditional PV technology, such as improved efficiency, the ability to work with infrared light and visible spectrum (making them operational at night), and stability at high temperatures. CNT-based solar panels also require less material for construction and have high electron mobility, resulting in increased output voltage (Kehang Cui et al. 2016).

Bifacial PV modules are also gaining interest among researchers due to their potential to generate 25% more power compared to traditional mono-facial modules (Molin et al. 2018). These modules perform better under shaded conditions and have shown to lose less power compared to mono-facial modules (Bhang et al. 2019; Zhang et al. 2020). However, there are reliability concerns regarding partial shading on the front side and shading on the rear side, which need to be addressed.

Passivated Emitter and Rear Cell (PERC) (Schulte-Huxel et al. 2017) technology has demonstrated higher efficiency than Aluminum Back Surface Field (Al-BSF) modules (Rodriguez-Gallegos et al. 2019). Researchers have optimized the metallization design for both mono-facial and bifacial PERC modules, considering real-world conditions (Vogt et al. 2017). Although PERC modules are prone to potential induced degradation (PID), modifications to the antireflection coating have been proposed to minimize this issue (Luo et al. 2018).

Researchers are also exploring the use of thin Si wafers to create single junction m-Si cells, aiming for cost-effectiveness, dependability, efficiency, and flexibility. Luminescent solar concentrator PV (LSC PV) modules have been developed as well, which operate at lower temperatures compared to traditional glass-sheet-based c-Si PV modules (Reinders, Debije, and Rosemann 2017).

Despite these technological advancements, PV systems still face challenges, such as the generation of hot spots (see Table 1) (Simon and Meyer 2010). Hot spots occur when a PV cell is shaded or less productive than other cells in the same string, leading to reduced system efficiency, cell degradation, and potential failures. The chances of reverse breakdown and generation of hot spots increase if the inactive cell area is greater than 8% (Kim and Krein 2015). Under such cases, the affected cell is forced into reverse bias condition to work in the second quadrant (see Figure 2) and dissipates power which causes local overheating. Mitigating the impacts of partial shading and hot spot generation is crucial to maintain the desired performance and lifespan of PV modules (Deng et al. 2017; Solheim et al. 2013).

The failure rate of PV panels within the first year of installation is approximately 25%, as shown in Figure 1. To address this, strict quality control and quality audits during manufacturing can help reduce defect-related hot spots and failures (Köntges et al. 2014). Techniques like ultrasonic testing, heat flux thermography, and electroluminescence should be employed to identify and reject defective cells or cell strings during module production. However, variations in quality control among manufacturers and the possibility of defects during transportation,

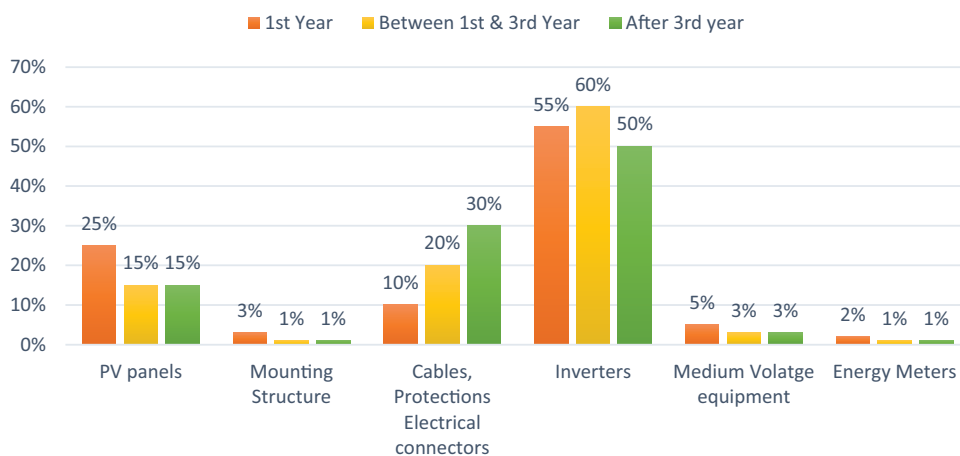
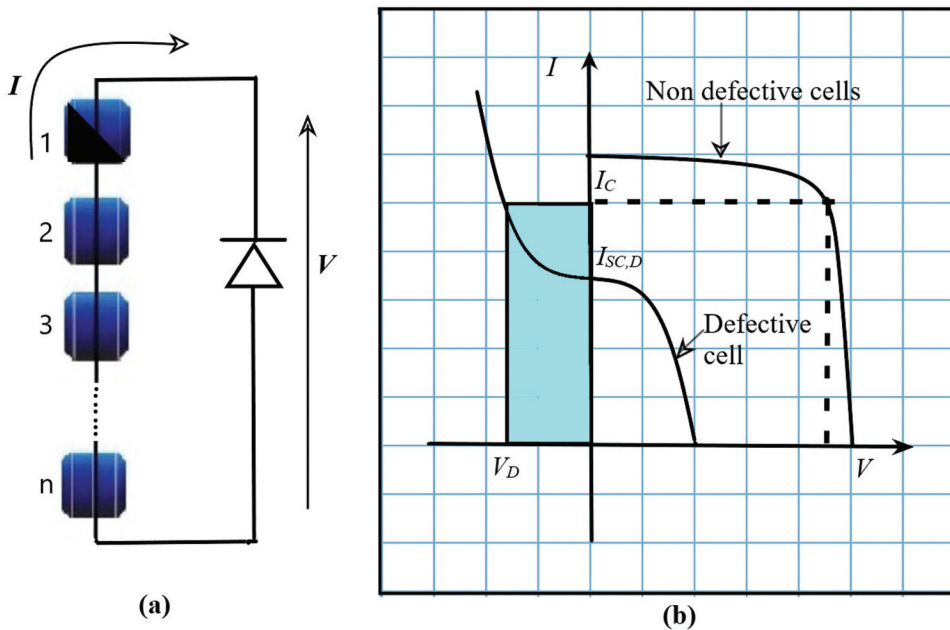


Figure 1. Failure incident rate concerning the components of PV plant (Fernandes et al. 2016; Kaplanis and Kaplani 2011).



**Figure 2.** (a) solar cells secured by a BPD with the first cell is concealed or damaged. (b) I-V characteristics of defective or shaded and non-defective cells (Boxwell 2017; Köntges et al. 2014).

handling, and installation can still lead to hot spot generation (Ghosh, Yadav, and Mukherjee 2019b). Factors like partial shading conditions from nearby buildings, trees, or aerosol deposition further contribute to the challenge (Moretón, Lorenzo, and Narvarte 2015). Researchers have proposed various solutions to mitigate this issue. The objective of this article is to summarize the existing research gaps and practices aimed at enhancing the performance of PV systems under PSC:

- (1) The various methods that are proposed to enhance the performance of PV systems under partial shading conditions, including distributed maximum power point tracking (DMPPT) (Pendem, Mikkili, and Katru 2018), bypass diodes, and advanced MPPT methods based on machine learning.
- (2) To explore the phenomenon of reverse breakdown and hot spot, its causes, and impact on the performance and reliability of PV systems.

Research in the field of partial shading in PV systems should focus on several key areas. First, addressing hot spots in PV systems under partial shade conditions is crucial, along with developing effective methods to mitigate their negative effects on system efficiency and reliability. Second, there is a need to explore innovative PV module designs that can minimize the impact of partial shade and hot spots, as well as develop reliable interconnection methods and optimize PV array configurations. Integration with energy storage systems is another important direction to enhance the stability and dependability of PV plants, enabling regular and predictable power output. Additionally, research on novel materials, such as semiconductors and thin films, should be conducted to improve the performance and dependability of PV systems.

To advance these areas, researchers should conduct detailed analyses of factors that affect PV system performance under partial shading, including shading patterns, module configurations, and operating conditions. Exploring emerging technologies like intelligent inverters, cascaded energy storage systems, and advanced control algorithms can mitigate the negative effects of partial shading and improve overall system efficiency and reliability. Furthermore, there should be an assessment of

unmet research needs and challenges related to partial shading, emphasizing the development of precise and reliable modeling and simulation tools to enable effective experimental studies for minimizing its impact.

By addressing these research gaps and focusing on these future directions, advancements in modeling, technology, and material applications can lead to more efficient and reliable PV systems, mitigating the effects of partial shading and maximizing solar energy generation.

### Theoretical background: reverse breakdown of PV cells

A solar module's internal circuit has a considerable impact on the development of hot spots. A solar module is made up of subpanels, each of which is made up of a sequence of PV cells connected by an antiparallel bypass diode (BPD) (Ghosh, Yadav, and Mukherjee 2019b; Winston 2019). The number of cells in a sub-panel  $N$  is 20 for 20  $V_{\text{NOM}}$  (60 cell module), 24 for 24  $V_{\text{NOM}}$  (72 cell module), and 32 for 32  $V_{\text{NOM}}$  (96 cell module).

The photo-induced current of one or more cells decreases under PSC and is forced into reverse-bias mode. Therefore, the reverse breakdown mechanism of PV cells plays a crucial role in PV performance under PSC. In reverse bias mode, the antiparallel diode protects these PV cells by mitigating the reverse voltage and providing an alternate path for the current. But standard BPD only limits the reverse bias voltage and does not eliminate it across the shaded cell (Guerriero et al. 2019). The reverse voltage across the shaded cell can be calculated using Kirchhoff's Voltage Law on the subpanel circuit (see Figure 3(b)):

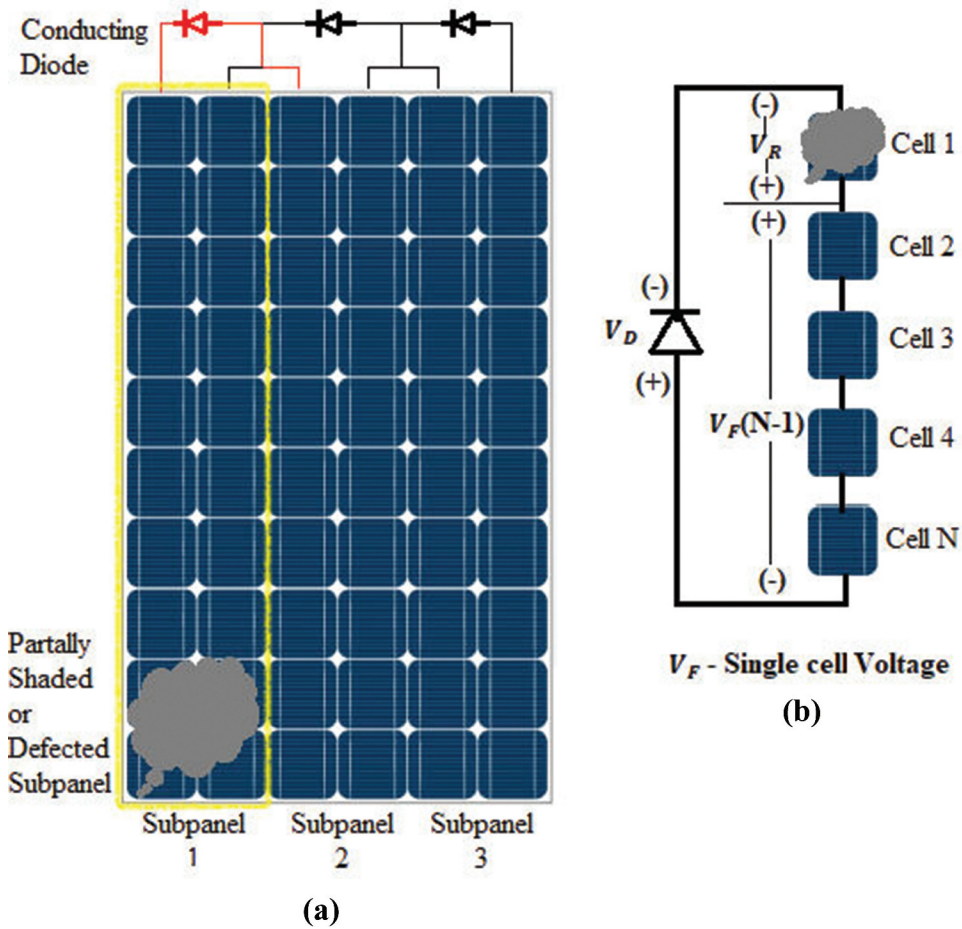
$$V_R = \sum_{n=0}^{N-1} V_F, n + V_D \quad (1)$$

where  $V_F$  is the forward (open circuit) voltage of the fully illuminated cell, and the forward voltage drop across BPD is denoted by  $V_D$ . For p-n junction semiconductor diodes,  $V_D$  generally varies from 0.6 V to 1.7 V. As a result; the number  $N$  should be kept as low as possible to avoid the reverse voltage surpassing the breakdown voltage. Figure 4(a) depicts the reverse bias characteristics of shaded/defective and non-defective cells. The reverse breakdown characteristics also vary with temperature (see Figure 4(b)). The temperature coefficient of reverse current is positive, resulting in a higher reverse current at higher temperatures (Breitenstein et al. 2011). Table 2 lists the different types of reverse breakdown mechanisms identified by Breitenstein et al (Breitenstein et al. 2011; Breitenstein et al. 2009). Table 3 lists the Precipitating factors of the reverse breakdown of the PV cells.

Because the current rises linearly during Stage I breakdown (see Figure 4 (a)), this occurrence is rarely lethal to PV cells. Stage II breakdown occurs at  $-7$  V to  $-11$  V in the vicinity of recombinative crystallographic defects such as iron precipitates in grain boundaries, where the defect-induced breakdown occurs. The reverse current develops exponentially in Stage II. Hot spots can emerge even if a considerable quantity of current does not flow in Stage I and II breakdowns under standard test conditions (STC). Due to avalanche breakdown, the reverse current exhibits a substantially exponential growth in Stage III and occurs in the voltage range of  $-13$  V to  $-18$  V (see Figure 4(a)). When one BPD is utilized over one-third of the cells in a module, the reverse voltage supplied across the shaded cell correlates to the voltage range where the most severe breakdown (i.e., Avalanche Breakdown) occurs.

Apart from hotspot, the PSC also alters the P-V and I-V characteristics (Lappalainen and Valkealahti 2017; Yadav et al. 2017; Yadav, Pachauri, and Chauhan 2016) of the PV modules.

Mismatch losses are caused by multiple peaks in the I-V and P-V characteristics (see Figure 5). According to previous studies, mismatch losses under PSC can affect annual energy yield by up to 25% (Jahn 2019). When PV modules are under 20% PSC, and the irradiance is between 1000 and 700  $\text{W/m}^2$ , maximum power declines by around 6.22% for every 100  $\text{W/m}^2$  drop in irradiance, according to another experimental investigation by Teo et al (Teo et al. 2018). Maximum power declines just 0.24% for every 100  $\text{W/m}^2$  drop in irradiance at lower irradiance levels (i.e., 700–0  $\text{W/m}^2$ ). As the irradiance falls below 700  $\text{W/m}^2$ , the PV system becomes immune to PSC. According to a related study (Islam et al. 2018; Javed et al. 2019;



**Figure 3.** Conventional bypass circuits of PV modules, (a) Provision of a BPD across each sub-panels, (b) voltage distribution in subpanel under PSC.

Sarwar et al. 2022), in reverse bias, less than twice the MPP Power is a “Safe” operating area, where the secondary breakdown is less likely.

Focus on actual junction breakdown mechanisms while classifying causes of failure (Breitenstein et al. 2011; Pachauri et al. 2021). It (see Figure 6.) will compile the most significant findings from a number of the authors’ earlier works, each of which focused on a particular aspect of the overall breakdown behavior and causes (Jia et al. 2021; Yadav and Mukherjee 2021).

## Hot spot mitigation techniques

### Introduction of smart circuits and elemental upgradation

As delineated in Section 2, solar PV modules use BPD across each sub-panel as the protective circuit against hot spots; however, BPD has limited protection capability. The maximum group size per diode that can be used without causing harm is around 15 cells/bypass diode for silicon cells. In a typical 36-cell module, two bypass diodes are used to ensure that the module does not suffer from “hot spot” damage (Honsberg and Bowden). As a result, it is ineffective in preventing the reverse disintegration of PV cells and the creation of hot spots. Researchers have proposed various smart circuits to mitigate hot spots using power semiconductor devices and elemental upgradation of the PV modules in the past (see Table 4).



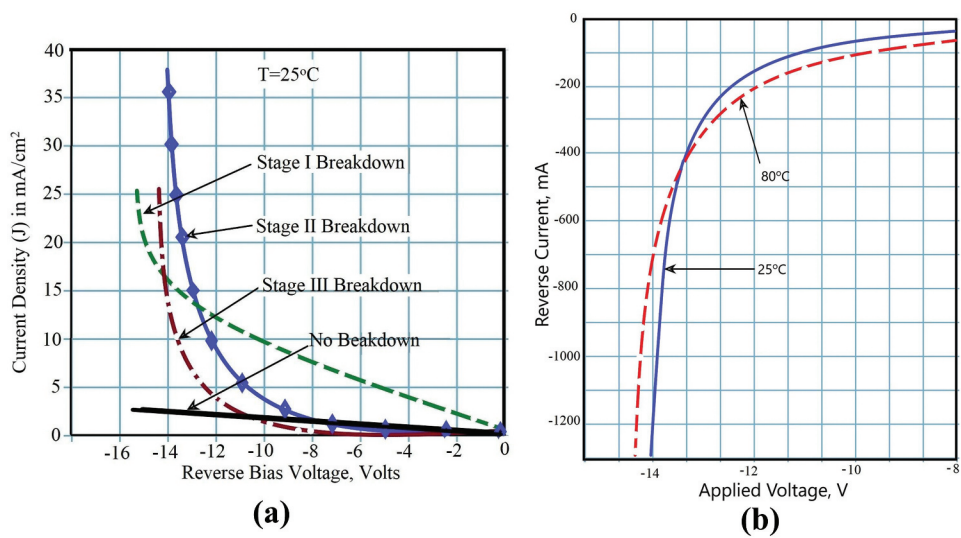


Figure 4. (a) different reverse breakdowns mechanisms and (b) effect of ambient temperature on reverse breakdown characteristics.

Table 1. Causes of the hot spot in PV Modules.

Type of factor	Factors causing the hot spot	Reference
External	Shade due to snow, leaves, droppings of birds, soot, and trees & buildings nearby.	(DuPontTM 2017; Fernandes et al. 2016; Gallardo-Saavedra and Karlsson 2018; Ghosh, Yadav, and Mukherjee 2019a; Gupta et al. 2019; Kim and Krein 2013; Rajput et al. 2018; Reddy, Reddy, and Kumar 2017; Suresh Kumar, Sarkar, and K S 2014; Zheng et al. 2013)
	Dust/dirt Deposition.	(Bouaichi et al. 2019; Ghosh, Yadav, and Mukherjee 2019a; Gupta et al. 2019, 2019; Kim and Krein 2013; Suresh Kumar, Sarkar, and K S 2014)
	Glass encapsulation discoloration	(Bouaichi et al. 2019; Boxwell 2017; Deng et al. 2017; Fernandes et al. 2016; Suresh Kumar, Sarkar, and K S 2014; Waqar Akram et al. 2019)
	Wear & Tear Due to Mechanical Loading/ rooftop conditions	(Bouaichi et al. 2019; Deng et al. 2017; DuPontTM 2017; Suresh Kumar, Sarkar, and K S 2014)
Internal	Micro Cracks	(Boxwell 2017; Deng et al. 2017; Moretón, Lorenzo, and Narvarte 2015; Suresh Kumar, Sarkar, and K S 2014; Waqar Akram et al. 2019)
	Defective Soldering	(Bouaichi et al. 2019; Boxwell 2017; Deng et al. 2017; Moretón, Lorenzo, and Narvarte 2015; Suresh Kumar, Sarkar, and K S 2014; Tsanakas et al. 2015; Waqar Akram et al. 2019)
	Potential Induced degradation	(Bright 2008; Deng et al. 2017; Moretón, Lorenzo, and Narvarte 2015; Suresh Kumar, Sarkar, and K S 2014; Waqar Akram et al. 2019)
	Material Imperfections	(Moretón, Lorenzo, and Narvarte 2015; Simon and Meyer 2010; Suresh Kumar, Sarkar, and K S 2014; Tsanakas et al. 2015; Waqar Akram et al. 2019)
	Cell Mismatch	(Deng et al. 2017; DuPontTM 2017; Ghosh, Yadav, and Mukherjee 2019a; Moretón, Lorenzo, and Narvarte 2015; Suresh Kumar, Sarkar, and K S 2014; Zheng et al. 2013)

Table 2. Type of breakdowns.

Stage	Type of Breakdown	Temperature coefficient	Slope	Voltage Level
I	Early pre-breakdown	Strongly Negative	Low	Below -5V
I	Edge Breakdown	Positive	Low	-5 V
II	Weak defect-induced	Zero or weakly negative	Moderate	-6 V to - 11 V
II	Strong defect-induced	Zero or weakly negative	Moderate	Below -12 V
III	Avalanche Breakdown	Negative	High	-13 V to - 18 V

**Table 3.** Precipitating factors of the reverse breakdown of the PV cells.

Precipitating Factors	Effects	Causes	Mitigating Schemes
High reverse-bias voltage (Breitenstein et al. 2011; Jaeun et al. 2021; Jordan and Kurtz 2013)	Reverse breakdown, increased leakage current, reduced efficiency	Inadequate design, incorrect installation, partial shading, high-temperature operation	Use of bypass diodes, proper design and installation, avoiding partial shading, reducing operating temperature
Light-induced degradation (GREEN et al. 2012; Jaeun et al. 2021; Jordan and Kurtz 2013; Köntges et al. 2014)	Reduced efficiency, increased leakage current	Exposure to light and high temperatures	Use of anti-reflective coatings, encapsulation with UV-resistant materials, reducing operating temperature
Hot carrier effects (Jordan and Kurtz 2013; Köntges et al. 2017)	Reduced efficiency, increased leakage current, decreased carrier lifetime	High electric field, high-temperature operation	Use of high-quality materials, proper design and installation, reducing operating temperature, optimization of doping concentration
Temperature effects (Jordan and Kurtz 2013; Köntges et al. 2017)	Reduced efficiency, increased leakage current	High operating temperature	Thermal management, use of cooling systems, proper design and installation, optimization of cell materials and structure
Cell/module defects (Köntges et al. 2017)	Reduced efficiency, increased leakage current	Manufacturing defects, material impurities, improper handling	Quality control, material purification, proper handling
Corrosion and moisture ingress (Köntges et al. 2017)	Reduced efficiency, increased leakage current, reduced lifetime	Exposure to moisture and corrosive substances, inadequate sealing	Improved module encapsulation, use of protective coatings
Cell/module damage	Reduced efficiency, increased leakage current, reduced lifetime	Physical damage, mechanical stress, impact from external objects	Improved module design, enhanced installation practices, use of protective layers
Electromagnetic interference (Zhou et al. 2017)	Reduced efficiency, increased noise, potential damage	Electromagnetic radiation from nearby sources, improper grounding, inadequate shielding	Proper grounding and shielding, use of surge protectors
Aging and degradation (Köntges et al. 2017)	Reduced efficiency, increased leakage current, reduced lifetime	Exposure to environmental factors, material degradation, wear and tear	Proper maintenance and cleaning, use of protective coating

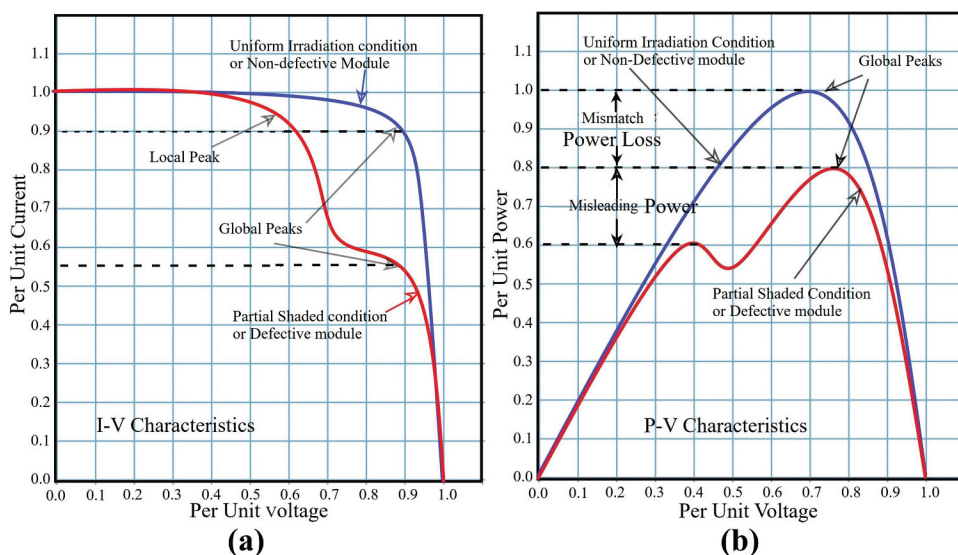
### **Elemental upgradation effects on PV under PSC**

Rajvikram et al (M et al. 2019), proposed a solution based on PV panels with phase-changing material (PCM). These materials have low thermal conductivity nature, so Thermal Conductivity Enhancers are utilized to lower the panel's working temperature and maximize power output. PV-PCM with an aluminum layer on the backside of the panel boosted the panel's conversion efficiency by an average of 24.4%, according to their findings. Their proposed method results in a maximum temperature drop of 13°C and an average temperature drop of 10.35°C, resulting in a 2% gain in panel electrical efficiency.

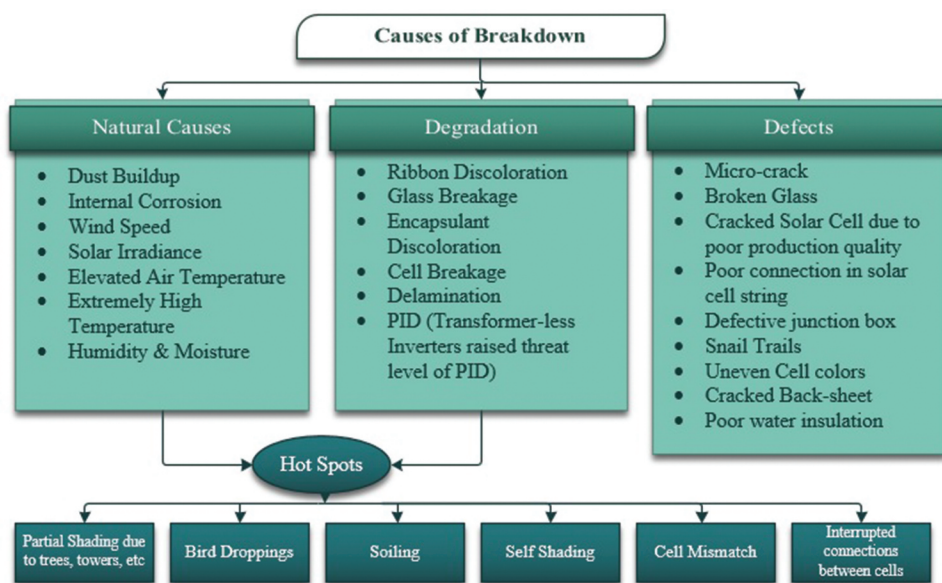
The technique proposed by D.Prince Winston (Winston 2019) has been proven to be capable of lowering the temperature of the impacted region and producing more power than typical BPD. During hot spot conditions, the suggested technique uses relay switches to disconnect the afflicted sub-string. In comparison to BPD, the proposed technique reduces the temperature of the hot-spotted cell from 69°C to 57°C and increases output power by 51%. Table 5 summarizes advantages, limitations, and cost consideration of elemental upgradation:

### **Hot spot mitigation through array configuration**

PV modules are typically connected in a power plant in a series-parallel (SP) configuration. The number of series-connected PV modules (string length) is determined by the plant's design voltage. And, depending on the plant capacity, numerous strings are connected in parallel (design output). The



**Figure 5.** Normalized I-V and P-V characteristics of partially shaded or defective cell (red curves) with Uniformly irradiated or non-defective module (blue curve) (Ghosh, Yadav, and Mukherjee 2019a; Moretón, Lorenzo, and Narvarte 2015).



**Figure 6.** Causes of breakdown.

impact of PSC can be minimized by modifying PV array topologies from the traditional SP layout, according to the literature. Table 6 summarizes related literature.

### MPPT techniques for minimization of mismatch losses

MPPT techniques for PV systems have been extensively investigated and developed by power engineers, and several algorithms have been created for MPPT to determine Global peaks under non-uniform solar irradiation. Table 7 summarizes a comparison of MPPT techniques based on complexity, tracking speed,

**Table 4.** Hot spot mitigation using smart circuits.

Smart Circuit	Important outcome of the experiment	Remarks	Ref.
Submodule Integrated Converter (SubMICs)-enhanced micro-converters (Distributed Power Electronics)	Sub-module level converter and differential power processing SubMICs – hot spot occurrences are reduced by 88.5% and 97.1%, respectively.	Distributed Power Electronics Reduces hot spots as well as increase energy yield in PV system.	(Olalla et al. 2018)
Power MOSFET is used as a voltage divider for shaded solar cells	For Heavily-stressing PSC in Modules with Low Shunt Resistance – 24°C temperature is reduced, and with High Shunt resistance – 20°C temperature is reduced compared to standard BPD.	EN 61,215 Qualification Procedure is the reference for PSCs.	(Daliento et al. 2016)
Conduction State Detection circuit (CSD) parallel with each BPD in addition with Controlled MOSFETs through each panel subgroup.	The MOSFETs in the shaded subgroup are turned off when the control system detects BPD conduction conditions. As a result, thermal stress and the creation of hot spots are no longer a problem.	Low cost and efficient design.	(Ayache, Chandra, and Chériti 2020)
Modified bypass circuit having IGBT and standard BPD	The bypass circuit reduces thermal stress and hot spots while lowering system complexity, power loss, and expense.	IGBTs can only be used at switching speeds below 200 kHz.	(Ghosh, Yadav, and Mukherjee 2019b)
Smart bypass circuits consist of NMOS, Gate Protection Circuit, Charge Pump, and Storage capacitors.	Smart bypass detects the cell's status and activates NMOS to bypass the failing cell or substring. It can be employed at the low voltage cell level or at the high voltage substring level to reduce the solar panel's power loss under PSC.	The Smart Bypass prototype has a few heating issues.	(Bauwens and Doutrelaigne 2014)
Relay circuits to open circuit the hot spotted series segment	The proposed technique reduces the temperature of the hot-spotted cell from 69 to 57 degrees Celsius and increases output power by 51% as compared to BPD. In medium hot spot circumstances, the proposed approach increases output power and voltage by 173% and 177%, respectively.	Only the proposed technique generates output power during extremely severe hot spot conditions, while the other techniques are unable to do so.	(Winston 2019)
MOSFET M1, feedback MOSFET M2 driven by digital oscillator TLC555	Bypass circuit completely prevents the rise in the temperature of shaded cells. It's also worth mentioning that during bypass events, the TLC555 consumes very little power because it's resting the rest of the time. When mismatch conditions arise, the bypass circuit self-activates, eliminating the need for microprocessors or other complex logic circuits. When mismatch conditions arise, the bypass circuit self-activates, eliminating the need for microprocessors or other complex logic circuits. When mismatch conditions arise, the bypass circuit self-activates, eliminating the need for microprocessors or other complex logic circuits.	In the worst-case operating conditions, as described by EN 61,215, a shaded PV cell protected by a normal BPD and a PV cell protected by the new bypass circuit had a temperature differential of roughly 50°C.	(Guerriero et al. 2019)

(Continued)

**Table 4.** (Continued).

Smart Circuit	Important outcome of the experiment	Remarks	Ref.
A low-power 8-bit Microcontroller Unit for wireless connectivity, data acquisition, measurement phases/durations. Power MOSFET as a switch.	The sensor is designed to measure the operating voltage, operational current, open-circuit voltage, and short circuit current of string-connected PV in real-time. The reliability of the system is enhanced with the sensor under PSC.	The proposed circuit is an effective method for high-granularity diagnostics and real-time PV plant performance evaluation.	(Guerriero et al. 2016)
To prevent the hot spotting, a voltage-threshold control is used in conjunction with an existing MPPT.	The voltage-threshold control works in tandem with the MPPT control to ensure that the string is functioning at the high-voltage local MPP without the BPD turned on. Imposing voltage-threshold regulation on a damaged cell reduces the intensity of hot spots.	With no additional hardware and limited computing requirements, a panel-level strategy for decreasing hot spots that combines MPPT and voltage-threshold control is effective.	(Kim et al. 2016)

precise monitoring, and cost. For PV system non-uniform irradiation, A robust global MPPT technique based on the wind-driven optimization (WDO) algorithm was proposed by Abdalla et al (Abdalla, Rezk, and Ahmed 2019). Under various shading conditions, they compared their new WDO algorithm to Particle Swarm Optimization (PSO), Differential Evolution (DE), Harmony Search Algorithm (HSA), Bat Algorithm, Sine-Cosine Algorithm (SCA), Cuckoo Search (CS), and Genetic Algorithm (GA). As far as statistical metrics go, the authors employed relative error, root mean square error, mean absolute error, standard deviation, success rate, tracking time, and efficiency. WDO has a higher success rate of 97.5% and the highest efficiency rate of 99.44%, according to their data, followed by DE and GA, and HSA and Bat algorithm have the lowest performance.

Using Enhanced Leader Particle Swarm Optimization(EL-PSO), Gavhane et al (Gavhane et al. 2017). provided a simulated analysis of MPPT under PSC. They employed a Siemens S75 panel under three distinct shading conditions to test the proposed approach. They concluded that EL-PSO is effective in detecting global optimization zones based on sequential mutations and that peaks with smaller power differences are also recognized by EL-PSO. The superiority of EL-PSO over PSO is demonstrated by its fast convergence, greater dynamic performance, ease of implementation, and high efficiency. Mahmoud Dhimish (Dhimish 2019) presented a comprehensive review of MPPT techniques to mitigate hot-spotting and the effect of PSC. The tracking precision of seven distinct state-of-the-art MPPT techniques was tested in a point-to-point analysis. Performance, control, circuit, and economic benefits must all be considered when selecting an MPPT approach. Seven techniques that were analyzed for comparison are Fast-Changing MPPT, Linear Extrapolation based MPPT, Modified Beta, I-V curve MPPT, Enhanced Adaptive Perturb and Observe Static Conductance-based MPPT, and Direct PWM voltage controller.

## Modern power converters under mismatch conditions

Modern power converters, including MLPEs, DC Optimizers, and micro-inverters, play a crucial role in improving power generation and system safety for Solar PV modules. Mismatches in power production can occur even in unshaded arrays, impacting the overall performance of a PV system (TIGO 2019). These mismatches, caused by variations in module performance, can result in energy losses of 2–5% initially, with increasing losses over time. However, the use of module-level power electronics offers a solution to recover these losses. Researchers are focusing on developing smart modules with embedded power optimizers, replacing the traditional junction box, to further enhance the efficiency and performance of PV systems (Fishelov and Adest 2022). Below is a list of some of the power converters provided in Table 8.

**Table 5.** Advantages, limitations, and cost consideration of elemental upgradation.

Elemental Upgradation	Advantages	Limitations	Cost Considerations
Phase Changing Materials (Awad et al. 2022; Browne, Norton, and McCormack 2015; M et al. 2019; Sivashankar, Selvam, and Manikandan 2021)	<p>Phase changing materials can undergo reversible phase transitions, allowing them to switch between high and low conductivity states.</p> <p>Enable self-healing and bypassing of shaded areas, minimizing power losses and maintaining higher overall system performance.</p> <p>Reduce the impact of partial shading on the overall PV system, enhancing energy yield and efficiency.</p>	<p>Requires precise control of temperature to activate the phase transition and ensure optimal conductivity changes.</p> <p>Compatibility issues with existing PV cell materials and structures may arise, requiring careful integration and testing.</p> <p>Thermal management challenges may arise due to the heat generated during the phase transition process.</p>	<p>Costs associated with acquiring and integrating phase changing materials into PV cell design.</p> <p>Additional fabrication and processing costs associated with incorporating phase changing materials.</p> <p>The cost of phase changing materials can vary depending on the specific material and required quantities.</p>
Contact Materials (Ahmad et al. 2018; Dwivedi et al. 2020; M et al. 2019)	<p>On average, the conversion efficiency was increased by 24.4%. The average overall electrical efficiency has increased by 2% due to a 10.35°C drop in average temperature.</p> <p>Enhances system performance, reliability, and efficiency under partial shading conditions.</p>	<p>Selection of appropriate contact materials like Aluminium sheet depends on cell technology, processing requirements, and compatibility factors.</p> <p>Compatibility issues may arise with different material combinations, requiring careful material selection and integration.</p>	<p>Costs associated with optimizing contact materials and any additional fabrication steps required.</p> <p>Contact material costs may vary depending on the specific materials used and fabrication requirements.</p>
Doping (Boxwell 2017; Breitenstein et al. 2011; Breitenstein et al. 2009; Venkateswari and Sreejith 2019)	<p>Enhances conductivity and carrier mobility in PV material, reducing power losses under partial shading.</p> <p>Improves charge transport and mitigates performance degradation in shaded areas.</p>	<p>Requires careful optimization of dopant concentration and distribution for desired results.</p> <p>Incorrect doping parameters can lead to undesired material properties and decreased performance.</p>	<p>Cost associated with doping agents and additional processing steps.</p> <p>Doping costs are typically integrated into the overall PV cell manufacturing process.</p>
Bandgap Engineering (Andreani et al. 2019; Chaves et al. 2020; Dharmadasa 2005; Honsberg and Bowden)	<p>Enables better absorption and utilization of different wavelengths of light, maximizing energy conversion.</p> <p>Facilitates increased power generation and improved performance under partial shading.</p>	<p>Challenging to achieve optimal bandgap engineering due to material compatibility and complex device structures.</p> <p>Fine-tuning bandgap may require advanced material characterization techniques and complex fabrication processes.</p>	<p>Costs may arise from specialized material deposition techniques and additional manufacturing steps.</p> <p>Bandgap engineering costs are usually incorporated into the overall fabrication process.</p>
Passivation Layers (Chen et al. 2018; Janssen et al. 2019; Luo et al. 2018)	<p>Reduces surface recombination, enhances carrier lifetime, and improves PV cell performance under partial shading.</p> <p>Minimizes power losses caused by surface defects and shading conditions.</p>	<p>Selection and deposition of appropriate passivation materials can be challenging for different cell technologies.</p> <p>Passivation layers may introduce additional process complexity and require careful optimization for desired performance.</p>	<p>Cost associated with specialized passivation materials and deposition techniques.</p> <p>Passivation layer costs are typically included in the overall PV cell manufacturing process.</p>

(Continued)



**Table 5.** (Continued).

Elemental Upgradation	Advantages	Limitations	Cost Considerations
Anti-Reflective Coating (Kaplan <a href="#">2016</a> ; Schulte-Huxel et al. <a href="#">2017</a> ; Vogt et al. <a href="#">2017</a> )	Reduces light reflection, increases light absorption, and compensates for power losses due to partial shading.	Coating design and optimization depend on the specific PV cell structure and material requirements.	Costs associated with anti-reflective coating materials, deposition methods, and additional manufacturing steps.
	Improves overall system performance, energy yield, and efficiency under varying shading conditions.	Anti-reflective coatings may introduce additional maintenance requirements and can be susceptible to wear and degradation.	Anti-reflective coating costs are typically integrated into the overall PV cell manufacturing process.
Tandem Cells (Bremner, Levy, and Honsberg <a href="#">2008</a> ; Cheng and Ding <a href="#">2021</a> ; Dharmadasa <a href="#">2005</a> ; Murayama and Mori <a href="#">2007</a> ; Yamaguchi et al. <a href="#">2021</a> )	Tandem cells stack multiple PV cells with varying bandgap energies, allowing for more efficient use of the solar spectrum.	Require complex manufacturing and integration processes, including the use of specialized materials and deposition techniques.	Costs associated with tandem cell design, manufacturing, and integration can be higher than traditional single-junction cells.
	Mitigate the impact of partial shading by utilizing cells with different bandgap energies that respond to different wavelengths.	May require advanced electrical and optical design considerations, such as optical coupling and series-parallel connection of cells.	Potential for higher energy yields and efficiencies may offset the higher initial investment in tandem cell technology.
	Offer higher energy conversion efficiencies compared to single-junction cells under both full and partial shading conditions.	Tandem cells may be sensitive to variations in illumination, temperature, and other environmental factors, requiring careful system design and control.	The cost of tandem cells can vary depending on the specific materials, design, and manufacturing requirements.

A boost DC-DC converter, which is applied directly to the PV module and incorporates an MPPT algorithm, a PLC system to communicate the information to a supervision control unit(SCU), a control unit with supervision and fault detection functions make up the prototype as in [Figure 7](#) (Orduz et al. [2011](#)).

The annual performance boost of 5.8% was observed by (Hanson et al. [2014](#)) after the module-level electronics were installed, equating to a recovery of about 30% of the shading losses in the system. Partial shade caused an average power loss of 8.3%, which would have climbed to 13% without optimization, according to a study of over 500 systems. NREL calculated that module-level optimization could recover 36% of the power wasted to partial shadow on average (TIGO [2019](#)). Compared to a current string inverter-based system with comprehensive MPP tracking, optimizers for PV installations with no shading features deliver less value in terms of energy production (Franke [2019](#)).

[Figure 8](#) summarizes the partial shading mitigation techniques using smart circuits, elemental upgradation, MPPT techniques, PV array configuration, and modern power converters, which offers effective solutions to combat the negative effects of shading on PV systems. By intelligently optimizing power distribution, improving component efficiency, and utilizing advanced algorithms, these techniques contribute to maximizing energy yield and overall system performance in partially shaded environments.

## Conclusions and future directions in the field

This article discusses in detail the detrimental effect of partial shading conditions on PV performance and reliability. Many researchers have recommended switches and bypass circuits to reduce hot spot temperature in articles published on this subject. However, recent studies indicate that the antiparallel bypass diode (BPD) is not highly efficient in addressing the reverse breakdown of PV cells unless a diode is connected across each cell, which is not cost-effective. Factoring the real issues involved in the partial shading, this article has made an in-depth analysis of the various path breaking research works published in the recent literature of high repute, with the research findings and solutions. The

**Table 6.** Impact of PV array configuration on partially shaded PV system.

Configuration	Key Findings/advantages	References
Series (S)	Lower Electrical losses due to smaller wire size and cable length higher mismatch losses and lesser current generation	(Ovidiu Popescu 2019) (Desai and Mikkili 2019; Wang and Hsu 2011)
Parallel (P)	The key benefit of this setup is its dependability. For different temperatures, 0°C, 45°C, 60°C, and 75°C parallel configuration promises the highest fill factor and power output during lightly shaded conditions.	(Ovidiu Popescu 2019)
Series-Parallel (SP)	Offers the highest value in terms of fill factor. Under PSC, SP-based Hierarchical reconfiguration can boost system efficiency by up to 50% while reducing switch usage and lowering fabrication costs. Due to short circuit currents under PSC, there is a loss of coherence between the MPPs of modules and the MPP of the array in the case of SP. low efficiency & high mismatch losses under PSC. SP setups are as good as more advanced electrical arrangements when it comes to output power changes induced by cloud shadings.	(Ngoc et al. 2019; Wang and Hsu 2011) (Shams El-Dein, Kazerani, and Salama 2013) (Desai and Mikkili 2019) (Lappalainen and Valkealahti 2017)
TCT	highest peak power value under PSC Under 9 different shading condition have highest average array efficiency i.e. 10.53% more than S, SP, BL and HC. But more number of wires is required makes it prone to faults. Under diagonal shading patterns, TCT yields 21.54% more power than SP configuration. TCT arrangement has lower power loss and a higher Fill factor under PSC, according to experiments.	(Desai and Mikkili 2019; Wang and Hsu 2011) (Darussalam, Pramana, and Rajani 2017) (Pachauri et al. 2019, 2020) (Jha and Triar 2018)
BL	Ranked third in terms of maximum power under PSC lower wiring cost compared to TCT. When shading area $\geq 50\%$ of total area the reconfiguration of PV array is not advisable.	(Wang and Hsu 2011) (Desai and Mikkili 2019) (Tubniyom et al. 2018)
HC	second in terms of maximum power under PSC wiring and wiring losses are more in comparison to S, SP.	(Wang and Hsu 2011) (Pendem and Mikkili 2018)
RSP	According to a prototype model, a reconfigurable PV module topology can produce up to 12.7% more energy than a shade tolerant PV module architecture with 6 bypass diodes when a PV module is shaded for 32% of the time. Interpolation is used in the proposed approach to limit the number of arithmetic operations performed in the embedded system, hence minimising the amount of time required to evaluate each configuration as much as possible. All possible configurations can be examined in an acceptable length of time using low-cost embedded electronics due to the reduction in calculation time.	(Calcabrini et al. 2021) (Serna-Garcés, Bastidas-Rodríguez, and Ramos-Paja 2016)
SP-TCT	Lags in Performance under PSC than Su-Do-Ku. Using Jig saw puzzle method SP-TCT yields 12.2% more power than conventional methods.	(Rani, Ilango, and Nagamani 2013) (Palpandian and David)
O-TCT	30% more power yield than SP & TCT under shaded conditions. Significantly reduce mismatch losses compared to SP & TCT. Array P-V characteristics are smoother with lower local maxima.	(Shams El-Dein, Kazerani, and Salama 2013) (Pachauri et al. 2020)
BL-TCT	Second in terms of array efficiency compared to S, SP and TCT.i.e average 10.29% more power than S and SP under 9 shading conditions. BL-TCT proved superior to S, SP and HC without escalating implementation cost.	(Desai and Mikkili 2019) (Kaushika and Gautam 2003)
HC-TCT	TCT and HC-TCT both obtained highest Combined Efficiency Scores (CES)	(Ghosh, Yadav, and Mukherjee 2018)
BL-HC	This design has the advantage of outperforming Total-Cross-Tied for asymmetrical array sizes and row wise shading schemes.	(Pendem and Mikkili 2018)
R-TCT	MS based Reconfigurations have low power loss and higher fill factor.	(Mishra et al. 2017)
RSP-TCT	Enhanced performance under shading conditions than conventional configurations	(Mishra et al. 2017) (Pachauri et al. 2020)
LS-TCT	Power loss is reduced, fill factor improved and higher maximum power is observed in LS-TCT than conventional TCT	(Pachauri et al. 2018, 2020)
M-TCT	GMPP is readily determined. Improved performances than SP and TCT under PSC. Shortcoming of M-TCT is wasted space due to shifting.	(Djilali et al. 2017)

(Continued)

**Table 6.** (Continued).

Configuration	Key Findings/advantages	References
Su-Do-Ku	Under PSC, the SU-DO-Ku puzzle configuration outperformed the TCT and SP-TCT puzzle configurations. Complicated for large array size and have more wiring loss than TCT.	(Rani, Ilango, and Nagamani 2013) (Chandrakant and Mikkili 2020)
LSP	LSP with TCT shows more power yielding capability than conventional TCT and other array configurations.	(Pachauri et al. 2018; Palpandian and David)
KKSP	When compared to SDKP configuration, KKSP and LSP give 18.86% lower wire losses; nevertheless, under PSC, KKSP outperforms SDKP and LSP.	(Yadav and Mukherjee 2018)
CDV	CDV extended with TCT show 23.97% increase in maximum power compared to SP, TCT and SDK. Average power generation increases by 21.67% under three cases of PSC.	(John Bosco and Carolin Mabel 2017)
Odd-Even	For Dwarf Broad Shading Pattern Odd-Even structure has 30.88% increased power output in comparison to TCT	(Nasiruddin et al. 2019)
Novel Structure (NS)	more efficient than TCT under PSC, with a maximum improved power of 13.2%.	(Mishra et al. 2017)
Non-Symmetrical Puzzle Pattern 1 & 2	having less power loss, a greater fill factor, and increased maximum power by 13.11%, 19.44%, and 10.7% for three different shading patterns compared to SP, BL, TCT, HC, BL-TCT, and SP-TCT.	(Yadav, Pachauri, and Chauhan 2016)

present analysis would add a new dimension to promote further research in future in the proposed field. The significant outcome of various literature survey has been summarized as,

- Active bypass circuits designed using solid-state switches such as IGBTs, MOSFETs, etc., have been shown to be more effective than traditional BPDs. However, these switches tend to attain higher temperatures during bypass action, so these innovative circuits require adequate thermal management arrangements for system reliability.
- Switching circuits, when applied to address hot spots, offer a better solution than BPDs. However, they can result in power drops due to the use of power semiconductor devices, reducing overall power during normal operation.
- Under partial shading conditions, the configuration of the PV array also plays an important role, and experimental evidence has shown that the TCT configuration is the most effective.
- Elemental upgrades and new switching techniques, such as distributed Power Electronics, can improve the performance ratio of PV systems under partial shading conditions.
- So far, the distributed MPPT PV array configuration is considered to be the best configuration for PV systems under partial shading conditions due to its ability to mitigate the effects of partial shading and maintain high efficiency.

In general, ongoing research and development into partial shade mitigation strategies for PV systems will continue to spur further innovation and boost efficient PV system developments, allowing more solar energy utilization even in difficult shading conditions.

Based on the review of various solutions for mitigating partial shading in PV systems, the following perspective and future guidelines are suggested:

- (1) **Holistic System Design:** Involves integrating advanced shading analysis tools, high-efficiency modules, and optimized system layouts to minimize the impact of shading and maximize energy yield.
- (2) **Advanced Module Technologies:** Includes exploring innovative bypass diode configurations, smart cell interconnections, and novel module materials to minimize power losses and improve overall system performance.

**Table 7.** Comparison of MPPT techniques under PSC.

MPPT Technique	Advantages	limitations	Cost of Implementation	Improved efficiency compared to conventional methods
Perturb & Observe Techniue (Abdalla, Rezk, and Ahmed 2019; Femia et al. 2005; Gaga, Errahimi, and Es-Sbai 2014; Jordehi 2016; Kjaer, Pedersen, and Blaabjerg 2005)	Simple and easy to imlement	May get stuck in a local MPP, low accuracy	Low-cost	3.3%
Fractional Short Circuit Current (FSCC) (Jordehi 2016; Sher et al. 2015, 2015)	High efficiency under varying shading conditions	Complex algorithm with non-linear equations	High-cost	4.6%
Particle Swarm Optimization (PSO) (Abdalla, Rezk, and Ahmed 2019; Gavhane et al. 2017; Gökmen et al. 2016; Jordehi 2016; Li et al. 2019; Liu et al. 2012; Pillai et al. 2018; Singh et al. 2021)	High accuracy, robust to environmental changes	May converge to local minima	High	5.1%
Artificial Neural Networks (ANNs) (Bouselham et al. 2017; Li et al. 2019)	High accuracy, adaptable to changing conditions	Require a large amount of training data, computationally expensive	High-cost	4.8%
Fuzzy Logic (Ibrahim, Nasr, and Enany 2021; Pervez et al. 2021)	Robust to uncertainty and imprecise data	Poor accuracy in highly dynamic conditions	Moderate	3.9%
Heterojunction with Intrinsic Thin layer (HIT) MPPT (Bansal, Jaiswal, and Singh 2021; Islam et al. 2018; Venkateswari and Sreejith 2019)	High efficiency even in low irradiance and shading	Limited availability of HIT solar panels	High-cost	1.3%
Current Sweep (Bhukya, Kedika, and Salkuti 2022; Ishaque and Salam 2013; Tsang and Chan 2015)	High efficiency under varying shading conditions	Low accuracy under partial shading conditions	Low-cost	4.2%
Hybrid MPPT (Kumar et al. 2023; Sarwar et al. 2022; Sher et al. 2015)	High efficiency and robustness	Complexity of algorithm with increased computational requirements	High-cost	4.9%
Improved Gradient Descent MPPT (Kamal et al. 2019; Kofinas et al. 2015)	High efficiency under varying shading conditions	High sensitivity to initial conditions and noisy measurements	Low-cost	5.5%
Modified Conductance Incremental MPPT (Islam et al. 2018; Tey and Mekhilef 2014)	High efficiency under varying shading conditions	Limited improvement in efficiency compared to other MPPT methods	Low-cost	3.5%
Grey Wolf Optimizer MPPT (Javed et al. 2019; Mohanty, Subudhi, and Ray 2016)	High accuracy and efficiency	Limited availability of grey wolf optimizer software	Low-cost	3.3%
Wind-Driven Optimization MPPT (Abdalla, Rezk, and Ahmed 2019)	High accuracy and efficiency	Limited testing under practical conditions	Low-cost	3.9%
Bacterial Foraging Optimization MPPT (Kumar, Puttamadappa, and Chandrashekar 2020; Zhu 2018)	High accuracy and efficiency	Limited testing under practical conditions	Low-cost	4.2%
Sine Cosine Algorithm MPPT (Abdalla, Rezk, and Ahmed 2019; Karmouni et al. 2022)	High accuracy and efficiency	Limited availability of sine cosine algorithm software	Low-cost	3.6%
Harmony Search Algorithm MPPT (Abdalla, Rezk, and Ahmed 2019)	High accuracy and efficiency	Limited availability of harmony search algorithm software	Low-cost	4.9%
Ant Colony Optimization MPPT (Sarwar et al. 2022)	High accuracy and efficiency	Limited testing under practical conditions	Low-cost	4.4%

(Continued)

**Table 7.** (Continued).

MPPT Technique	Advantages	limitations	Cost of Implementation	Improved efficiency compared to conventional methods
Adaptive Fuzzy PSO MPPT (Ibrahim, Nasr, and Enany 2021; Ishaque and Salam 2013)	High accuracy and efficiency	High computational requirements	High-cost	4.2%
Shuffled Frog Leaping Algorithm MPPT (Maarof and Ismail 2022; Sridhar et al. 2017)	High accuracy and efficiency	Limited availability of shuffled frog leaping algorithm software	Low-cost	4.1%
Cat Swarm Optimization MPPT (Guo et al. 2018)	High accuracy and efficiency	Limited testing under practical conditions	Low-cost	5.5%
Water Cycle Algorithm MPPT (Sarvi, Soltani, and Avanaki 2014)	High accuracy and efficiency	High computational requirements	High-cost	5.2%
Improved Shuffled Frog Leaping Algorithm MPPT (Aldosary et al. 2021; Nie and Nie 2017)	High accuracy and efficiency	Limited availability of improved shuffled frog leaping algorithm software	Low-cost	4.7%
Genetic Algorithm MPPT (Abdalla, Rezk, and Ahmed 2019; Hadji, Gaubert, and Krim 2018; Kumar et al. 2023; Li et al. 2018)	High accuracy and efficiency	High computational requirements	High-cost	4.6%
Fruit Fly Optimization MPPT (Megantoro et al. 2022; Sarwar et al. 2022)	High accuracy and efficiency	Limited availability of fruit fly optimization algorithm software	Low-cost	4.5%
Flower Pollination Algorithm (Kumar et al. 2023)	Simple implementation, high accuracy, robustness to local minima	Large number of iterations required, sensitive to initial conditions	Low cost	2.87% improvement in Pmax
Bat Algorithm (Abdalla, Rezk, and Ahmed 2019)	Robustness to noise, fast convergence	Requires fine-tuning of parameters, may get stuck in local minima	Low cost	2.2% improvement in Pmax
Improved Particle Swarm Optimization (Abdulkadir, Yatim, and Yusuf 2014; Pervez et al. 2021)	Fast convergence, robustness to local minima, low sensitivity to initial conditions	May converge to suboptimal solutions, requires fine-tuning of parameters	Low cost	2.7% improvement in Pmax
Bat-inspired Algorithm with Wavelet Transform (Alyasseri et al. 2022)	High accuracy, robustness to noise and local minima	Requires fine-tuning of parameters, computationally intensive	High cost	2.5% improvement in Pmax
Grey Relational Analysis-based Technique (Javed et al. 2019)	High accuracy, robustness to noise and local minima	Requires fine-tuning of parameters, computationally intensive	High cost	2.4% improvement in Pmax
Electromagnetism-Like Mechanism Algorithm (EM) (Tan et al. 2018)	Good convergence speed, robustness, efficiency	Sensitive to parameter tuning, limited exploration ability	Medium	Higher Pmax for EM-optimized system compared to GA-optimized system

- (3) Intelligent Monitoring and Control: This would involve employing advanced algorithms and machine learning techniques to dynamically optimize system operation, reconfigure module connections, and mitigate shading-induced performance degradation.
- (4) Integration with Energy Storage: As PV systems become more prevalent, integrating energy storage technologies will play a crucial role in managing intermittent power output caused by

**Table 8.** Modern power converters.

Device	Key Features	Functions
Module Optimizer (Brown 2021; Franke 2019) ()	Designed to withstand harsh environmental conditions. Advanced, in-the-moment performance evaluation. For installer and firefighter safety, the module DC voltage is automatically reduced to a safe level when the inverter or grid is turned off.	Optimizer replaces the typical junction box on a PV panel, resulting in a smart module that produces more power. Smart modules include module-level power electronics, which improve power harvesting, safety, and module-level monitoring.
DC Optimizers (Brown 2021) (Casey 2019; Hanson et al. 2014; Moorthy et al. 2020)	DC to DC energy is “conditioned” by optimizers and sent to the central inverter. Independent optimization technology enables operation with any inverter and does not necessitate the purchase of additional interface gear.	DC optimizer enhances the current at its output to match the current flowing through the unshaded modules when a shaded module produces power with a lower current; to compensate, the optimizer drops its output voltage by the same amount it boosts the current.
Module Level Power Electronics (Brown 2021; Hanson et al. 2014; Saur News Bureau)	benefits such as mismatch mitigation and module-level monitoring, are also available. MLPE have an energy self-consumption that results in higher power losses in both the additional connectors and the internal power electronics, which is more substantial.	Individual modules have MLPEs attached to them to improve performance in the shaded conditions. MPP tracking for each module is key function of MLPEs.
Micro Inverters (Brown 2021; Casey 2019; Saur News Bureau)	At the module’s location, micro-inverters convert DC to AC electricity. Micro-inverters require maintenance, but they are less likely to degrade your system’s performance.	Micro-inverters boost each module’s output performance at the individual level, resulting in a slew of advantages for system owners.
1- Phase Inverters with HD wave (Casey 2019; SolarEdge Solution)	Weighted efficiency of 99%. Small, light, and simple to set up. Low heat dissipation ensures high reliability. Integrated module-level monitoring. IP65 – suited for both outdoor and indoor use.	Single-phase inverters are built employing a cutting-edge power conversion technique based on distributed switching and high-performance DSP processing. The inverter may provide a pure sine wave, resulting in a significant reduction in the magnetics and heavy cooling elements.
MPPT Converters (Orduz et al. 2011)	The performance is based on three factors: 1) Maximum decoupling current conversion ratio. 2) BPD effect on PV module. 3) grid-connected PV inverter’s input voltage.	The MPPT converter will keep the PV generator’s working point as close as MPP under all operating conditions, including varying irradiance, temperature, and load characteristics.

partial shading. Future guidelines should explore the optimal sizing, control strategies, and economic viability of PV systems combined with energy storage to enhance system performance and grid integration.

By incorporating these perspectives and future guidelines, a roadmap is provided for researchers, industry professionals, and policymakers to advance the field of partial shading mitigation in PV systems.

## Nomenclature

### Abbreviations

PV	Photovoltaic
PV-TE	PV-thermometric
MC-FDTD	Monte Carlo-Finite Difference Time Domain
CNT	Carbon Nanotube
MLPEs	Module-Level Power Electronics
PERC	Passivated Emitter and Rear Cell
Al-BSF	Aluminum Back Surface Field
PID	Potential Induced Degradation
n-PERT	N-Type Passivated Emitter Rear Totally diffused
LSC PV	luminescent solar concentrator PV



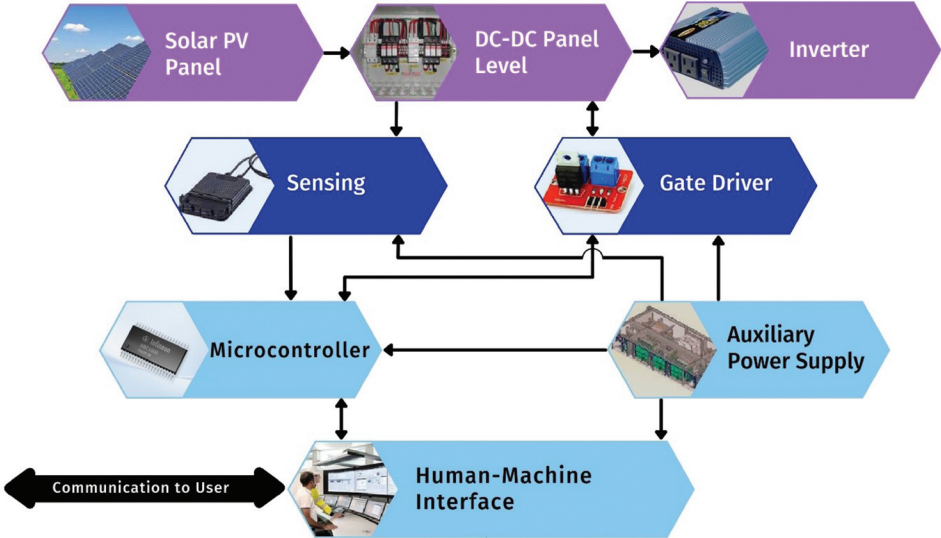


Figure 7. DC-DC power optimizer.

Smart Circuits	Elemental Upgradation	PV Array Configuration	MPP T Techniques	Modern Power Converters
<ul style="list-style-type: none"><li>• SubMICs enhanced DPE</li><li>• Power MOSFET</li><li>• Conduction state Detection Circuit</li><li>• Modified bypass circuit having IGBT &amp; BPD</li><li>• Smart bypass circuit consist of NDMOS</li><li>• Relay circuits</li><li>• MOSFET driven by TLC555</li><li>• 8 bit Microcontroller</li><li>• Voltage Threshold Control</li></ul>	<ul style="list-style-type: none"><li>• Phase Changing Materials</li><li>• Contact Materials</li><li>• Doping</li><li>• Bandgap Engineering</li><li>• Passivation Layers</li><li>• Anti-Reflective Coating</li><li>• Tandem Cells</li></ul>	<ul style="list-style-type: none"><li>• Series</li><li>• Parallel</li><li>• SP</li><li>• TCT</li><li>• HC</li><li>• BL</li><li>• SP-TCT</li><li>• BL-HC</li><li>• Su-Do-Ku</li><li>• LSP</li><li>• KKSP</li><li>• SDV</li><li>• Odd-Even</li><li>• NS</li><li>• Non-Symmetrical Puzzle Pattern 1&amp;2</li></ul>	<ul style="list-style-type: none"><li>• Perturb &amp; Observe</li><li>• IC</li><li>• FSOC</li><li>• PSO</li><li>• ANN</li><li>• Fuzzy Logic</li><li>• HIT</li><li>• Current Sweep</li><li>• Hybrid MPPT</li><li>• IGD</li><li>• MCI</li><li>• GWO</li><li>• WDO</li><li>• EM</li><li>• GOA</li></ul>	<ul style="list-style-type: none"><li>• Module Optimizer</li><li>• DC Optimizer</li><li>• Module Level Power Electronics</li><li>• Micro Inverters</li><li>• 1-Phase Inverter with HD wave</li><li>• MPPT Converters</li></ul>

Figure 8. Classification of partial shading mitigation techniques.

c-Si	Crystalline silicon
PSC	partial shading conditions
MPP	Maximum Power Point
MPPT	MPP Tracking
DMPPT	Distributed MPPT
BPD	Bypass Diode
STC	standard test conditions
SubMICs	Submodule Integrated Converters
MOSFET	Metal-Oxide-Semiconductor Field-Effect Transistor
CSD	Conduction State Detection
IGBT	Insulated-Gate Bipolar Transistor

NMOS	N-channel Metal-oxide Semiconductor
PCM	phase-changing material
TCT	Total-Cross-Tied
BL	Bridge-Link
HC	Honey Comb
SP	Series-Parallel
PLC	Programmable Logic Controller
SCU	Supervision Control Unit
SDKP	SuDoKu puzzled
IC	Incremental Conductance
O-TCT	Optimal TCT
RSP	Reconfigurable SP
LS-TCT	Latin-based puzzle-based TCT
M-TCT	Modified TCT
NS	Novel Structure
CDV	Cross Diagonal View
KKSP	Ken-Ken Square puzzled
WDO	Wind-Driven Optimization
DE	Differential Evolution
CS	Cuckoo Search
SCA	Sine-Cosine Algorithm
GA	Genetic Algorithm
HSA	Harmony Search Algorithm
PSO	Particle Swarm Optimization
EL-PSO	Enhanced Leader-PSO
ANN	Artificial Neural Network
PWM	Pulse Width Modulation
FSCC	Fractional Short Circuit Current
EM	Electromagnetism-Like Mechanism Algorithm
HIT	Heterojunction with Intrinsic Thin layer
GWO	Grey Wolf Optimizer
BFO	Bacterial Foraging Optimization
IGD	Improved Gradient Descent
GOA	Grasshopper Optimization Algorithm
P&O	Perturb & Observe
<b>Symbol</b>	
$V_R$	Reverse Voltage
$V_F$	Forward (Open Circuit) Voltage
$V_D$	Forward Voltage Drop
$I-V$	Current Voltage

## Disclosure statement

We hereby declare that there is no conflict of interest with regards to this article.

## Notes on contributors



**Nikhil Kushwaha** received the B.Tech Degree in Electrical & Electronics Engineering from UPTU, Uttar Pradesh, India in 2010, The M.Tech Degree In Power System Engineering from National Institute of Technology, Hamirpur, India, in 2012. He is currently working toward the Ph.D. degree with Delhi Technological University, Delhi, India. His research interests include Solar Array PV Reconfiguration, Hot-spot mitigation, diagnostic and monitoring techniques for photovoltaic devices and systems.



**Vinod Kumar Yadav** (Senior Member, IEEE) received the B. Tech. degree in electrical engineering from the Institute of Engineering and Technology, Bareilly, India, in 2003, the M. Tech. degree in power system engineering from the National Institute of Technology, Jamshedpur, India, in 2005, and the Ph.D. degree in power system engineering from the Indian Institute of Technology, Roorkee, India, in 2011. His research interests include renewable energy systems, power system planning and optimization, distributed generations, and smart grid.



**Radheshyam Saha** worked as the Chief Engineer at the Central Electricity Authority and is currently serving as a Professor in the Electrical Engineering Department at Delhi Technical University (DTU) in Delhi, India. He received his Ph.D. degree in FACTS Technology from the Indian Institute of Technology, Delhi, India, in 2008. His research interests include HVDC and Power Systems.

## ORCID

Nikhil Kushwaha  <http://orcid.org/0000-0003-3458-6284>

## References

- Abdalla, O., H. Rezk, and E. M. Ahmed. 2019. Wind driven optimization algorithm based global MPPT for PV system under non-uniform solar irradiance. *Solar Energy* 180 (August 2018):429–44. doi:10.1016/j.solener.2019.01.056.
- Abdulkadir, M., A. H. M. Yatim, and S. T. Yusuf. 2014. An improved PSO-based MPPT control strategy for photovoltaic systems. *International Journal of Photoenergy* 2014 (1c):1–11. doi:10.1155/2014/818232.
- Ahmad, N., A. Khandakar, A. El-Tayeb, K. Benhmed, A. Iqbal, and F. Touati. 2018. Novel design for thermal management of PV cells in harsh environmental conditions. *Energies* 11 (11). doi: 10.3390/en11113231.
- Aldosary, A., Z. M. Ali, M. M. Alhaider, M. Ghahremani, S. Dadfar, and K. Suzuki. 2021. A modified shuffled frog algorithm to improve MPPT controller in PV system with storage batteries under variable atmospheric conditions. *Control Engineering Practice* 112 (April):104831. doi:10.1016/j.conengprac.2021.104831.
- Alyasseri, Z. A. A., O. A. Alomari, M. A. Al-Betar, S. N. Makhadmeh, I. A. Doush, M. A. Awadallah, A. K. Abasi, and A. Elnagar. 2022. Recent advances of bat-inspired algorithm, its versions and applications. *Neural Computing & Applications*. 34 (19):16387–422. Springer London. doi:10.1007/s00521-022-07662-y.
- Andreani, L. C., A. Bozzola, P. Kowalczewski, M. Liscidini, and L. Redorici. 2019. Silicon solar cells: Toward the efficiency limits. *Advances in Physics* 4 (1):1548305. doi:10.1080/23746149.2018.1548305.
- Awad, M. M., O. K. Ahmed, O. M. Ali, N. T. Alwan, S. J. Yaqoob, A. Nayyar, M. Abouhawwash, and A. F. Alrasheedi. 2022. Photovoltaic thermal collectors integrated with phase change materials: A comprehensive analysis. *Electron* 11 (3):337. doi:10.3390/electronics11030337.
- Ayache, K., A. Chandra, and A. Chérity. 2020. An embedded reconfiguration for reliability enhancement of photovoltaic shaded panels against hot spots. *IEEE Transactions on Industry Applications* 56 (2):1815–26. doi:10.1109/TIA.2019.2956912.
- Bansal, N., S. P. Jaiswal, and G. Singh. 2021. Comparative investigation of performance evaluation, degradation causes, impact and corrective measures for ground mount and rooftop solar PV plants – a review. *Sustain Energy Technol Assessments* 47 (August):101526. doi:10.1016/j.seta.2021.101526.
- Bauwens, P., and J. Doutreloigne. 2014. Reducing partial shading power loss with an integrated smart bypass. *Solar Energy* 103:134–42. doi:10.1016/j.solener.2014.01.040.
- Bhang, B. G., W. Lee, G. G. Kim, J. H. Choi, S. Y. Park, and H. K. Ahn. 2019. Power performance of bifacial c-Si PV modules with different shading ratios. *IEEE Journal of Photovoltaics* 9 (5):1413–20. doi:10.1109/JPHOTOV.2019.2928461.
- Bhukya, L., N. R. Kedika, and S. R. Salkuti. 2022. Enhanced maximum power point techniques for solar photovoltaic system under uniform insolation and partial shading conditions: A review. *Algorithms* 15 (10). doi:10.3390/a15100365.

- Bouaichi, A., and Ahmed Alami, M. 2019. In-situ evaluation of the early PV module degradation of various technologies under harsh climatic conditions: The case of Morocco. *Renewable Energy* 143:1500–18. doi:10.1016/j.renene.2019.05.091.
- Bouselham, L., M. Hajji, B. Hajji, and H. Bouali. 2017. A new MPPT-based ANN for photovoltaic system under partial shading conditions. *Energy Procedia* 111 (September 2016):924–33. doi:10.1016/j.egypro.2017.03.255.
- Boxwell, M. 2017. *Solar electricity handbook - a simple, practical guide to solar energy-designing and installing solar PV systems*. 2017th ed. Greenstream Publishing. <http://www.solarelectricityhandbook.com/>.
- Breitenstein, O., J. Bauer, K. Bothe, W. Kwapił, D. Lausch, U. Rau, J. Schmidt, M. Schneemann, M. C. Schubert, J.-M. Wagner, et al. 2011. Understanding junction breakdown in multicrystalline solar cells. *Journal of Applied Physics* 109 (7). doi: 10.1063/1.3562200.
- Breltenstein, O. et al. 2009. Physical mechanisms of breakdown in multicrystalline silicon solar cells. *Conf. Rec. IEEE Photovolt. Spec. Conf.* 000181–86. doi: 10.1109/PVSC.2009.5411700.
- Bremner, S. P., M. Y. Levy, and C. B. Honsberg. May, 2008. Analysis of tandem solar cell efficiencies under AM1.5G spectrum using a rapid flux calculation method. *Progress in Photovoltaics: Research and Applications* 16(3):225–33. doi: 10.1002/ppp.799.
- Bright, R. 2008. Selecting cable strain reliefs. *Electronic Production (Garden City, New York)* 50 (3):33–51.
- Brown, C. 2021. Shading losses in PV systems, and techniques to mitigate them. <https://www.aurorasolar.com/blog/shading-losses-for-pv-systems-and-techniques-to-mitigate-them/>.
- Browne, M. C., B. Norton, and S. J. McCormack. 2015. Phase change materials for photovoltaic thermal management. *Renewable and Sustainable Energy Reviews* 47:762–82. doi:10.1016/j.rser.2015.03.050.
- Calcabrini, A., M. Muttillio, R. Weegink, P. Manganiello, M. Zeman, and O. Isabella. 2021. A fully reconfigurable series-parallel photovoltaic module for higher energy yields in urban environments. *Renewable Energy* 179:1–11. doi:10.1016/j.renene.2021.07.010.
- Casey, B. 2019. Power optimizers: Everything you need to know. <https://www.solaris-shop.com/blog/power-optimizers-everything-you-need-to-know/>.
- Chandrakant, C. V., and S. Mikkili. 2020. A typical review on static reconfiguration strategies in photovoltaic array under non-uniform shading conditions. *CSEE Journal of Power and Energy Systems*. doi:10.17775/cseejpes.2020.02520.
- Chaves, A., J. G. Azadani, H. Alsalman, D. R. da Costa, R. Frisenda, A. J. Chaves, S. H. Song, Y. D. Kim, D. He, J. Zhou, et al. 2020. Bandgap engineering of two-dimensional semiconductor materials. *Npj 2D Materials and Applications* 4 (1). doi: 10.1038/s41699-020-00162-4.
- Chen, Y., P. P. Altermatt, D. Chen, X. Zhang, G. Xu, Y. Yang, Y. Wang, Z. Feng, H. Shen, P. J. Verlinden, et al. 2018. From laboratory to production: Learning models of efficiency and manufacturing cost of industrial crystalline silicon and thin-film photovoltaic technologies. *IEEE Journal of Photovoltaics*. 8(6):1531–38. doi:10.1109/JPHOTOV.2018.2871858.
- Cheng, Y., and L. Ding. 2021. Perovskite/Si tandem solar cells: Fundamentals, advances, challenges, and novel applications. *SusMat* 1 (3):324–44. doi:10.1002/sus2.25.
- Daliento, S., F. Di Napoli, P. Guerriero, and V. d'Alessandro. 2016. A modified bypass circuit for improved hot spot reliability of solar panels subject to partial shading. *Solar Energy* 134:211–18. doi:10.1016/j.solener.2016.05.001.
- Darussalam, R., R. I. Pramana, and A. Rajani. 2017. Experimental investigation of serial parallel and total-cross-tied configuration photovoltaic under partial shading conditions. *Proceeding - ICSEEA 2017 Int. Conf. Sustain. Energy Eng. Appl. Continuous Improv. Sustain. Energy Eco-Mobility* 2018-Janua:140–44. doi:10.1109/ICSEEA.2017.8267699.
- Deng, S., Zhang, Z., and Ju, C. 2017. Research on hot spot risk for high-efficiency solar module. *Energy Procedia* 130:77–86. doi:10.1016/j.egypro.2017.09.399.
- Desai, A. A., and S. Mikkili. 2019. Modelling and analysis of PV configurations (alternate TCT-BL, total cross tied, series, series parallel, bridge linked and honey comb) to extract maximum power under partial shading conditions. *CSEE Journal of Power and Energy Systems* (99). doi:10.17775/CSEEJPES.2020.00900.
- Dharmadasa, I. M. 2005. Third generation multi-layer tandem solar cells for achieving high conversion efficiencies. *Solar Energy Materials and Solar Cells* 85 (2):293–300. doi:10.1016/j.solmat.2004.08.008.
- Dhimish, M. 2019. Assessing MPPT techniques on hot-spotted and partially shaded photovoltaic modules: Comprehensive review based on experimental data. *IEEE Transactions on Electron Devices* 66 (3):1132–44. doi:10.1109/TED.2019.2894009.
- Djilali, N., N. Belhaouas, P. Agathoklis, B. Amrouche, B. Amrouche, K. Sedraoui, and N. Djilali. 2017. PV array power output maximization under partial shading using new shifted PV array arrangements. *Applied Energy* 187:326–37. doi:10.1016/j.apenergy.2016.11.038.
- DuPont™. 2017. *Mitigating strategies for hot spots in crystalline silicon solar panels*. <https://pdf4pro.com/view/mitigating-strategies-for-hot-spots-in-crystalline-silicon-59726e.html>.
- Dwivedi, P., K. Sudhakar, A. Soni, E. Solomin, and I. Kirpichnikova. 2020. Advanced cooling techniques of P. V. modules: A state of art. *Case Studies in Thermal Engineering* 21 (June):100674. doi:10.1016/j.csite.2020.100674.
- Femia, N., G. Petrone, G. Spagnuolo, and M. Vitelli. 2005. Optimization of perturb and observe maximum power point tracking method. *IEEE Transactions on Power Electronics* 20 (4):963–73. doi:10.1109/TPEL.2005.850975.
- Fernandes, C. A. F., J. P. N. Torres, M. Morgado, and J. A. P. Morgado. 2016. Aging of solar PV plants and mitigation of their consequences. *Proc. - 2016 IEEE Int. Power Electron. Motion Control Conf. PEMC 2016*. 1240–47 doi: 10.1109/EPEPEMC.2016.7752174.

- Fishelov, Amir, and Adest, Meir. 2022. "Embedded power optimizers for smart modules." <https://www.solaredge.com/products/power-optimizer/smart-modules#/>.
- Franke, W. T. 2019. The impact of optimizers for PV modules. *SDU Electr. Eng.* May. [Online]. Available: [https://www.sdu.dk/-/media/files/om\\_sdu/centre/cie/optimizer+for+pv+modules+ver11\\_final.pdf](https://www.sdu.dk/-/media/files/om_sdu/centre/cie/optimizer+for+pv+modules+ver11_final.pdf).
- Gaga, A., F. Errahimi, and N. Es-Sbai. 2014. Design and implementation of MPPT solar system based on the enhanced P&O algorithm using Labview. 2014 *International Renewable and Sustainable Energy Conference (IRSEC)*. no. 1. 203–08, Oct, doi: [10.1109/IRSEC.2014.7059786](https://doi.org/10.1109/IRSEC.2014.7059786).
- Gallardo-Saavedra, S., and B. Karlsson. 2018. Simulation, validation and analysis of shading effects on a PV system. *Solar Energy* 170 (May):828–39. doi:[10.1016/j.solener.2018.06.035](https://doi.org/10.1016/j.solener.2018.06.035).
- Gavhane, P. S., S. Krishnamurthy, R. Dixit, J. P. Ram, and N. Rajasekar. 2017. EL-PSO based MPPT for solar PV under partial shaded condition. *Energy Procedia* 117:1047–53. doi:[10.1016/j.egypro.2017.05.227](https://doi.org/10.1016/j.egypro.2017.05.227).
- Ghosh, S., V. K. Yadav, and V. Mukherjee. 2018. Evaluation of cumulative impact of partial shading and aerosols on different PV array topologies through combined Shannon's entropy and DEA. *Energy* 144:765–75. doi:[10.1016/j.energy.2017.12.040](https://doi.org/10.1016/j.energy.2017.12.040).
- Ghosh, S., V. K. Yadav, and V. Mukherjee. 2019a. Impact of environmental factors on photovoltaic performance and their mitigation strategies—A holistic review. *Renewable Energy Focus* 28 (March):153–72. doi:[10.1016/j.ref.2018.12.005](https://doi.org/10.1016/j.ref.2018.12.005).
- Ghosh, S., V. K. Yadav, and V. Mukherjee. 2019b. Improvement of partial shading resilience of PV array through modified bypass arrangement. *Renewable Energy* 143:1079–93. doi:[10.1016/j.renene.2019.05.062](https://doi.org/10.1016/j.renene.2019.05.062).
- Gökmen, N., W. Hu, P. Hou, Z. Chen, D. Sera, and S. Spataru. 2016. Investigation of wind speed cooling effect on PV panels in windy locations. *Renewable Energy* 90:283–90. doi:[10.1016/j.renene.2016.01.017](https://doi.org/10.1016/j.renene.2016.01.017).
- GREEN, M., K. Emery, Y. Hishikawa, W. Warta, and E. D. Dunlop. 2012. Solar cell efficiency tables (version 40). *IEEE Transactions on Fuzzy Systems* 20 (5):606–14. doi:[10.1002/pip](https://doi.org/10.1002/pip).
- Guerriero, P., F. Di Napoli, G. Vallone, V. Dalessandro, and S. Daliento. 2016. Monitoring and diagnostics of PV plants by a wireless self-powered sensor for individual panels. *IEEE Journal of Photovoltaics* 6 (1):286–94. doi:[10.1109/JPHOTOV.2015.2484961](https://doi.org/10.1109/JPHOTOV.2015.2484961).
- Guerriero, P., Tricoli, S. Daliento, and P. Tricoli. 2019. A bypass circuit for avoiding the hot spot in PV modules. *Solar Energy* 181 (November 2018):430–38. doi:[10.1016/j.solener.2019.02.010](https://doi.org/10.1016/j.solener.2019.02.010).
- Guo, L., Z. Meng, Y. Sun, and L. Wang. 2018. A modified cat swarm optimization based maximum power point tracking method for photovoltaic system under partially shaded condition. *Energy* 144:501–14. doi:[10.1016/j.energy.2017.12.059](https://doi.org/10.1016/j.energy.2017.12.059).
- Gupta, V., M. Sharma, R. Pachauri, and K. N. D. Babu. 2019. Impact of hailstorm on the performance of PV module: A review. *Energy Sources, Part A: Recovery, Utilization and Environmental Effects*. 1–22. doi:[10.1080/15567036.2019.1648597](https://doi.org/10.1080/15567036.2019.1648597).
- Gupta, V., M. Sharma, R. K. Pachauri, and K. N. Dinesh Babu. Oct 2019. Comprehensive review on effect of dust on solar photovoltaic system and mitigation techniques. *Solar Energy* 191:596–622. doi: [10.1016/j.solener.2019.08.079](https://doi.org/10.1016/j.solener.2019.08.079).
- Hadji, S., J. P. Gaubert, and F. Krim. 2018. Real-time Genetic algorithms-based MPPT: Study and comparison (theoretical and experimental) with conventional methods. *Energies* 11 (2). doi:[10.3390/en11020459](https://doi.org/10.3390/en11020459).
- Hanson, A. J., C. A. Deline, S. M. Macalpine, J. T. Stauth, C. R. Sullivan, and S. Member. 2014. "Partial-shading Assessment of photovoltaic Installations via module-level monitoring." *IEEE Journal of Photovoltaics* 4 (6):1618–24. doi:[10.1109/JPHOTOV.2014.2351623](https://doi.org/10.1109/JPHOTOV.2014.2351623).
- Honsberg, C. and S. Bowden. Bypass diodes. <https://www.pveducation.org/pvcdrom/modules-and-arrays/bypass-diodes>.
- Honsberg, C. and S. Bowden. Tandem cells. <https://www.pveducation.org/pvcdrom/tandem-cells>.
- Ibrahim, S. A., A. Nasr, and M. A. Enany. 2021. Maximum power point tracking using ANFIS for a reconfigurable PV-Based Battery Charger under non-uniform operating conditions. *Institute of Electrical and Electronics Engineers Access* 9:114457–67. doi:[10.1109/ACCESS.2021.3103039](https://doi.org/10.1109/ACCESS.2021.3103039).
- Ishaque, K., and Z. Salam. 2013. A review of maximum power point tracking techniques of PV system for uniform insolation and partial shading condition. *Renewable and Sustainable Energy Reviews* 19:475–88. doi:[10.1016/j.rser.2012.11.032](https://doi.org/10.1016/j.rser.2012.11.032).
- Islam, H., S. Mekhilef, N. Shah, T. Soon, M. Seyedmahmoudian, B. Horan, and A. Stojcevski. 2018. Performance evaluation of maximum power point tracking approaches and photovoltaic systems. *Energies* 11 (2):365–69. doi:[10.3390/en11020365](https://doi.org/10.3390/en11020365).
- Jaeeun, K., R. Matheus, P. P. Siva, Y. Hasnain, C. Eun-Chel, and Y. Junsin. Jul, 2021. A review of the degradation of photovoltaic modules for Life Expectancy. *Energies* 14(4278):1–21. doi: [10.3390/en14144278](https://doi.org/10.3390/en14144278).
- Jahn, U. 2019. Partial shading - an overview. <https://www.sciencedirect.com/topics/engineering/partial-shading>.
- Janssen, G. J. M., M. K. Stodolny, B. B. Van Aken, J. Löffler, M. W. P. E. Lamers, K. J. J. Tool, and I. G. Romijn. 2019. Minimizing the Polarization-type potential-induced degradation in PV modules by Modification of the Dielectric Antireflection and Passivation Stack. *IEEE Journal of Photovoltaics* 9 (3):608–14. doi:[10.1109/JPHOTOV.2019.2896944](https://doi.org/10.1109/JPHOTOV.2019.2896944).
- Javed, M. Y., A. F. Mirza, A. Hasan, S. T. H. Rizvi, Q. Ling, M. M. Gulzar, M. U. Safder, and M. Mansoor. 2019. A comprehensive review on a PV based system to harvest maximum power. *Electron* 8 (12):1480. doi:[10.3390/electronics8121480](https://doi.org/10.3390/electronics8121480).



- Jha, V., and U. S. Triar. 2018. Experimental Verification of different photovoltaic array configurations under partial shading condition. *WIECON-ECE 2017 -IEEE International Conference on Advanced Computing and Applications* 2017 (December):43–46. doi:10.1109/WIECON-ECE.2017.8468929.
- Jia, Y., Y. Wang, X. Hu, J. Xu, G. Weng, X. Luo, S. Chen, Z. Zhu, and H. Akiyama. 2021. Diagnosing breakdown mechanisms in monocrystalline silicon solar cells via electroluminescence imaging. *Solar Energy* 225 (July):463–70. doi:10.1016/j.solener.2021.07.052.
- John Bosco, M., and M. Carolin Mabel. 2017. A novel cross diagonal view configuration of a PV system under partial shading condition. *Solar Energy* 158 (October):760–73. doi:10.1016/j.solener.2017.10.047.
- Jordan, D. C., and S. R. Kurtz. 2013. Photovoltaic degradation rates - an Analytical review. *Progress in Photovoltaics: Research and Applications* 21 (1):12–29. doi:10.1002/pip.1182.
- Jordehi, A. R. Nov 2016. Maximum power point tracking in photovoltaic (PV) systems: A review of different approaches. *Renewable and Sustainable Energy Reviews* 65:1127–38. doi: 10.1016/j.rser.2016.07.053.
- Kamal, T., M. Karabacak, S. Z. Hassan, H. Li, and L. M. Fernandez-Ramirez. 2019. A robust online adaptive B-Spline MPPT control of three-phase grid-coupled photovoltaic systems under real partial shading condition. *IEEE Transactions on Energy Conversion* 34 (1):202–10. doi:10.1109/TEC.2018.2878358.
- Kaplanis, E. 2016. Degradation in field-aged crystalline silicon photovoltaic modules and diagnosis using electroluminescence imaging. *8th Int. Work. Teach. Photovoltaics*, no. April. 38–41 [Online]. Available: [https://ueaeprints.uea.ac.uk/58215/1/EKaplanis\\_PV\\_degradation\\_Electroluminescence\\_IWTPV2016.pdf](https://ueaeprints.uea.ac.uk/58215/1/EKaplanis_PV_degradation_Electroluminescence_IWTPV2016.pdf).
- Kaplanis, S., and E. Kaplanis. 2011. Energy performance and degradation over 20 years performance of BP c-Si PV modules. *Simulation Modelling Practice and Theory* 19 (4):1201–11. doi:10.1016/j.simpat.2010.07.009.
- Karmouni, H., M. Chouiekh, S. Motahhir, H. Qjidaa, M. Ouazzani Jamil, and M. Sayyouri. Aug. 2022. A fast and accurate sine-cosine MPPT algorithm under partial shading with implementation using arduino board. *Cleaner Engineering and Technology* 9(August):100535. doi: 10.1016/j.clet.2022.100535.
- Kaushika, N. D., and N. K. Gautam. 2003. Energy yield simulations of interconnected solar PV arrays. *IEEE Transactions on Energy Conversion* 18 (1):127–34. doi:10.1109/TEC.2002.805204.
- Kehang Cui, S. M., G. Dutta, C. Tan, and P. U. Arumugam. 2016. Carbon Nanotube– silicon solar cells. *IEEE Nanotechnology Magazine* 10:12–20. doi:10.1109/MNANO.2016.2572243.
- Kim, K. A. and P. T. Krein. 2013. Hot spotting and second breakdown effects on reverse I-V characteristics for mono-crystalline Si Photovoltaics. 2013 *IEEE Energy Convers. Congr. Expo. ECCE 2013*. 1007–14 doi: 10.1109/ECCE.2013.6646813.
- Kim, K. A., and P. T. Krein. 2015. Reexamination of photovoltaic hot spotting to Show Inadequacy of the bypass diode. *IEEE Journal of Photovoltaics* 5 (5):1435–41. doi:10.1109/JPHOTOV.2015.2444091.
- Kim, K. A., G. S. Seo, B. H. Cho, and P. T. Krein. 2016. Photovoltaic hot-spot Detection for solar panel Substrings using AC Parameter Characterization. *IEEE Transactions on Power Electronics* 31 (2):1121–30. doi:10.1109/TPEL.2015.2417548.
- Kjaer, S. B., J. K. Pedersen, and F. Blaabjerg. Sep 2005. A review of single-phase grid-connected inverters for photovoltaic modules. *IEEE Transactions on Industry Applications* 41 (5):1292–306. doi:10.1109/TIA.2005.853371.
- Kofinas, P., A. I. Dounis, G. Papadakis, and M. N. Assimakopoulos. 2015. An intelligent MPPT controller based on direct neural control for partially shaded PV system. *Energy Buildings* 90:51–64. doi:10.1016/j.enbuild.2014.12.055.
- Köntges, M. et al., *Performance and reliability of photovoltaic systems subtask 3.2: Review of failures of photovoltaic modules: IEA PVPS task 13: External final report IEA-PVPS*. 2014.
- Köntges, M., Oreski, G., and Jahn, U. 2017. *Assessment of photovoltaic module failures in the field*. Report IEA-PVPS T13-09:2017, 1–120. Germany (DEU). Accessed May, 2017. [https://www.google.com/search?q=http%3A%2F%2Fwww.iea-pvps.org.5%E2%80%8B&rlz=1C1GCEB\\_enIN1054IN1054&oq=http%3A%2F%2Fwww.iea-pvps.org.5%E2%80%8B&gs\\_l\\_c\\_r\\_p=EgZjaHJvbWUyBggAEEUYOTIJCAEQIRgKGKABMgkIAhAhGAoYoAHSaQgyNDI1ajBqN6gCALACAA&sourceid=chrome&ie=UTF-8&safe=active](https://www.google.com/search?q=http%3A%2F%2Fwww.iea-pvps.org.5%E2%80%8B&rlz=1C1GCEB_enIN1054IN1054&oq=http%3A%2F%2Fwww.iea-pvps.org.5%E2%80%8B&gs_l_c_r_p=EgZjaHJvbWUyBggAEEUYOTIJCAEQIRgKGKABMgkIAhAhGAoYoAHSaQgyNDI1ajBqN6gCALACAA&sourceid=chrome&ie=UTF-8&safe=active).
- Kumar, D., Y. K. Chauhan, A. S. Pandey, A. K. Srivastava, V. Kumar, F. Alsaif, R. M. Elavarasan, M. R. Islam, R. Kannadasan, M. H. Alsharif, et al. 2023. Mar. A novel hybrid MPPT approach for solar PV systems using Particle-Swarm-optimization-Trained machine learning and Flying Squirrel Search optimization. *Sustainability* 15 (6):5575. doi:10.3390/su15065575.
- Kumar, C. S., C. Puttamadappa, and Y. L. Chandrashekar. 2020. Bacterial Foraging optimization based maximum power point tracking for photovoltaic system under partially shaded condition with Interleaved Resonant Fly-back converter. (15776):15776–84. <http://www.testmagazine.biz/index.php/testmagazine/article/view/3311>
- Lappalainen, K., and S. Valkealahti. 2017. Output power variation of different PV array configurations during irradiance transitions caused by moving clouds. *Applied Energy* 190:902–10. doi:10.1016/j.apenergy.2017.01.013.
- Liu, Y. H., S. C. Huang, J. W. Huang, and W. C. Liang. 2012. A particle swarm optimization-based maximum power point tracking algorithm for PV systems operating under partially shaded conditions. *IEEE Transactions on Energy Conversion* 27 (4):1027–35. doi:10.1109/TEC.2012.2219533.
- Li, X., H. Wen, Y. Hu, L. Jiang, and W. Xiao. 2018. Modified Beta algorithm for GMPPT and partial shading Detection in photovoltaic systems. *IEEE Transactions on Power Electronics* 33 (3):2172–86. doi:10.1109/TPEL.2017.2697459.



- Li, H., D. Yang, W. Su, J. Lu, and X. Yu. 2019. An overall distribution Particle Swarm optimization MPPT algorithm for photovoltaic system under partial shading. *IEEE Transactions on Industrial Electronics* 66 (1):265–75. doi:10.1109/TIE.2018.2829668.
- Luo, W., Y. S. Khoo, J. P. Singh, J. K. C. Wong, Y. Wang, A. G. Aberle, and S. Ramakrishna. 2018. Investigation of potential-induced degradation in n-PERT bifacial silicon photovoltaic modules with a glass/glass structure. *IEEE Journal of Photovoltaics* 8 (1):16–22. doi:10.1109/JPHOTOV.2017.2762587.
- Maarouf, B. B., and Z. H. Ismail. 2022, Feb. Current studies and Applications of shuffled frog Leaping algorithm: A review. *Archives of Computational Methods in Engineering* 1–20. doi:10.1007/s11831-022-09722-x.
- Megantoro, P., H. F. A. Kusuma, L. J. Awal, Y. Afif, D. F. Priambodo, and P. Vigneshwaran. 2022. Comparative analysis of evolutionary-based maximum power point tracking for partial shaded photovoltaic. *International Journal of Electrical and Computer Engineering* 12 (6):5717–29. doi:10.11591/ijece.v12i6.pp5717-5729.
- Mishra, N., A. S. Yadav, R. Pachauri, Y. K. Chauhan, and V. K. Yadav. 2017. Performance enhancement of PV system using proposed array topologies under various shadow patterns. *Solar Energy* 157:641–56. doi:10.1016/j.solener.2017.08.021.
- Mohanty, S., B. Subudhi, and P. K. Ray. 2016. A new MPPT design using grey Wolf optimization technique for photovoltaic system under partial shading conditions. *IEEE Transactions on Sustainable Energy* 7 (1):181–88. doi:10.1109/TSTE.2015.2482120.
- Molin, E., B. Stridh, A. Molin, and E. Wackelgard. 2018. Experimental yield study of bifacial PV modules in nordic conditions. *IEEE Journal of Photovoltaics* 8 (6):1457–63. doi:10.1109/JPHOTOV.2018.2865168.
- Moorthy, J. G., S. Manual, S. M. P. Raja, and P. Raja. 2020. Performance analysis of solar PV based DC optimizer distributed system with simplified MPPT method. *SN Applied Sciences* 2 (2):1–8. doi:10.1007/s42452-020-2010-2.
- Moretón, R., E. Lorenzo, and L. Narvarte. 2015. Experimental observations on hot-spots and derived acceptance/rejection criteria. *Solar Energy* 118:28–40. doi:10.1016/j.solener.2015.05.009.
- M, R., L. S. R. S. A. H., and D. A. 2019. Experimental investigation on the abasement of operating temperature in solar photovoltaic panel using PCM and aluminium. *Solar Energy* 188 (February):327–38. doi:10.1016/j.solener.2019.05.067.
- Murayama, M., and T. Mori. 2007. Dye-sensitized solar cell using novel tandem cell structure. *Journal of Physics D: Applied Physics* 40 (6):1664–68. doi:10.1088/0022-3727/40/6/014.
- Nasiruddin, I., S. Khatoun, M. F. Jalil, and R. C. Bansal. 2019. Shade diffusion of partial shaded PV array by using odd-even structure. *Solar Energy* 181 (March 2018):519–29. doi:10.1016/j.solener.2019.01.076.
- Ngoc, T. N., E. R. Sanseverino, N. N. Quang, P. Romano, F. Viola, B. D. Van, H. N. Huy, T. T. Trong, and Q. N. Phung. 2019. A hierarchical architecture for increasing efficiency of large photovoltaic plants under non-homogeneous solar irradiation. *Solar Energy* 188 (July):1306–19. doi:10.1016/j.solener.2019.07.033.
- Nie, X., and H. Nie. 2017. MPPT control strategy of PV based on Improved shuffled frog Leaping algorithm under complex environments. *Journal of Control Science and Engineering* 2017:1–11. doi:10.1155/2017/2186420.
- Olalla, C., M. N. Hasan, C. Deline, and D. Maksimović. 2018. Mitigation of hot-spots in photovoltaic systems using distributed power electronics. *Energies* 11 (4):1–16. doi:10.3390/en11040726.
- Ordaz, R., J. Solórzano, M. A. Egido, and E. Román. 2011. Analytical study and evaluation results of power optimizers for distributed power conditioning in photovoltaic arrays. *Progress in Photovoltaics: Research and Applications* 21:359–73. doi:10.1002/pip.1188.
- Ovidiu Popescu, S. P. S. 2019. Connecting multiple solar panels – series vs. Parallel. <https://greentumble.com/connecting-multiple-solar-panels/>.
- Pachauri, R. K., I. Kansal, T. S. Babu, and H. H. Alhelou. 2021. Power losses reduction of solar PV systems under partial shading conditions using re-allocation of PV module-fixed electrical connections. *Institute of Electrical and Electronics Engineers Access* 9:94789–812. doi:10.1109/ACCESS.2021.3093954.
- Pachauri, R. K., Mahela, O.P., and Sharma, A. 2020. Impact of partial shading on various PV array configurations and different modeling approaches: A comprehensive review. *Institute of Electrical and Electronics Engineers Access* 8:181375–403. doi:10.1109/ACCESS.2020.3028473.
- Pachauri, R., R. Singh, A. Gehlot, R. Samakaria, and S. Choudhury. 2019. Experimental analysis to extract maximum power from PV array reconfiguration under partial shading conditions. *Engineering Science and Technology an International Journal* 22 (1):109–30. doi:10.1016/j.jestch.2017.11.013.
- Pachauri, R. K., A. Tanwar, S. R. Tripathi, and D. Kumar. 2020. Experimental study on SP and TCT connections of PV modules under realistic shading conditions. *International Journal of Scientific & Technology Research* 9 (2):6319–28.
- Pachauri, R., A. S. Yadav, Y. K. Chauhan, A. Sharma, and V. Kumar. 2018. Shade dispersion-based photovoltaic array configurations for performance enhancement under partial shading conditions. *International Transactions on Electrical Energy Systems* 28 (7):e2556. doi:10.1002/etep.2556.
- Palpandian, M., and P. W. David. A jigsaw puzzle based recon guration technique for enhancing maximum power in partial shaded hybrid photovoltaic array. Research Square. Accessed May 18, 2021. <https://www.researchsquare.com/article/rs-513975/v1>.
- Pendem, S. R., and S. Mikkili. 2018. Modeling, simulation, and performance analysis of PV array configurations (series, series-parallel, Bridge-Linked, and Honey-Comb) to harvest maximum power under various partial shading conditions. *International Journal of Green Energy* 15 (13):795–812. doi:10.1080/15435075.2018.1529577.

- Pendem, S. R., S. Mikkili, and V. V. Katru. 2018. Total-cross-tied configuration of MICs for enhancing the performance of PV D-MPPT systems under various PSCs using PO algorithm. *INDICON 2018 - 15th IEEE India Counc. Int. Conf.* doi:10.1109/INDICON45594.2018.8987172.
- Pervez, I., I. Shams, S. Mekhilef, A. Sarwar, M. Tariq, and B. Alamri. 2021. Most Valuable Player algorithm based maximum power point tracking for a partially shaded PV generation system. *IEEE Transactions on Sustainable Energy* 12 (4):1876–90. doi:10.1109/TSTE.2021.3069262.
- Pillai, D. S., N. Rajasekar, J. P. Ram, and V. K. Chinnaiyan. 2018. Design and testing of two phase array reconfiguration procedure for maximizing power in solar PV systems under partial shade conditions (PSC). *Energy Conversion and Management* 178 (October):92–110. doi:10.1016/j.enconman.2018.10.020.
- Rajput, P., V. T. Shyam, G. N. Tiwari, O. S. Sastry, T. S. Bhatti, and T. S. Bhatti. 2018. A thermal model for N series connected glass/cell/polymer sheet and glass/cell/glass crystalline silicon photovoltaic modules with hot solar cells connected in series and its thermal losses in real outdoor condition. *Renewable Energy* 126:370–86. doi:10.1016/j.renene.2018.03.040.
- Rani, B. I., G. S. Ilango, and C. Nagamani. 2013. Enhanced power generation from PV array under partial shading conditions by shade dispersion using Su Do Ku configuration. *IEEE Transactions on Sustainable Energy* 4 (3):594–601. doi:10.1109/TSTE.2012.2230033.
- Reddy, G. S., T. B. Reddy, and M. V. Kumar. 2017. A MATLAB based PV module models analysis under conditions of nonuniform irradiance. *Energy Procedia* 117:974–83. doi:10.1016/j.egypro.2017.05.218.
- Reinders, A., M. G. Debije, and A. Rosemann. 2017. Measured efficiency of a luminescent solar concentrator PV module called leaf roof. *IEEE Journal of Photovoltaics* 7 (6):1663–66. doi:10.1109/JPHOTOV.2017.2751513.
- Rodriguez-Gallegos, C. D., O. Gandhi, J. M. Y. Ali, V. Shanmugam, T. Reindl, and S. K. Panda. 2019. On the grid metallization optimization design for monofacial and bifacial Si-based PV modules for real-world conditions. *IEEE Journal of Photovoltaics* 9 (1):112–18. doi:10.1109/JPHOTOV.2018.2882188.
- Sarvi, M., I. Soltani, and I. N. Avanaki. 2014. A water cycle algorithm maximum power point tracker for photovoltaic energy conversion system under partial shading. *Applied Mathematics and Mechanics* 2 (1):103–16.
- Sarwar, S., M. Y. Javed, M. H. Jaffery, J. Arshad, A. Ur Rehman, M. Shafiq, and J.-G. Choi. 2022. A novel hybrid MPPT technique to maximize power harvesting from pv system under partial and complex partial shading. *Applied Science* 12 (2):587. doi:10.3390/app12020587.
- Saur News Bureau. Power optimizer: Smart module technology. <https://www.saurenergy.com/solar-energy-articles/power-optimizer-smart-module-technology>.
- Schulte-Huxel, H., R. Witteck, H. Holst, M. R. Vogt, S. Blankemeyer, D. Hinken, T. Brendemuhl, T. Dullweber, K. Bothe, M. Kontges, et al. 2017. High-efficiency modules with passivated emitter and rear solar cells—an analysis of electrical and optical losses. *IEEE Journal of Photovoltaics* 7(1):25–31. doi:10.1109/JPHOTOV.2016.2614121.
- Serna-Garcés, S. I., J. D. Bastidas-Rodríguez, and C. A. Ramos-Paja. 2016. Reconfiguration of urban photovoltaic arrays using commercial devices. *Energies* 9 (1):1–23. doi:10.3390/en9010002.
- Shams El-Dein, M. Z., M. Kazerani, and M. M. A. Salama. 2013. An optimal total cross tied interconnection for reducing mismatch losses in photovoltaic arrays. *IEEE Transactions on Sustainable Energy* 4 (1):99–107. doi:10.1109/TSTE.2012.2202325.
- Sher, H. A., A. F. Murtaza, A. Noman, K. E. Addoweesh, K. Al-Haddad, and M. Chiaberge. Oct, 2015. A new sensorless hybrid MPPT algorithm based on fractional short-circuit current measurement and P&O MPPT. *IEEE Transactions on Sustainable Energy* 6(4):1426–34. doi:10.1109/TSTE.2015.2438781.
- Sher, H. A., A. F. Murtaza, A. Noman, K. E. Addoweesh, and M. Chiaberge. Jan, 2015. An intelligent control strategy of fractional short circuit current maximum power point tracking technique for photovoltaic applications. *Journal of Renewable and Sustainable Energy* 7(1):013114. doi:10.1063/1.4906982.
- Simon, M., and E. L. Meyer. 2010. Detection and analysis of hot-spot formation in solar cells. *Solar Energy Materials and Solar Cells* 94 (2):106–13. doi:10.1016/j.solmat.2009.09.016.
- Singh, N., K. K. Gupta, S. K. Jain, N. K. Dewangan, and P. Bhatnagar. Aug 2021. A flying squirrel search optimization for MPPT under partial shaded photovoltaic system. *IEEE Journal of Emerging and Selected Topics in Power Electronics* 9 (4):4963–78. doi:10.1109/JESTPE.2020.3024719.
- Sivashankar, M., C. Selvam, and S. Manikandan. 2021. A review on the selection of phase change materials for photovoltaic thermal management. *IOP Conference Series: Materials Science & Engineering* 1130 (1):012026. doi:10.1088/1757-899x/1130/1/012026.
- SolarEdge Solution. Single phase inverters with HD-wave technology. <https://www.solaredge.com/products/pv-inverter/sin-gle-phase/>.
- Solheim, H. J., H. G. Fjær, E. A. Sørheim, and S. E. Foss. 2013. Measurement and simulation of hot spots in solar cells. *Energy Procedia* 38 (1876):183–89. doi:10.1016/j.egypro.2013.07.266.
- Sridhar, R., S. Jeevananthan, S. S. Dash, and P. Vishnuram. 2017. A new maximum power tracking in PV system during partially shaded conditions based on shuffled frog leap algorithm. *Journal of Experimental & Theoretical Artificial Intelligence* 29 (3):481–93. doi:10.1080/0952813X.2016.1186750.
- Suresh Kumar, E., B. Sarkar, and N. K. S. 2014. Quality improvement of PV modules by electroluminescence and thermal imaging. *International Journal of Engineering and Advanced Technology* 3 (3):422–27. [Online]. Available. <http://www.ijeat.org/attachments/File/v3i3/C2698023314.pdf>.

- Tan, J. D., S. P. Koh, S. K. Tiong, K. Ali, and Y. Y. Koay. 2018. An electromagnetism-like mechanism algorithm approach for photovoltaic system optimization. *Indonesian Journal of Electrical Engineering and Computer Science* 12 (1):333–40. doi:10.11591/ijeecs.v12.i1.pp333-340.
- Teo, J. C., R. H. G. Tan, V. H. Mok, V. K. Ramachandaramurthy, and C. Tan. 2018. Impact of partial shading on the P-V characteristics and the maximum power of a photovoltaic string. *Energies* 11 (7). doi:10.3390/en11071860.
- Tey, K. S., and S. Mekhilef. 2014. Modified incremental conductance MPPT algorithm to mitigate inaccurate responses under fast-changing solar irradiation level. *Solar Energy* 101:333–42. doi:10.1016/j.solener.2014.01.003.
- TIGO. 2019. Maximizing energy harvest the roles of predictive IV and impedance matching in PV array optimization. [Online]. Available: [https://assets-global.website-files.com/5fad551d7419c7a0e9e4aba4/600eef99c91ec4bec971789\\_Optimization with Impedance Matching and Predictive IV.pdf](https://assets-global.website-files.com/5fad551d7419c7a0e9e4aba4/600eef99c91ec4bec971789_Optimization%20with%20Impedance%20Matching%20and%20Predictive%20IV.pdf).
- Tsanakas, J. A., D. Chrysostomou, P. N. Botsaris, and A. Gasteratos. 2015. Fault diagnosis of photovoltaic modules through image processing and Canny edge detection on field thermographic measurements. *International Journal of Sustainable Energy* 34 (6):351–72. doi:10.1080/14786451.2013.826223.
- Tsang, K. M., and W. L. Chan. 2015. Maximum power point tracking for PV systems under partial shading conditions using current sweeping. *Energy Conversion and Management* 93:249–58. doi:10.1016/j.enconman.2015.01.029.
- Tubniyom, C., W. Jaideaw, R. Chatthaworn, A. Suksri, and T. Wongwuttanasatian. 2018. Effect of partial shading patterns and degrees of shading on total cross-tied (TCT) photovoltaic array configuration. *Energy Procedia* 153:35–41. doi:10.1016/j.egypro.2018.10.028.
- Venkateswari, R., and S. Sreejith. 2019. Factors influencing the efficiency of photovoltaic system. *Renewable and Sustainable Energy Reviews* 101 (November 2018):376–94. doi:10.1016/j.rser.2018.11.012.
- Vogt, M. R., H. Schulte-Huxel, M. Offer, S. Blankemeyer, R. Witteck, M. Kontges, K. Bothe, and R. Brendel. 2017. Reduced module operating temperature and increased yield of modules with PERC instead of al-BSF solar cells. *IEEE Journal of Photovoltaics* 7 (1):44–50. doi:10.1109/JPHOTOV.2016.2616191.
- Wang, Y. J., and P. C. Hsu. 2011. An investigation on partial shading of PV modules with different connection configurations of PV cells. *Energy* 36 (5):3069–78. doi:10.1016/j.energy.2011.02.052.
- Waqar Akram, M., G. Li, Y. Jin, X. Chen, C. Zhu, X. Zhao, M. Aleem, and A. Ahmad. 2019. Improved outdoor thermography and processing of infrared images for defect detection in PV modules. *Solar Energy* 190 (March):549–60. doi:10.1016/j.solener.2019.08.061.
- Winston, D. P. 2019. Efficient output power enhancement and protection technique for hot spotted solar photovoltaic modules. *IEEE Transactions on Device and Materials Reliability* 19 (4):664–70. doi:10.1109/TDMR.2019.2945194.
- Yadav, A. S., and V. Mukherjee. 2018. Line losses reduction techniques in puzzled PV array configuration under different shading conditions. *Solar Energy* 171 (June):774–83. doi:10.1016/j.solener.2018.07.007.
- Yadav, A. S., and V. Mukherjee. 2021. Conventional and advanced PV array configurations to extract maximum power under partial shading conditions: A review. *Renewable Energy* 178:977–1005. doi:10.1016/j.renene.2021.06.029.
- Yadav, A. S., R. K. Pachauri, and Y. K. Chauhan. 2016. Comprehensive investigation of PV arrays with puzzle shade dispersion for improved performance. *Solar Energy* 129:256–85. doi:10.1016/j.solener.2016.01.056.
- Yadav, A. S., R. K. Pachauri, Y. K. Chauhan, S. Choudhury, and R. Singh. 2017. Performance enhancement of partially shaded PV array using novel shade dispersion effect on magic-square puzzle configuration. *Solar Energy* 144:780–97. doi:10.1016/j.solener.2017.01.011.
- Yamaguchi, M., F. Dimroth, J. F. Geisz, and N. J. Ekins-Daukes. 2021. Multi-junction solar cells paving the way for super high-efficiency. *Journal of Applied Physics* 129 (24). doi: 10.1063/5.0048653.
- Yerkar, S. A., M. C. Bisane, and D. Waghchore. 2017. Carbon nanotubes in solar panel technology. *IETE Zonal Seminar "Recent Trends in Engineering & Technology" - 2017 Special Issue of International Journal of Electronics, Communication & Soft Computing Science and Engineering* 93–97.
- Zhang, Y., Y. Yu, F. Meng, and Z. Liu. 2020. Experimental investigation of the shading and mismatch effects on the performance of bifacial photovoltaic modules. *IEEE Journal of Photovoltaics* 10 (1):296–305. doi:10.1109/JPHOTOV.2019.2949766.
- Zheng, H., S. Li, T. A. Haskew, and Y. Xiao. 2013. Impact of uneven shading and bypass diodes on energy extraction characteristics of solar photovoltaic modules and arrays. *International Journal of Sustainable Energy* 32 (5):351–65. doi:10.1080/14786451.2012.709857.
- Zhou, Y. P., Y. L. He, Y. Qiu, Q. Ren, and T. Xie. 2017. Multi-scale investigation on the absorbed irradiance distribution of the nanostructured front surface of the concentrated PV-TE device by a MC-FDTD coupled method. *Applied Energy* 207:18–26. doi:10.1016/j.apenergy.2017.05.115.
- Zhu, Z. 2018. MPPT control method for photovoltaic system based on particle swarm optimization and bacterial foraging algorithm. *International Journal of Electrical Components and Energy Conversion* 4 (1):45. doi:10.11648/j.ijecec.20180401.15.

ACCEPTED MANUSCRIPT • OPEN ACCESS

# Effects of Surface Modified Recycled Coarse Aggregates on Concrete's Mechanical Characteristics

To cite this article before publication: Harish Panghal *et al* 2023 *Mater. Res. Express* in press <https://doi.org/10.1088/2053-1591/acf915>

## Manuscript version: Accepted Manuscript

Accepted Manuscript is "the version of the article accepted for publication including all changes made as a result of the peer review process, and which may also include the addition to the article by IOP Publishing of a header, an article ID, a cover sheet and/or an 'Accepted Manuscript' watermark, but excluding any other editing, typesetting or other changes made by IOP Publishing and/or its licensors"

This Accepted Manuscript is © 2023 The Author(s). Published by IOP Publishing Ltd.



As the Version of Record of this article is going to be / has been published on a gold open access basis under a CC BY 4.0 licence, this Accepted Manuscript is available for reuse under a CC BY 4.0 licence immediately.

Everyone is permitted to use all or part of the original content in this article, provided that they adhere to all the terms of the licence <https://creativecommons.org/licenses/by/4.0>

Although reasonable endeavours have been taken to obtain all necessary permissions from third parties to include their copyrighted content within this article, their full citation and copyright line may not be present in this Accepted Manuscript version. Before using any content from this article, please refer to the Version of Record on IOPscience once published for full citation and copyright details, as permissions may be required. All third party content is fully copyright protected and is not published on a gold open access basis under a CC BY licence, unless that is specifically stated in the figure caption in the Version of Record.

View the [article online](#) for updates and enhancements.

# Effects of Surface Modified Recycled Coarse Aggregates on Concrete's Mechanical Characteristics

Harish Panghal<sup>1\*</sup>, Awadhesh Kumar<sup>2</sup>

<sup>1</sup>Ph.D Student, Department of Civil Engineering, Delhi Technological University, Delhi, 1100042, India  
Email: [harish\\_phd2k18@dtu.ac.in](mailto:harish_phd2k18@dtu.ac.in)

<sup>2</sup>Professor, Department of Civil Engineering, Delhi Technological University, Delhi, 1100042, India  
Email: [awadheshg@dtu.ac.in](mailto:awadheshg@dtu.ac.in)

\*CORRESPONDENCE: Harish [harish\\_phd2k18@dtu.ac.in](mailto:harish_phd2k18@dtu.ac.in)

Department of Civil Engineering, Delhi Technological University, Delhi, 1100042, India

## ABSTRACT

Sustainable concrete using recycled coarse aggregates from construction and demolition waste is gaining popularity in the construction industry, but has poor mechanical characteristics due to old cement mortar adhering to aggregate surfaces. This study uses two processes (abrasion treatment and cement slurry treatment) to modify the surface of recycled coarse aggregates (RCA) to minimize the strength loss of RCA and enhance the bonding properties of the concrete matrix and RCA. Surface-modified RCA replaced coarse aggregates in varying percentages, ranging from 0 to 100% in 25% increments. To comprehend the effects of surface-modified RCA, the workability, compressive strength, flexural strength, split tensile strength, microstructural characteristics (XRD, SEM, and EDAX), and modulus of elasticity of concrete are evaluated. Surface-modified RCA improves concrete's mechanical characteristics, but abrasion-treated RCA has significantly greater strength than reference concrete up to 50% replacement level, while cement slurry treatment has slightly lower strength. Test findings reveal that among all the two processes of surface modifications of RCA, abrasion treatment is more effective and efficient. At 100% replacement level, surface-modified RCA by abrasion treatment reduces compressive, flexural, and split tensile strength by 10.89%, 10.42%, and 09.92% compared to reference concrete, while surface-modified RCA by cement slurry treatment reduces these values by 14.80%, 13.27%, and 12.76%. Surface modifications improve bonding properties of RCA and cement matrix, reducing porosity and resulting in dense and strong ITZs compared to unmodified RCA.

**Keywords:** Mechanical Characteristics; Recycled Coarse Aggregates; Scanning Electron Microscopy; Surface Modification; X-Ray Diffraction

## 1. Introduction

Construction and demolition (C&D) waste generation is expected to reach 2.59 billion tonnes by 2030 and 3.40 billion tonnes by 2050 [1], causing environmental issues and a lack of disposal sites [2]. Recycling C&D waste as recycled aggregates (RA) is economically and environmentally advantageous [3]. Recycled aggregates are produced by processing construction and demolition waste materials like concrete, asphalt, bricks, and tiles [4][5]. These aggregates can be utilized to create both coarse and fine aggregates for reuse [6]. In this study, we employ recycled coarse aggregate (RCA) produced by using an impact crusher to crush waste concrete from IL&FS C&D Waste Recycling Plant, Delhi Metro Rail Corporation (DMRC), Delhi, India [7][8]. However, the quality of RCA is affected by cement mortar attached to the aggregate surface [9]. Proper removal or

reduction of cement mortar during recycling is crucial for high-quality RCA production [10]. Efficient methods for separating and cleaning aggregates can enhance performance and contribute to sustainable construction practices [11]. Experimental studies aim to enhance RCA concrete's strength, durability, and performance by incorporating additives and techniques [12][13][14].

Researchers investigate factors like chemical admixtures, fiber reinforcement, and curing methods to optimize the material's characteristics for sustainable construction practices [15]. The two-stage mixing method improves recycled aggregate interfacial zones by filling pores and cracks, resulting in dense concrete [16]. Pre-soaking of recycled aggregates with HCl, H<sub>2</sub>SO<sub>4</sub>, and H<sub>3</sub>PO<sub>4</sub> reduces water absorption without exceeding permissible limits for chloride and sulphate components [17]. A new mixing technique and stone-involved pozzolanic powder coating of recycled coarse aggregates enhance ITZ structure, achieving better workability and strength [18]. The treatment of RCA by soaking in pozzolanic materials enhances its mechanical properties [19]. Incorporating 25–30% fly ash improves the mechanical characteristics of recycled aggregate concrete [20]. The treatment of RCA with a sodium silicon-based polymer enhances its fragmentation resistance and decreases its water absorption capacity [21]. The particle density, water absorption, and mechanical strength of RCA are significantly improved by coating with calcium meta-silicate solution and soaking in HCL acid at a 0.5 mol concentration [22]. Carbonation-based surface modification of RCA increases density and reduces water absorption [23] while enhancing compressive strength [24]. Acetic acid solution treatment of RCA increases concrete's compressive strength by up to 25% within 28 days [25]. In comparison to untreated RCA, surfaces treated with pozzolanic slurry and CO<sub>2</sub> had superior mechanical strength and were more resistant to carbonation and chloride ion diffusion [26]. Recycled concrete aggregate (RCA) can be improved by reducing mortar attachment and using mineral admixtures as internal curing agents. This leads to increased mechanical strength and durability in RCA, with potential cost savings compared to natural aggregate (NA) mixes [27]. Cement mortar density improved by 5.7% after limewater-CO<sub>2</sub> treatment, reducing water absorption by 50%. Compressive and flexural strength improved by 22.8% and 42.4%, respectively, while total porosity decreased by 33% [28]. The combination of crushing and carbonation treatment increased crushing stress and decreased RCA's water absorption [29]. The accelerated carbonation process reduced cement mortar's water absorption, sorptivity, totally charged passed, and chlorine ions diffusion coefficients [30]. Adding treated RCA to a 0.1 mol HCL solution and coating with calcium meta-silicate slurry increased drop and brittle behaviour during post-cracking extension. TRCA improved crack resistance in concrete's processing zone for fractures compared to RCA [31].

Simultaneously, the research delves into the impact of carbonated recycled fine aggregate (CRFA) and recycled fine aggregate (RFA) on the properties of alkali-activated slag and glass powder mortar. An elevated RFA content contributes to a notable increase in compressive strength within the mortar. On the other hand, a higher CRFA content leads to an enhanced flow value, prolonged setting time, and decreased strength [32]. The study also sheds light on a strong correlation between reduced water absorption and heightened compressive strength in recycled concrete. This correlation is influenced by treatment methods, particle sizes, and processes [33]. Moreover, subjecting recycled concrete aggregate to an acid-mechanical treatment yields a remarkable 16.06% enhancement in compressive strength. This treatment positively affects various aspects such as surface density, transition zones, micro-cracks, pores, and mortar quality, consequently elevating concrete excellence



and reducing sorptivity [34]. Concrete's environmental impact prompts sustainable solutions—recycling, substituting Ordinary Portland Cement and natural aggregates with by-products like fly ash. The review covers advanced methods like optimizing cement hydration, introducing novel materials like carbon nanotubes, aiming for eco-friendly, sustainable concrete [35]. Study explores coal fly ash (CFA) use in concrete, replacing cement. Testing nano-silica (nS)-enhanced CFA with 5% nS and 0%, 15%, 25% CFA levels shows improved mechanical properties and microstructure. Optimal blend (5% nS, 15% CFA) enhances strengths by 37.68% and 36.21%, offering potential for eco-friendly concrete [36]. Research explores superabsorbent polymer (SAP) and expansive agent (EA) in ultra-high-performance concrete (UHPC) for shrinkage reduction and strength retention. Optimal blend (S1E1, 0.1% SAP, 1% CEA) achieves high strength (135 MPa) and significant shrinkage reduction (24%) in 7 days. Study highlights self-desiccation water retention, portlandite formation's interplay for UHPC shrinkage control. Ongoing hydration reduces microporosity, compacting microstructure, revealing SAP-induced voids limiting CEA expansion [37]. Six papers in this special issue focus on concrete with industrial waste and environmentally friendly variants. These cutting-edge studies aim to innovate the fabrication, properties, and development of eco-friendly concrete [38]. Study examines NaCl and gypsum influence on geopolymer concrete activated by quicklime. Tests show individual NaCl or gypsum boosts compressive strength (up to 245.3% at 3 days), combined generates salts, enhancing strength by 180.3%. 2% NaCl reduces mass loss, improves elastic modulus. 4% NaCl, 7.5% gypsum improves sulfate corrosion resistance by 38.8%. Tailored geopolymer blend reduces costs, emissions, improves corrosion resistance compared to slag Portland cement [39].

Recycling C&D waste as recycled aggregates is becoming popular for sustainable concrete production, reducing landfill waste, conserving natural resources, and supporting the circular economy [40]. However, old adhered cement mortar on recycled aggregates can negatively impact strength and durability, making proper removal crucial for optimal performance [41]. This study aims to minimize the strength loss of recycled coarse aggregates (RCA) and enhance bonding between RCA and concrete matrix. Previous studies have shown similar mechanical characteristics loss due to unmodified surfaces or other treatments [15]. This study analyses the mechanical characteristics after surface modification using abrasion treatment and cement slurry treatment, aiming to minimize strength loss.

The novelties of this work are as follows:

1. Investigating a gap by implementing two distinct surface modification techniques on Recycled Coarse Aggregates (RCA): Abrasion Treatment and Cement Slurry Treatment. These processes elucidate the impact of surface modifications on the properties of RCA.
2. Evaluating the effective utilization of surface-modified RCA in the production of eco-friendly concrete, serving as a substitute for natural coarse aggregates at varying proportions (0%, 25%, 50%, 75%, and 100%).
3. Analysing the effects of different proportions of surface-modified RCA on the fresh characteristics of concrete (Workability), its hardened (compressive strength, flexural strength, split tensile strength), and its microstructural attributes (XRD, SEM, and EDAX). A comparative study with a control mixture is conducted to identify noteworthy distinctions.

## 2. Materials and Testing

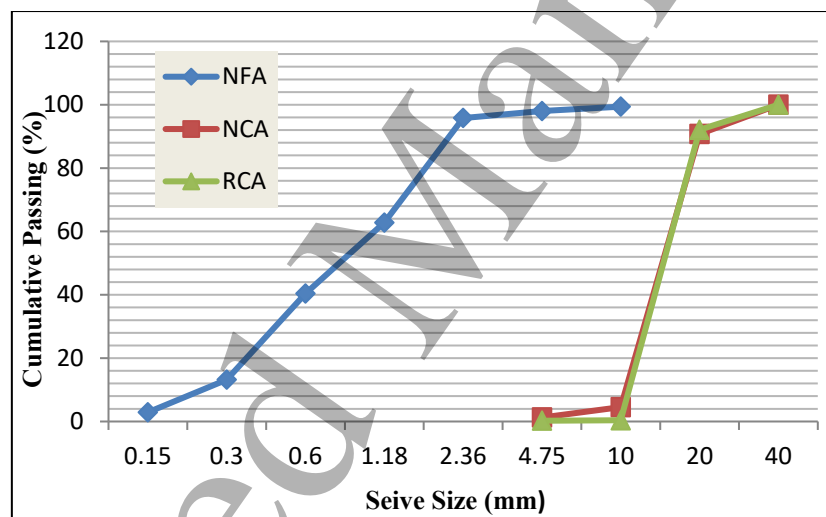
### 2.1 Material

The study uses ordinary Portland cement of grade 43, confirming IS 269-2015 [42]. Physical test results are presented in Table 1.

**Table 1** Physical Test Results on Cement

Types	Value Measured	As per IS 269-2015 [42]
Consistency	31%	-
Initial Setting Time	58	> 30 Minutes
Final setting Time	435	< 10 Hours
Specific Gravity	3.11	3.0 to 3.15

Natural sand with a 4.75 mm down size is used as a fine aggregate confirming IS 383-2016 (Reaffirmed 2021) [43]. Crushed stone aggregates of size 4.75 mm to 20 mm are used as natural coarse aggregates in mixed proportion confirming to IS 383-2016 (Reaffirmed 2021) [43]. With an impact crusher, crushed concrete from C&D waste is reduced to the necessary size and used as recycled coarse aggregates. Figure 1 depicts the aggregate particle size distribution.



**Figure 1** Gradation Curve for Aggregates of Different Mixtures

Analysis of physical and mechanical properties is crucial for determining RCA compatibility with concrete. Table 2 displays the physical and mechanical characteristics of aggregates. The RCA sample results revealed poor bonding due to old cement mortar on aggregate surfaces which causes a loss in strength. To address this issue, surface modification is necessary before adding RCA to concrete. Surface modification is needed to improve the connection between RCA particles and the cement matrix. RCA treated with abrasion has a rough surface, similar to coarse aggregates, while treated with cement slurry has a smooth surface. This treatment strengthens the bond with the cement matrix and improves the mechanical properties of concrete. Chemical admixture as super-plasticizers (C-MAX) is used 1% by weight of cement, and potable water is used for mixing and curing.

**Table 2** Physical and Mechanical Characteristics of Aggregates

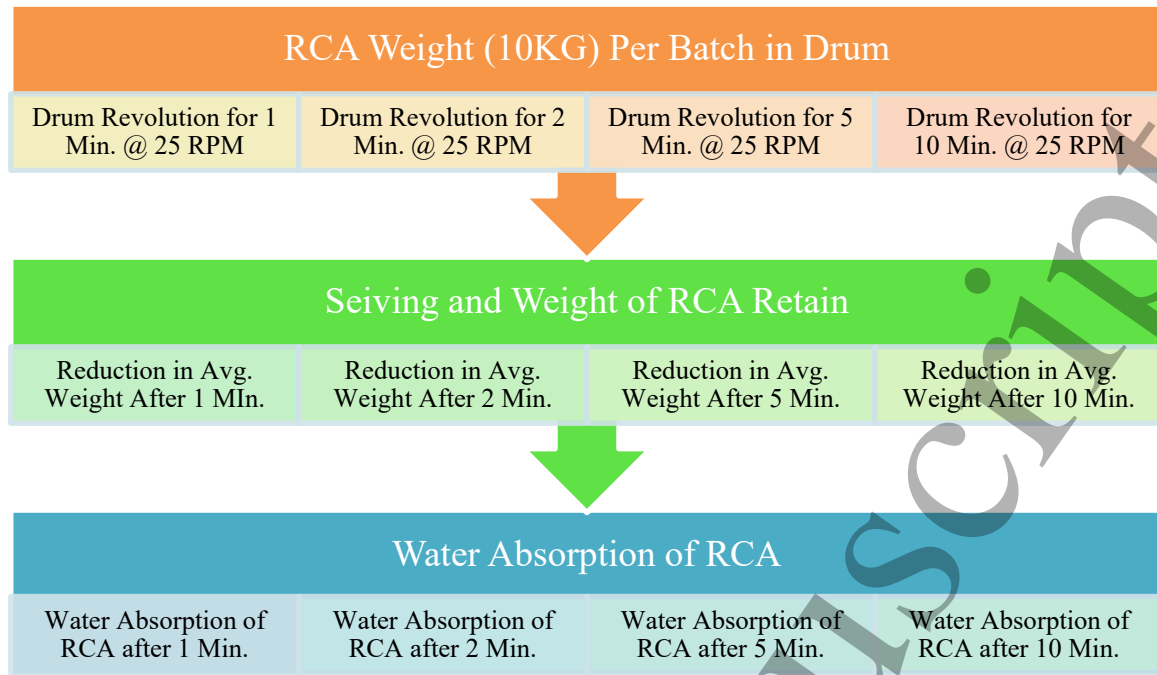
Property	NFA	NCA	RCA	RCAAT	RCACST	Standard Limits
Bulk Density (kg/m <sup>3</sup> )	1625	1740	1660	1685	1704	1200-1750
Specific gravity	2.675	2.754	2.681	2.721	2.708	2.30-2.90
Water Absorption (%)	0.51	1.12	2.35	1.87	2.05	≤ 2.0 (IS 2386 Part 3) [44]
Abrasion Loss (%)	15.52	25.43	28.76	26.13	27.46	< 30 (IS 2386 Part 4) [45]
Crushing Value (%)	16.21	26.24	27.71	24.36	26.36	< 30 (IS 2386 Part 4) [45]
Impact Value (%)	15.31	17.31	20.68	14.23	15.26	< 30 (IS 2386 Part 4) [45]
NFA- Natural Fine Aggregates, NCA- Natural Coarse Aggregates, RCA- Recycled Coarse Aggregates, RCAAT- Recycled Coarse Aggregates with Abrasion Treatment, RCACST- Recycled Coarse Aggregates with Cement Slurry Treatment						

**2.2 Surface Modification of RCA**

Recycled aggregate concrete (RAC) suffers from strength reduction due to weak bonding between RCA and cement matrix. Researchers found significant losses in RAC's mechanical characteristics when using unmodified coarse aggregates. This is due to the reduction in strength due to weak bonding. Old mortar on RCA surfaces causes weak bonding with the cement matrix, requiring treatment to enhance compressive strength. This experimental study uses abrasion treatment and cement slurry treatment to modify RCA surfaces for improved properties.

**2.2.1 Abrasion Treatment of RCA**

An abrasion treatment method was used to reduce the quantity of mortar that was adhered to the surface of the RCA. A Los Angeles Abrasion machine is utilized in this procedure; it comprises a hollow steel cylinder that is closed at both ends, has an internal diameter of 711 mm, and can revolve around its horizontal axis. The device was kept spinning at a speed of 25 revolutions per minute for 5 minutes while being filled with coarse recycled aggregates. The rotating drum causes aggregate particles to rub against each other, removing any attached mortar in the process. The surface modification process from the abrasion treatment is depicted in Figure 2.



**Figure 2** Process of surface modification from abrasion treatment of the RCA

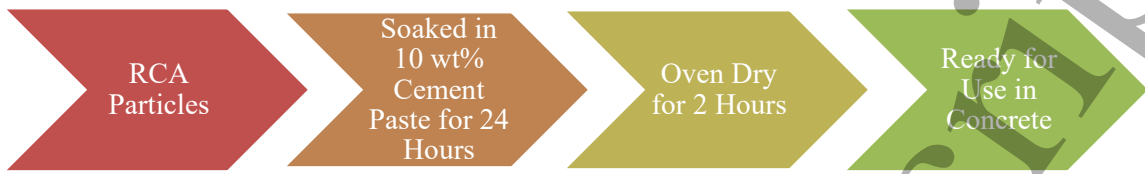
Numerous abrasion machine RPM trials were conducted to optimize drum rotation duration. Table 3 displays the outcomes of the trials. The percentage of RCA's that could absorb water after treatment was a criterion used to choose the rotation of the drum. The percentage of RCAs able to absorb water after treatment was used as a criterion. The treated materials absorbed 1.87% water after 5 minutes of revolutions, making 5 minutes the optimal treatment time for RA.

**Table 3** Results of the Trails of Drum Rotation and RCA Percentages

S. No.	Weight of RCA (KG)	Weight of RCA After 1 Minute	Reduction in Weight of RCA After 1 Minute (%)	Weight of RCA After 2 Minute	Reduction in Weight of RCA After 2 Minute (%)	Weight of RCA After 5 Minute	Reduction in Weight of RCA After 5 Minute (%)	Weight of RCA After 10 Minute	Reduction in Weight of RCA After 10 Minute (%)
1	10	9.63	3.7	9.07	09.25	8.57	14.25	8.06	19.4
2	10	9.39	6.1	8.83	11.71	8.57	14.30	8.12	18.8
3	10	9.38	6.2	8.88	11.15	8.77	12.30	8.07	18.0
4	10	9.67	3.3	8.71	12.25	8.82	11.75	7.91	20.9
5	10	9.57	4.3	8.77	12.31	8.81	11.90	8.17	19.1
6	10	9.58	3.7	9.02	09.25	8.52	14.25	8.01	19.4
7	10	9.32	6.1	8.78	11.72	8.52	14.30	8.07	18.8
8	10	9.33	6.2	8.83	11.15	8.72	12.30	8.02	18.0
9	10	9.62	3.3	8.72	12.25	8.57	11.75	7.86	20.9
10	10	9.52	4.3	8.72	12.31	8.76	11.90	8.12	19.1
Average Reduction in Weight (%)		04.72		11.33		12.90		15.49	
Water Absorption (%)		02.34		02.16		01.87		01.98	

### 2.2.2 Cement Slurry Treatment of RCA

The technique involves creating cement paste using water, dissolved in 10% water, and agitated for 10-15 minutes. Recycled aggregates are immersed in the paste for 24 hours, and then dried in the oven for optimal particle penetration. This dried recycled aggregate is used in concrete preparations. The process of surface modification from cement slurry treatment of the RCA is shown in Figure 3.



**Figure 3** Process of surface modification from cement slurry treatment of the RCA

### 3. Mix Proportions

Nine concrete mixtures were produced for 27MPa target strength to investigate the mechanical behavior of surface-modified recycled coarse aggregates. Table 4 lists all of the compositions of the concrete mixtures. Reference mixture (RC) is created with natural aggregates, and mixtures RCAAT-25, RCAAT-50, RCAAT-75, and RCAAT-100 are produced by replacing natural coarse aggregates with abrasion modifies recycled coarse aggregates and RCACST-25, RCACST-50, RCACST-75, and RCACST-100 are produced by replacing natural coarse aggregates with cement slurry modifies recycled coarse aggregates at varying replacement percentages of 25%, 50%, 75%, and 100% respectively. All mixtures were produced using the weight batching method at a constant water-cement ratio of 0.50.

**Table 4** Composition of Concrete Mixtures

Mix. No.	Mixture ID	NFA (kg/m <sup>3</sup> )	NCA (kg/m <sup>3</sup> )	RCA (kg/m <sup>3</sup> )	Cement (kg/m <sup>3</sup> )	W/C Ratio	Admixture (%)	Slump (mm)
M1	RC	444.48	1511	0	400	0.5	4	116
M2	RCAAT 25	444.48	377.75	1133.25	400	0.5	4	109
M3	RCAAT 50	444.48	755.5	755.5	400	0.5	4	103
M4	RCAAT 75	444.48	1133.25	377.75	400	0.5	4	92
M5	RCAAT 100	444.48	0	1511	400	0.5	4	81
M6	RCACST 25	444.48	377.75	1133.25	400	0.5	4	106
M7	RCACST 50	444.48	755.5	755.5	400	0.5	4	100
M8	RCACST 75	444.48	1133.25	377.75	400	0.5	4	88
M9	RCACST 100	444.48	0	1511	400	0.5	4	75

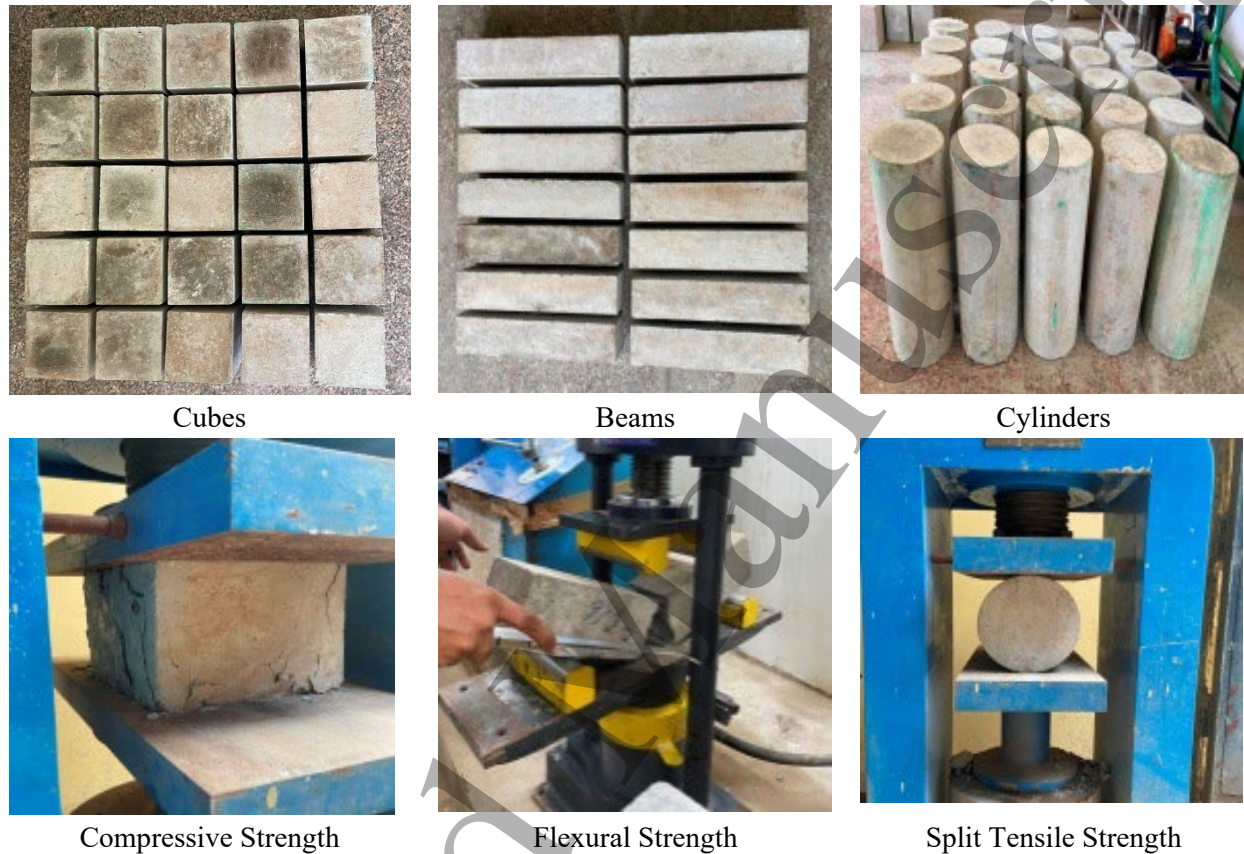
RC-Reference Concrete, NFA-Natural Fine Aggregates, NCA-Natural Coarse Aggregates, RCA-Recycled Coarse Aggregates, RCAAT- Recycled Coarse Aggregates with Abrasion Treatment, RCACST- Recycled Coarse Aggregates with Cement Slurry Treatment

### 4. Testing Programs

The study examines the mechanical characteristics of structural concrete by measuring its workability and hardened characteristics. 162 samples were cast, cured, and tested in steel cube (15×15×15 cm) molds for compressive strength, rectangular (50×10×10 cm) molds for flexural



strength, and cylindrical ( $\phi = 15$  cm, height ( $H$ ) = 30 cm) molds for split tensile strength. The strength of concrete mixtures was determined using the average of three specimens (as shown in Figure 4) for each mix. These characteristics are measured at two different intervals of time, 7 and 28 days. The modulus of elasticity was measured on cylindrical (150×300 mm) specimens with a height/diameter ratio of 2.0 as per IS 516-1959 [46]. Scanning electron microscopy (SEM), energy-dispersive X-ray spectroscopy (EDAX), and X-ray diffraction (XRD) were used to examine the micro-structural characteristics of concrete samples for different mixtures.



**Figure 4** Different Specimens and their Testing

## 5. RESULTS AND DISCUSSION

### 5.1 Workability

In order to evaluate the feasibility of mixes created using varying replacement percentages of surface-modified recycled coarse aggregates (RCA), a slump test was conducted according to the guidelines of IS 1199-1959 [47]. The results of the slump test, illustrating the variations in slump for different concrete mixtures, are presented in Figure 5. The observed trend in Figure 4 indicates a reduction in the slump of concrete as the proportion of surface-modified recycled coarse aggregates increases. The workability of concrete containing surface-modified RCA falls within the medium range, spanning from 50 to 100 mm, as evidenced by the slump values across all mixtures. This decline in slump value is attributed to the greater water absorption capacity of RCA compared to natural aggregates, resulting from their rougher surface texture and larger surface area. Comparing the concrete mixes prepared using surface-modified RCA subjected to abrasion treatment and those



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

treated with a cement slurry process, it is noted that the former exhibits a slightly higher slump value. This suggests that the abrasion treatment imparts a relatively more favourable workability to the concrete mixes. The consistent trend of diminishing slump with an increasing proportion of surface-modified RCA aligns with findings from a prior study [48][49].

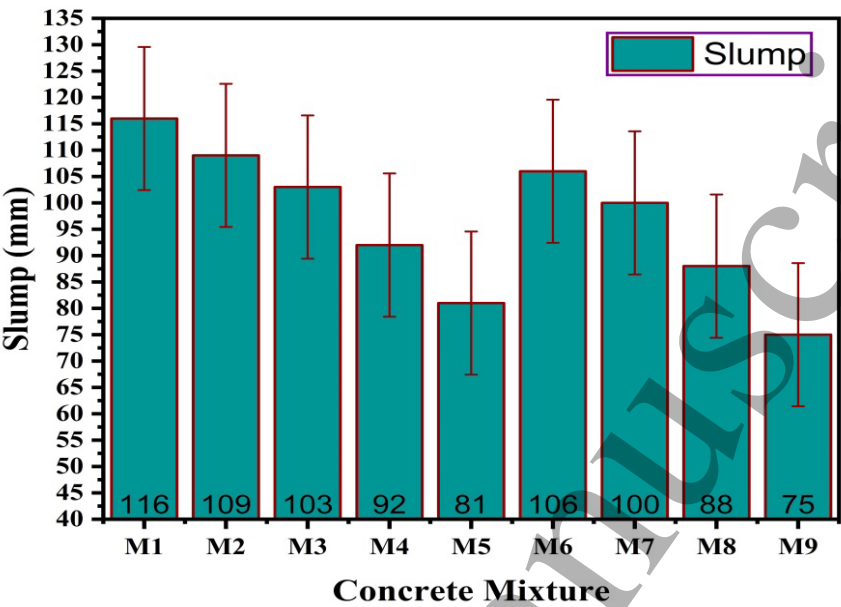


Figure 5 Slump Value for Different Concrete Mixtures

5.2 Compressive Strength

The variations of the 7 and 28 days compressive strength for different replacement percentages of surface-modified RCA concerning the reference mixture are shown in Figure 6. From the compressive strength test results as per Table 5, the concrete mixture with a higher percentage replacement of surface-modified RCA has lower compressive strength than the reference mixture. An optimal replacement percentage is observed for each surface modification technique. Abrasion treatment achieves optimal efficiency at 50% replacement, while cement slurry treatment reaches its peak effectiveness at 25%. The application of simple abrasion proves to be more efficient in enhancing compressive strength. Particularly noteworthy is the highest compressive strength achieved at 28 days, resulting from the addition of surface-modified RCA through the abrasion treatment process at a 50% replacement rate. This enhancement is attributed to the effective removal of adhered mortar and the reduction of voids, leading to increased particle density. Thus, RCAAT 50 (M3) exhibits the highest strength. However, the efficiency of surface modification through pre-soaking recycled aggregates in cement slurry is found to be comparatively lower than that of abrasion treatment.

Table 5 Percentage Variation of Compressive Strength for Different Mixtures

Concrete Mix	Compressive Strength (MPa)			
	7 Days	% Variation to RC	28 Days	% Variation to RC
M1	23.05	-	34.78	-
M2	29.96	+29.97	41.88	+20.41
M3	26.39	+14.49	38.73	+11.35
M4	21.42	-07.07	33.21	-04.51

M5	20.31	-11.88	30.99	-10.89
M6	26.27	+13.96	37.81	+08.71
M7	22.15	-03.90	33.64	-03.27
M8	20.10	-12.79	31.33	-09.91
M9	18.96	-17.78	29.63	-14.80
+ Sign Represents an Increase in Strength and – Sign Represents a Decrease in Strength				

Among the concrete mixtures examined, M3 (RCAAT25) demonstrates the highest compressive strength at 41.88 MPa, while M9 (RCACST100) exhibits the lowest at 26.29 MPa. These findings concur with prior research, such as the observations made by Kessal et al. [50]. Another study by Ashraf M. Wagih et al. demonstrated a 20-34% and 18-28% decrease in compressive strength for concrete mixtures containing RCA at 7 and 28 days, respectively [51]. Importantly, after undergoing surface modification treatment, the decline is mitigated to 12-18% at 7 days and 10-15% at 28 days, even for 100% replacement of RCA. This favorable outcome underscores the efficacy of surface modification treatments in attenuating the loss of compressive strength in recycled aggregates concrete. Notably, the abrasion treatment of RCA emerges as a pragmatic and effective approach, delivering high-quality aggregates with reduced water absorption and efficient removal of adhered mortar.

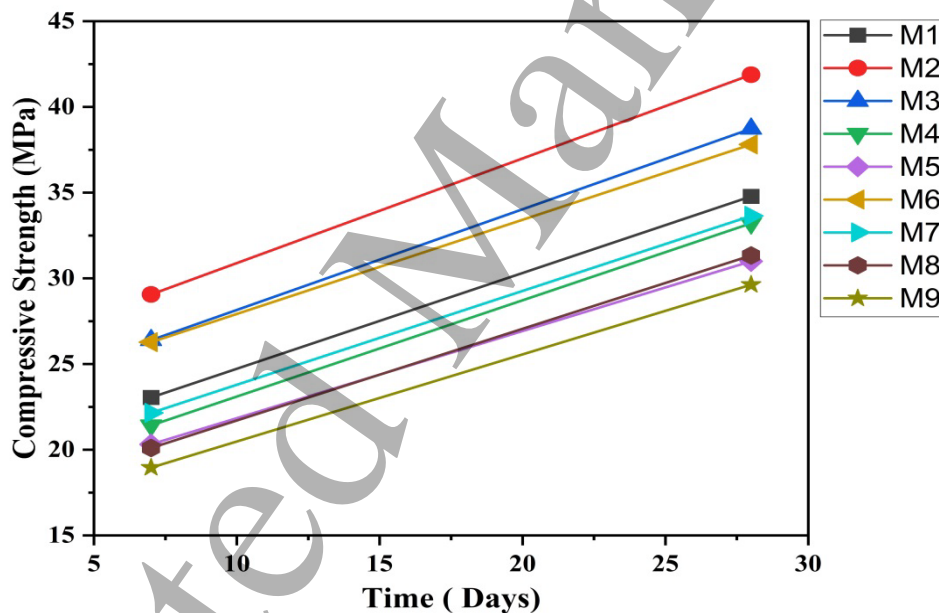


Figure 6 Variations of Compressive Strength for Different Concrete Mixtures

### 5.3 Flexural Strength

Figure 7 illustrates the variations in flexural strength at 7 and 28 days for different replacement percentages of surface-modified RCA in comparison to the reference mixture. These results closely follow the trend observed in Table 6, echoing the patterns evident in compressive strength. Notably, the data indicates that concrete mixtures incorporating a higher replacement percentage of surface-modified RCA exhibit reduced flexural strength compared to the reference mixture. Table 6 quantifies the percentage variation in flexural strength for the various concrete mixtures. It presents the flexural strength values at 7 and 28 days; along with their respective percentage variations from the reference concrete (RC). The table highlights that the use of a greater proportion of surface-

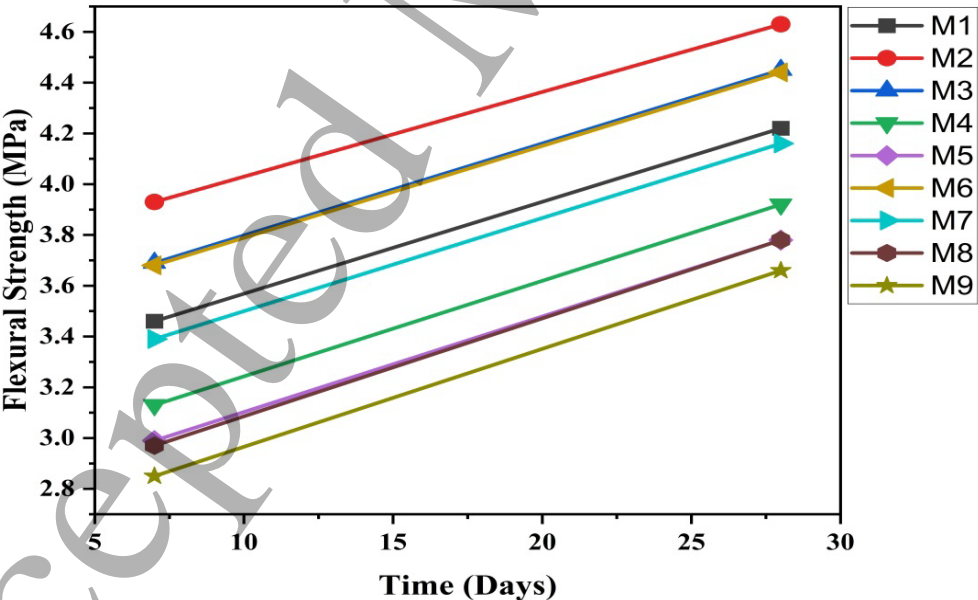
1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

modified RCA for replacement leads to a decrease in flexural strength relative to the reference mixture.

**Table 6** Percentage Variation of Flexural Strength for Different Mixtures

Concrete Mix	Flexural Strength (MPa)			
	7 Days	% Variation to RC	28 Days	% Variation to RC
M1	3.46	-	4.22	-
M2	3.93	+13.58	4.63	+09.71
M3	3.69	+06.64	4.45	+05.45
M4	3.13	-09.53	3.92	-07.10
M5	2.99	-13.58	3.78	-10.42
M6	3.68	+06.35	4.44	+05.21
M7	3.39	-02.02	4.16	-01.42
M8	2.97	-14.16	3.78	-10.42
M9	2.85	-17.63	3.66	-13.27
+ Sign Represents an Increase in Strength and – Sign Represents a Decrease in Strength				

Among the mixtures tested, M3 (RCAAT25) demonstrates the highest flexural strength at 4.45 MPa, while M9 (RCACST100) shows the lowest at 3.66 MPa. This outcome aligns with prior research, which has reported a 5-10% reduction in flexural strength for 50% RCA replacement and a 15-20% reduction for 100% replacement [52]. However, this study deviates from the observed patterns by indicating a more moderate 13% reduction at a 100% surface-modified RCA replacement level. This suggests that the surface modification treatments applied to RCA have yielded positive outcomes in terms of flexural strength. These findings contribute to a nuanced understanding of the impact of surface-modified RCA content on the flexural behaviour of concrete mixtures and highlight the potential benefits of surface modification techniques.



**Figure 7** Variations of Flexural Strength for Different Concrete Mixtures

#### 5.4 Split Tensile Strength

Figure 8 illustrates the variation in split tensile strength at 7 and 28 days for different replacement percentages of surface-modified RCA in comparison to the reference mixture. This trend aligns with the observed results in Table 7, mirroring the behaviour seen in compressive strength. Notably, a consistent pattern emerges where the split tensile strength of the concrete mixture is notably lower when a higher proportion of surface-modified RCA is introduced as a replacement for conventional aggregates. Table 7 quantifies the percentage variation in split tensile strength for the various concrete mixtures. It reveals that the use of surface-modified RCA at higher replacement levels results in a decrease in split tensile strength compared to the reference mixture. The table showcases the split tensile strength values at 7 and 28 days; along with their respective percentage variations from the reference concrete (RC).

**Table 7** Percentage Variation of Split Tensile Strength for Different Mixtures

Concrete Mix	Split Tensile Strength (MPa)			
	7 Days	% Variation to RC	28 Days	% Variation to RC
M1	2.31	-	2.82	-
M2	2.62	+13.41	3.11	+10.28
M3	2.47	+06.92	2.97	+05.31
M4	2.12	-08.22	2.62	-07.63
M5	1.99	-13.91	2.54	-09.92
M6	2.46	+06.49	2.94	+04.25
M7	2.27	-01.73	2.78	-01.41
M8	2.09	-09.52	2.59	-08.15
M9	1.91	-17.31	2.46	-12.76
+ Sign Represents an Increase in Strength and – Sign Represents a Decrease in Strength				

Among the mixtures tested, M3 (RCAAT25) displays the highest split tensile strength at 2.97 MPa, while M9 (RCACST100) exhibits the lowest at 2.46 MPa. These findings are in agreement with prior research presented by Kessal et al. [50]. Notably, this study diverges from the observations made by Ashraf M. Wagih, who reported a 24% reduction in split tensile strength [51]. In contrast, the surface-modified RCA examined in this study showcases a more modest 12% reduction at a 100% replacement level. These results underscore the influence of surface-modified RCA content on the split tensile strength of the resulting concrete mixtures and contribute to the broader understanding of the mechanical behaviour of such compositions.

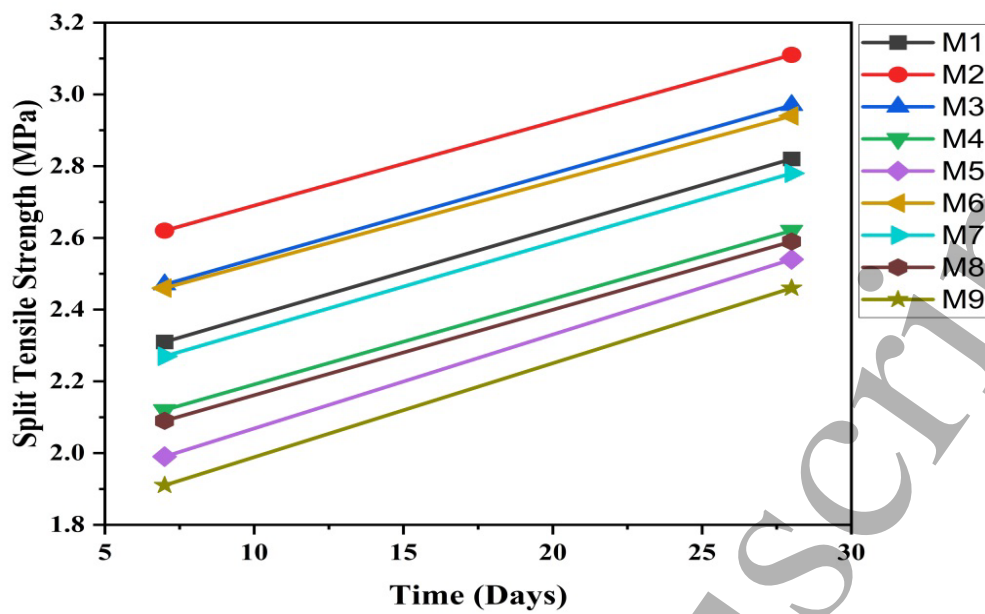


Figure 8 Variations of Split Tensile Strength for Different Concrete Mixtures

5.5 Modulus of Elasticity

The modulus of elasticity (MOE) serves as a critical indicator for assessing the deformation capacity of both Recycled Aggregate Concrete (RAC) and standard concrete. The MOE measurements conducted at the 28-day mark reveal a notable trend: an inverse relationship between the MOE and the incorporation of surface-modified RCA in the concrete matrix. This decrease in MOE is prominently presented in Table 8. The observed reduction is attributed to the inherent characteristics of surface-modified RCA, such as its water absorption tendency and brittleness. This decline in MOE is particularly noteworthy due to the heightened susceptibility of RCA to deformation. Consequently, structural elements constructed with RCA exhibit larger actual deformations compared to those composed of natural aggregates. The empirical data presented in Table 8 showcases the MOE values for various concrete mixtures.

Table 8 Modulus of Elasticity of Concrete with Different Codes

Mix	Compressive Strength (MPa)	Modulus of Elasticity (GPa)		
		Experimental ( $E_c$ )	As Per IS 456-2000 ( $E_c$ )	As Per ACI Code ( $E_c$ )
M1	34.78	32.21	29.48	27.91
M2	41.88	35.48	32.35	30.63
M3	38.73	33.98	31.11	29.46
M4	33.21	31.82	28.81	27.28
M5	27.99	29.34	26.45	25.04
M6	37.81	32.77	30.74	29.10
M7	33.64	31.87	29.00	27.45
M8	27.70	28.15	26.31	24.91
M9	26.29	27.91	25.83	24.27

Figure 9 further elucidates the connection between experimental results and different codes, illustrating how while surface modification of RCA contributes to increased strength through enhanced nucleation sites for hydration, the MOE consistently diminishes with higher levels of surface-modified RCA content. This trend is underscored by a comparative analysis of the



compressive strength and MOE of plain concrete across different levels of surface-modified RCA replacement. The factors contributing to the MOE reduction include the amount and rigidity of the binder phase, the volume and stiffness of aggregates, and the characteristics of the interfacial transition zone (ITZ) between aggregates and the cementations paste. Similar patterns of MOE decline have been observed in prior studies [50][51]. The observed decrease in MOE serves as a crucial consideration when evaluating the implications of surface-modified RCA incorporation on the overall mechanical behavior of the resulting concrete.

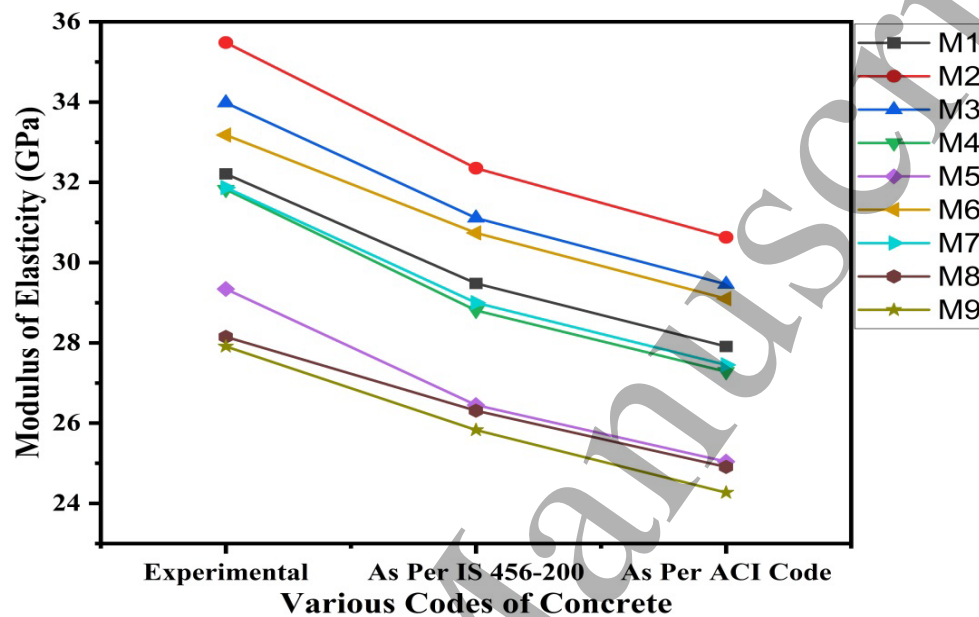
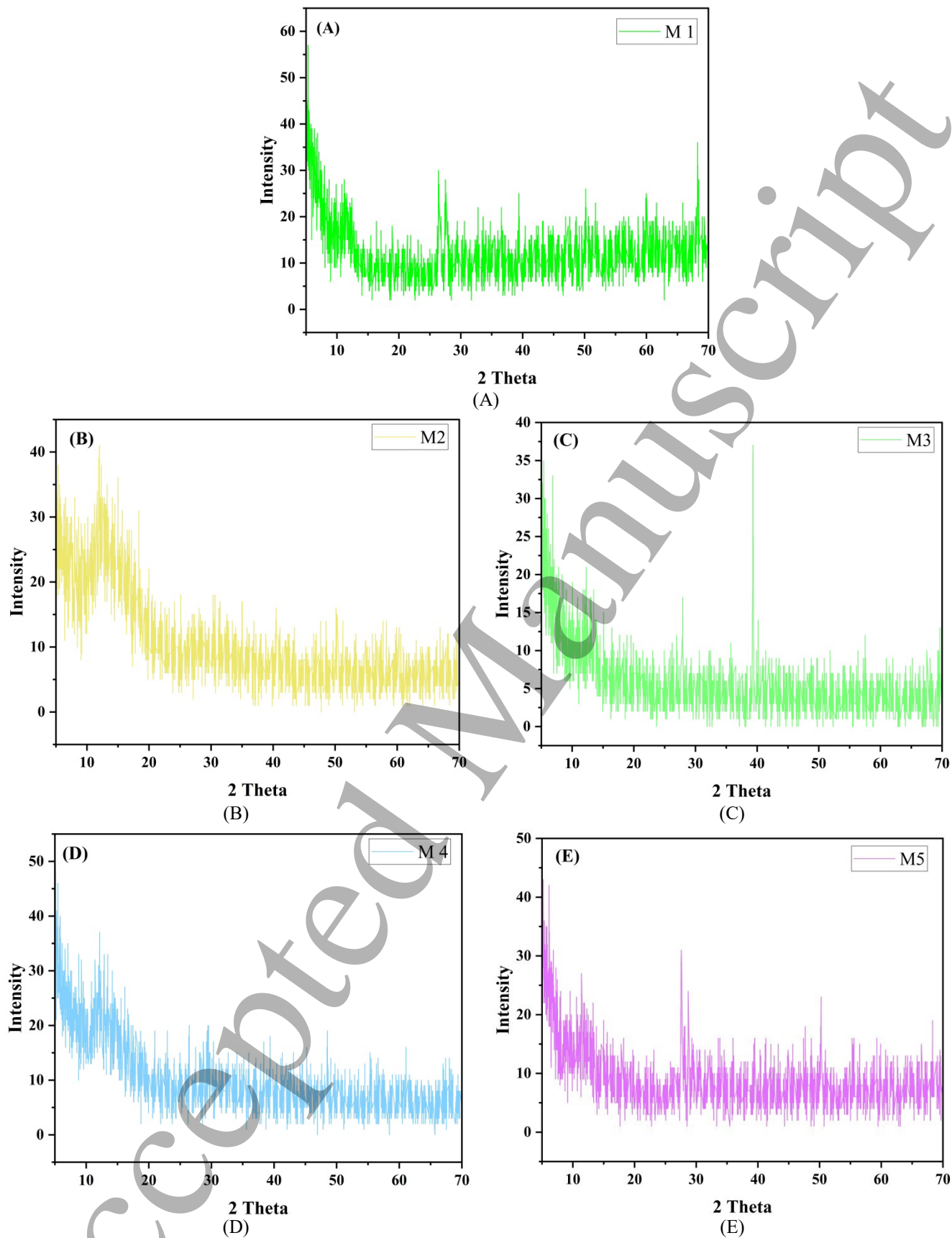


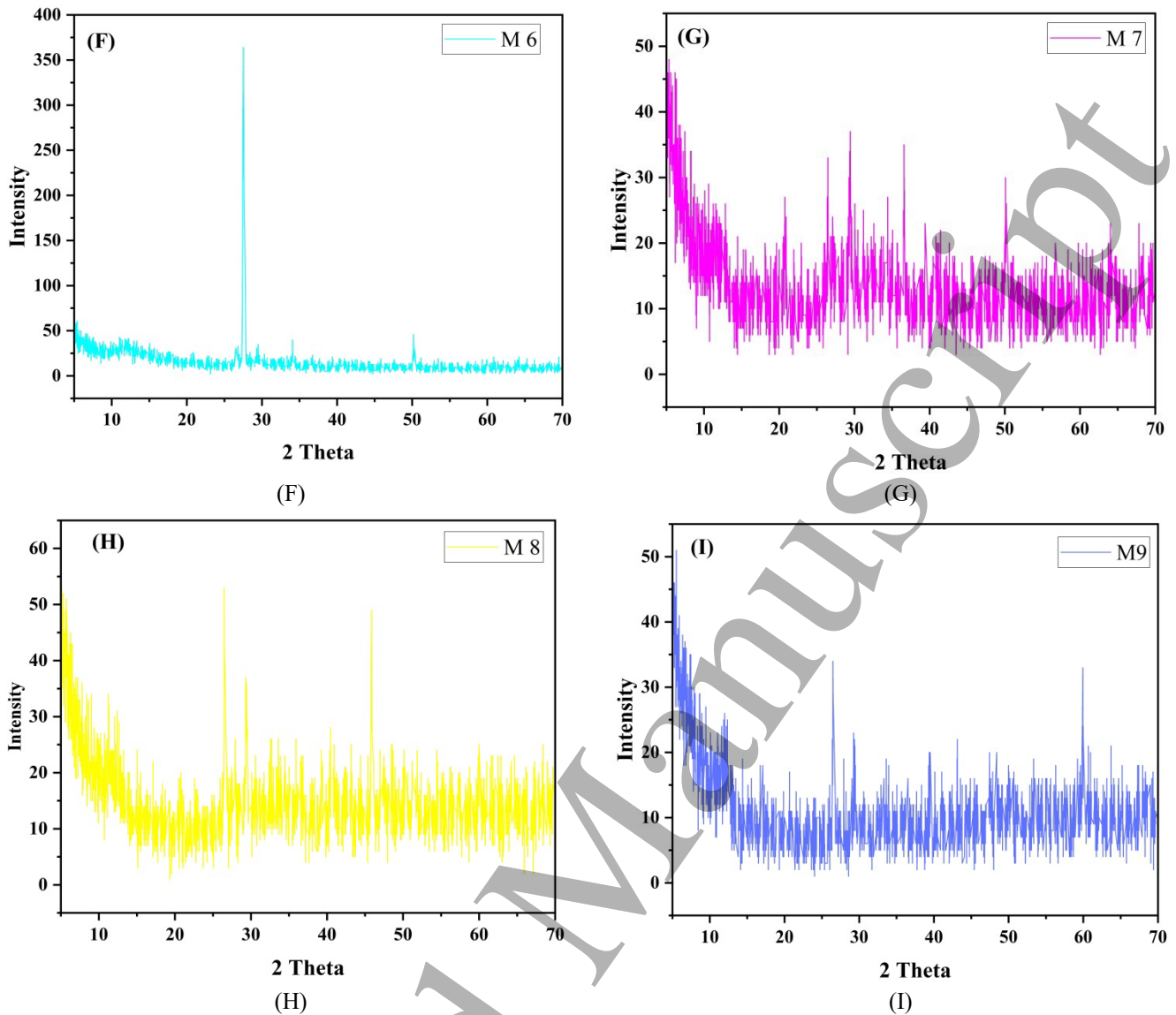
Figure 9 Variation of modulus of elasticity with different codes

## 5.6 X-Ray Diffraction (XRD)

X-Ray Diffraction (XRD) is a method for analyzing the crystallographic arrangement of materials. By directing X-rays at a crystalline sample, the X-rays diffract based on the crystal lattice arrangement. Measurement of the angles and intensities of these diffracted X-rays provides valuable insights into the material's crystal structure. The Bruker D-8 diffractometer scans samples at a 2-degree angle from 3 to 70 degrees, with a scan speed of 2 degrees per minute and a 0.005-degree sampling interval. The Jade 7 X-ray diffraction software analyzes the scans, presenting peak intensities in a graph against 2 degrees on the x-axis and intensity on the y-axis (Figure 10). This equipment comprises a high-intensity X-ray source, a goniometer for sample rotation, and a detector for capturing diffracted X-rays. Samples for XRD analysis were ground into a fine powder, carefully loaded onto a sample holder, or mounted to align correctly with the X-ray beam. The XRD instrument was calibrated to specified measurement conditions, including X-ray wavelength and scan range. Placing the sample holder within the instrument, the X-ray beam was directed onto the sample. Diffracted X-rays were collected across various angles, and the resulting diffraction pattern was scrutinized to deduce the material's crystal structure and identify its constituent phases.







**Figure 10** XRD for Different Concrete Mixtures (A) for M1, (B) for M2, (C) for M3, (D) for M4, (E) for M5, (F) for M6, (G) for M7, (H) for M8, and (I) for M9

The analysis of X-ray diffraction (XRD) results elucidates the presence of well-defined crystalline formations characterized by consistent geometrical patterns, influenced by the incorporation of surface-modified recycled coarse aggregates (RCA). This influence extends to the phase composition of minerals, including calcium silicate hydrate (CSH), calcium alumina silicate hydrate (CASH), ettringite, and calcium hydroxide (CH). The XRD peaks' characteristics, including their positions, corresponding d-spacing values, chemical formulas, chemical names, and crystal system classifications, are comprehensively detailed in Table 9, providing a comprehensive understanding of the structural changes and mineral transformations induced by the utilization of surface-modified RCA. This study is remarkably similar to the previous research [52].

**Table 9** XRD data for various concrete mixtures

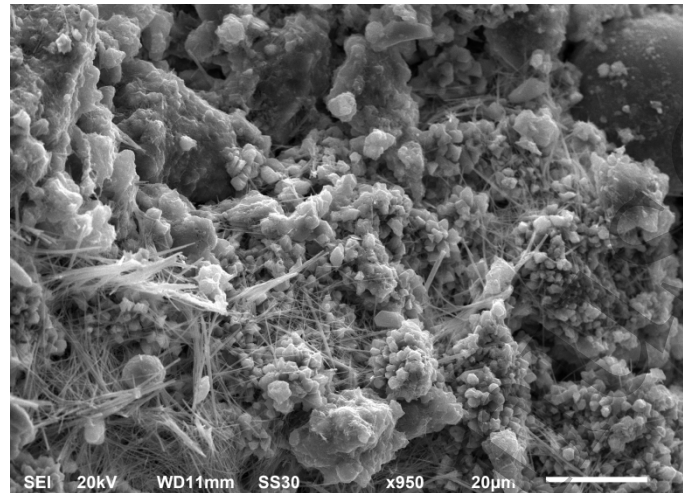
S. No.	Material	Peak No.	Peak Angle	d-spacing	Chemical Formula	Crystal System
1	M1	1	26.652	3.34510	CSH	Tobermorite
		2	27.020	3.29730	CASH	Tobermorite

		3	34.569	2.59178	ETTRINGITE	Hexagonal
		4	61.577	1.50480	CH	Hexagonal
2	M2	1	25.722	3.39863	CSH	Tobermorite
		2	27.201	3.27137	CASH	Tobermorite
		3	34.980	2.48364	ETTRINGITE	Hexagonal
		4	60.600	1.49825	CH	Hexagonal
3	M3	1	26.915	3.30932	CSH	Tobermorite
		2	27.340	3.25943	CASH	Tobermorite
		3	39.324	2.28957	ETTRINGITE	Hexagonal
		4	59.861	1.54387	CH	Hexagonal
4	M4	1	26.621	3.34593	CSH	Tobermorite
		2	27.140	3.28299	CASH	Tobermorite
		3	42.684	2.11678	ETTRINGITE	Hexagonal
		4	54.626	1.67893	CH	Hexagonal
5	M5	1	25.061	3.55057	CSH	Tobermorite
		2	27.300	3.26411	CASH	Tobermorite
		3	35.980	2.49408	ETTRINGITE	Hexagonal
		4	60.693	1.52450	CH	Hexagonal
6	M6	1	26.637	3.34831	CSH	Tobermorite
		2	27.521	3.23846	CASH	Tobermorite
		3	34.146	2.62371	ETTRINGITE	Hexagonal
		4	50.175	1.81673	CH	Hexagonal
7	M7	1	26.596	3.34884	CSH	Tobermorite
		2	29.440	3.03153	CASH	Tobermorite
		3	36.606	2.45287	ETTRINGITE	Hexagonal
		4	50.110	1.81894	CH	Hexagonal
8	M8	1	26.575	3.35146	CSH	Tobermorite
		2	29.402	3.03534	CASH	Tobermorite
		3	39.385	2.28593	ETTRINGITE	Hexagonal
		4	60.001	1.54058	CH	Hexagonal
9	M9	1	26.571	3.35196	CSH	Tobermorite
		2	29.281	3.04761	CASH	Tobermorite
		3	39.523	2.27830	ETTRINGITE	Hexagonal
		4	59.930	1.54223	CH	Hexagonal
CSH- Calcium Silicate Hydroxide, CASH- Calcium Aluminate Silicate Hydrates, CH- Calcium Hydroxide, Ettringite- Hydrated Calcium Aluminum Sulfate Hydroxide						

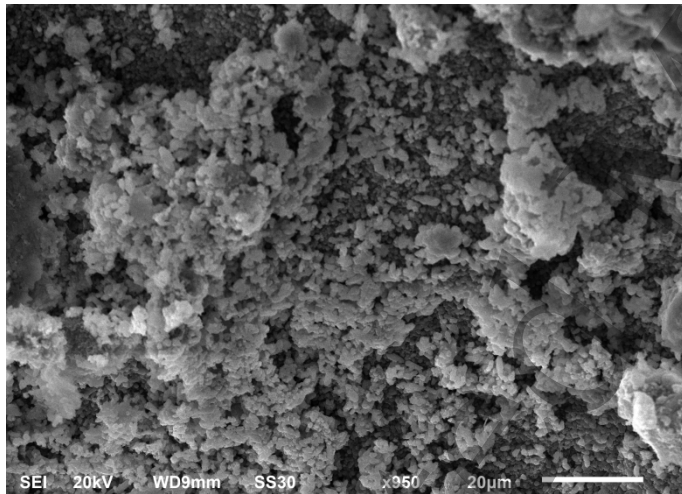
## 5.7 Scanning Electron Microscopy (SEM)

Scanning Electron Microscopy (SEM) is a potent imaging method utilizing focused electron beams to capture high-resolution surface images of materials, with magnifications ranging from modest to extensive. SEM is applied to characterize RCA materials, examining particle microstructure and surface morphology in concrete samples. The JSM 6610V SEM at the University Science Instrumentation Center (USIC) Delhi was used for sample analysis. This SEM boasts a high-resolution electron gun, electromagnetic lenses, and detectors for secondary and backscattered electrons. Sample preparation involves meticulous steps like cutting, polishing, and conductive coating to ensure accurate imaging and sample integrity. Prepared samples are placed in the SEM stage within a vacuum chamber, where a focused electron beam produces high-resolution images by

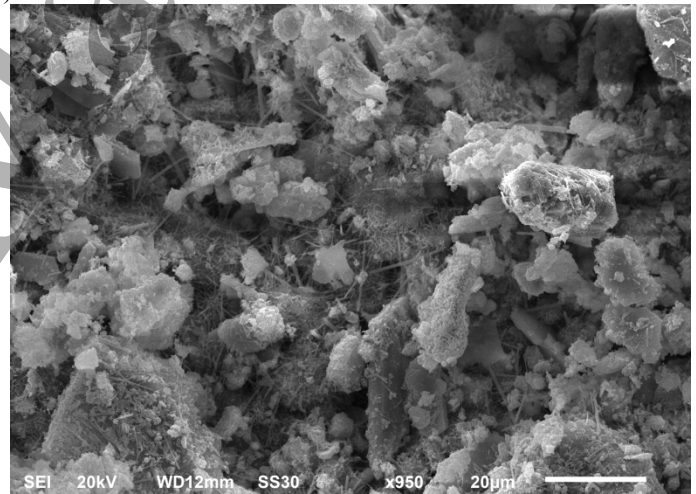
detecting emitted secondary and backscattered electrons. Figure 11 showcases micrographs from SEM analysis of distinct mixtures at the 28-day mark. The formation of hydration products at the microstructure level in different concrete mixtures at 28 days is seen in the micrographs, which are responsible for the strength of concrete. The main compounds present during the hydration process are calcium hydroxide (CH), calcium silicate hydroxide (CSH), and ettringite. The hexagonal crystals indicate CH, the flower-shaped structure indicates CSH gel and the needle-like structure indicates ettringite.



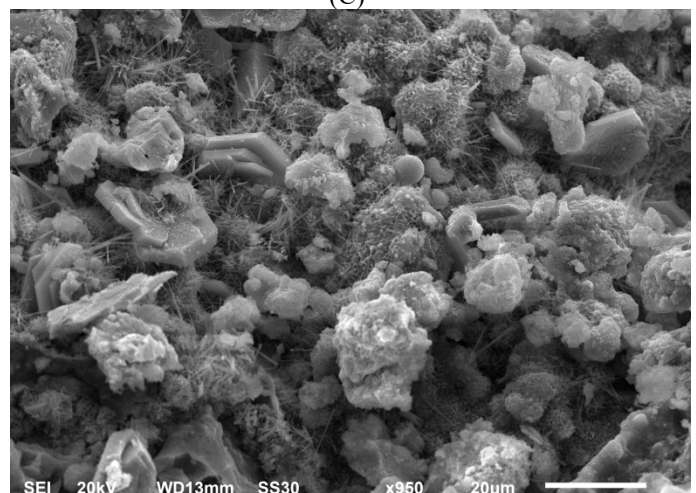
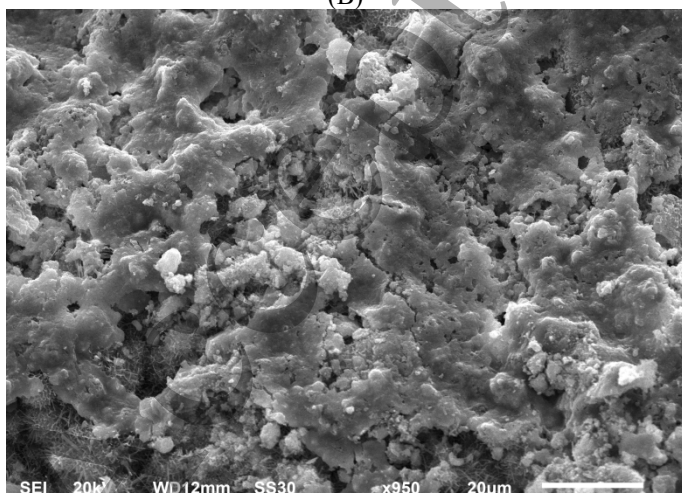
(A)



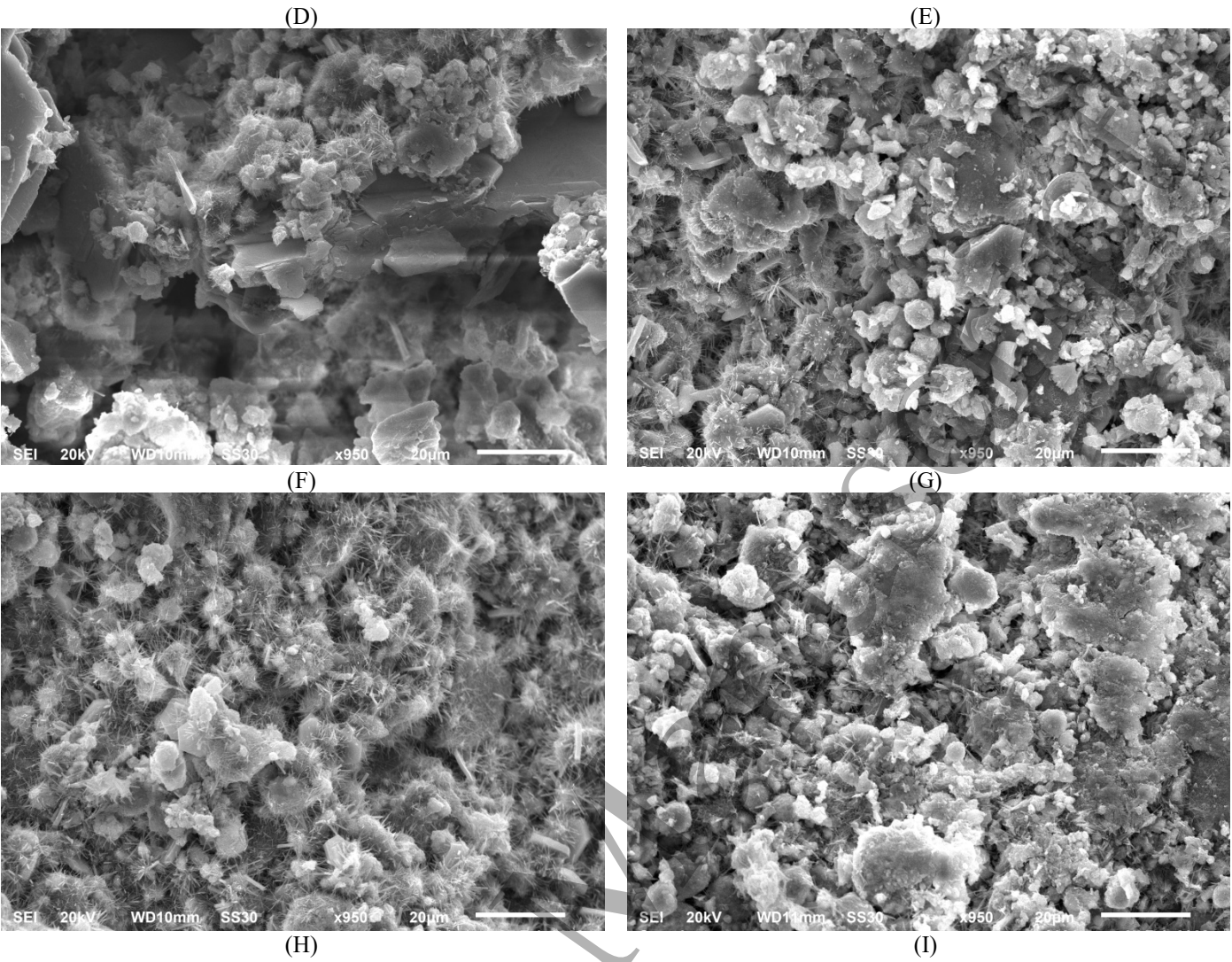
(B)



(C)





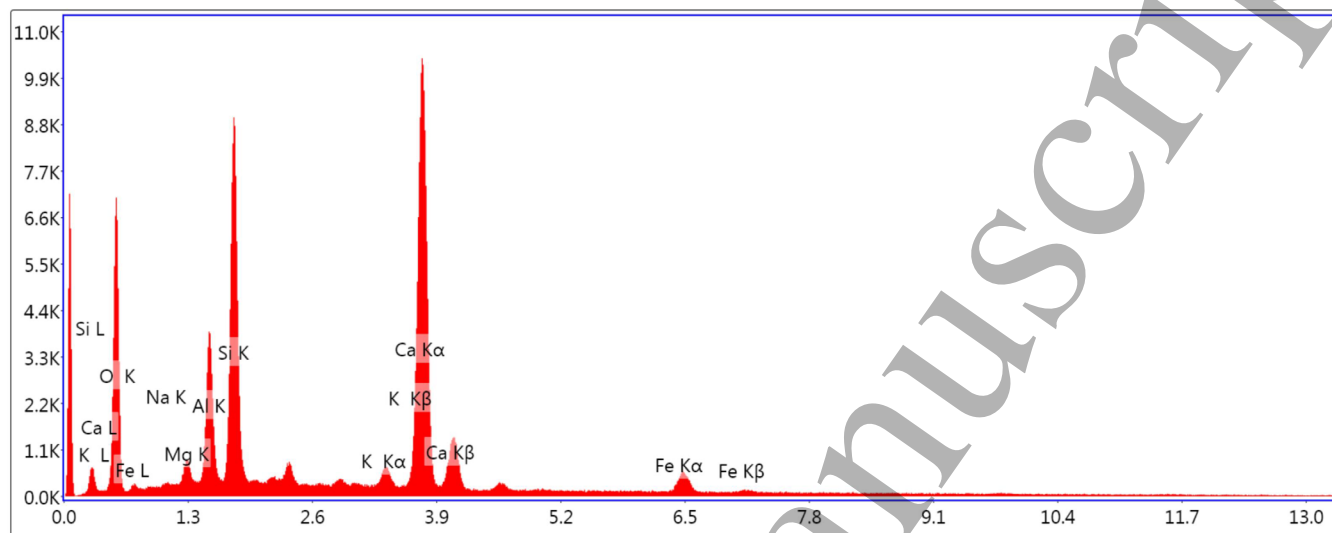


**Figure 11** SEM Micrographs for Different Concrete Mixtures (A) for M1, (B) for M2, (C) for M3, (D) for M4, (E) for M5, (F) for M6, (G) for M7, (H) for M8, and (I) for M9

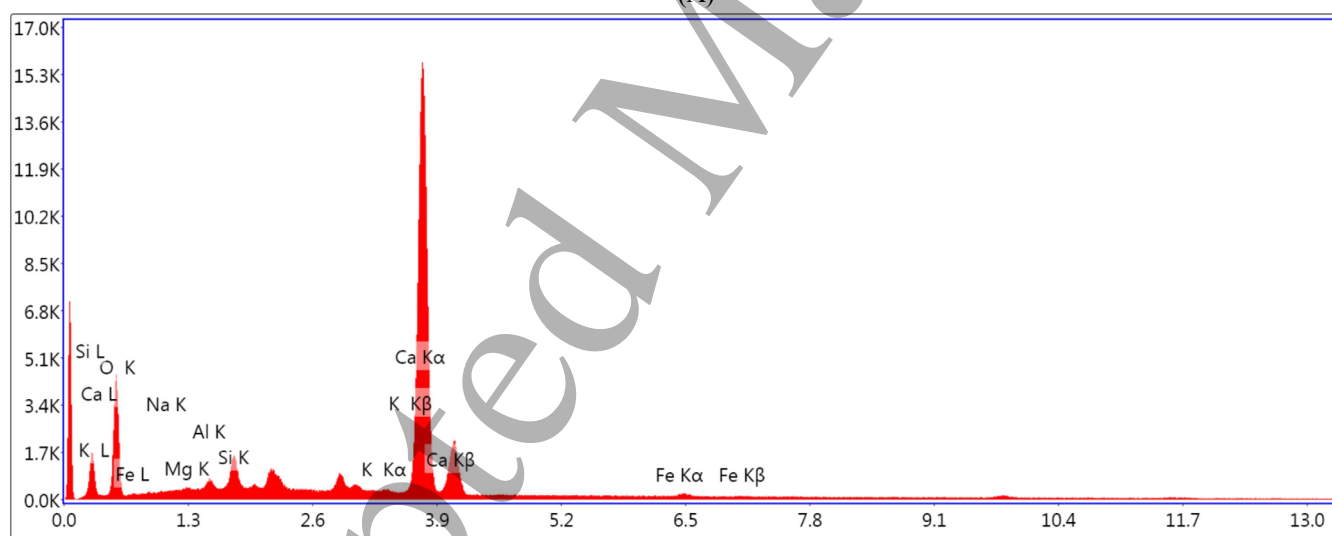
The acquired test outcomes elucidate that the integration of surface-modified recycled coarse aggregates (RCA) through abrasion treatment yields substantial enhancements in the concrete's microstructure. This alteration results in a more compact cement paste, thereby facilitating improved adhesion between the aggregates and cement paste. Conversely, the introduction of surface-modified RCA via cement slurry treatment diminishes the porosity of RCA and amplifies the density and robustness of the interfacial transition zones (ITZs). Complementary scanning electron microscope (SEM) analysis corroborates these findings by confirming that the removal of adhered mortar from RCA through abrasion treatment, coupled with the application of a cement slurry coating, synergistically contributes to the refinement of RCA's microstructural attributes. Consequently, these modifications exert a positive influence on the caliber and potency of recycled aggregates concrete, highlighting the significance of surface modification techniques in augmenting the performance of environmentally sustainable concrete materials. Comparable interpretations are also done in the previous study [52].

## 5.8 Energy Dispersive X-Ray Spectroscopy (EDAX)

Energy-dispersive X-ray spectroscopy (EDAX) is employed for the comprehensive chemical composition analysis of concrete samples. This research utilizes EDAX to ascertain both the qualitative and quantitative evaluations of distinct elements within different concrete blends. Figure 12 depicts the quantitative and qualitative analyses of diverse elements present in various mixtures, as determined through EDAX analysis.

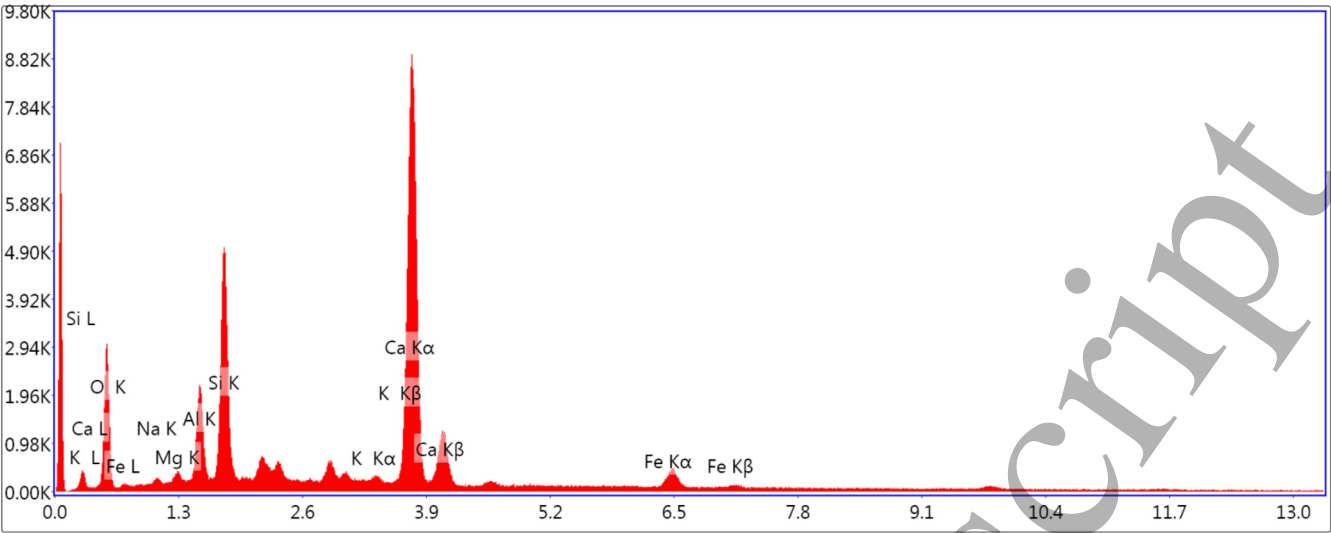


(A)



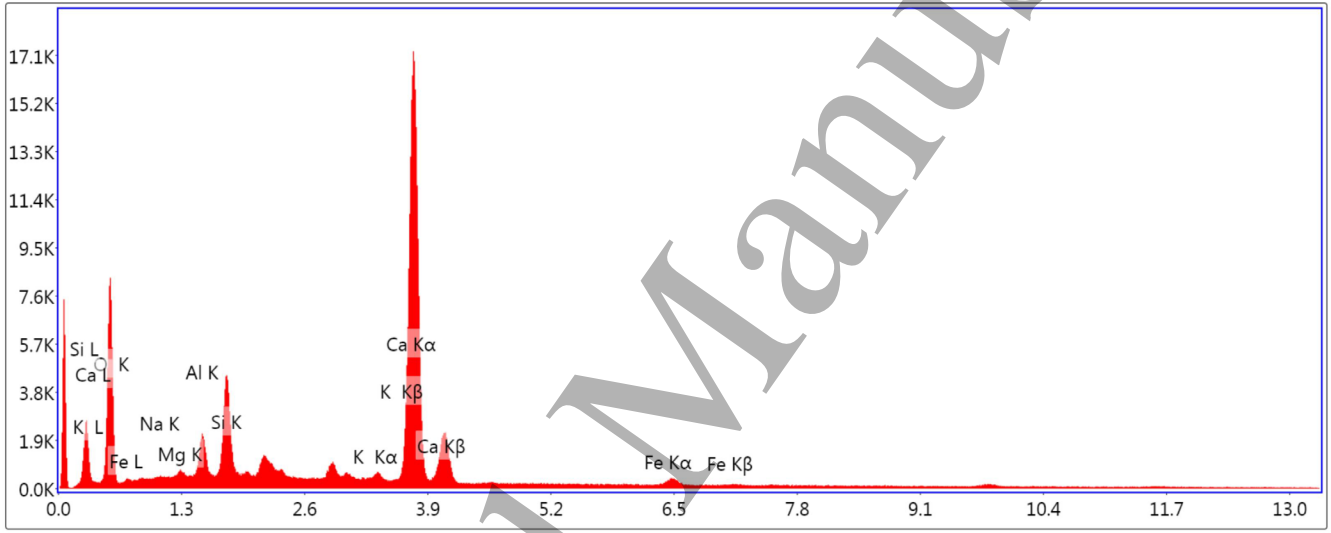
(B)





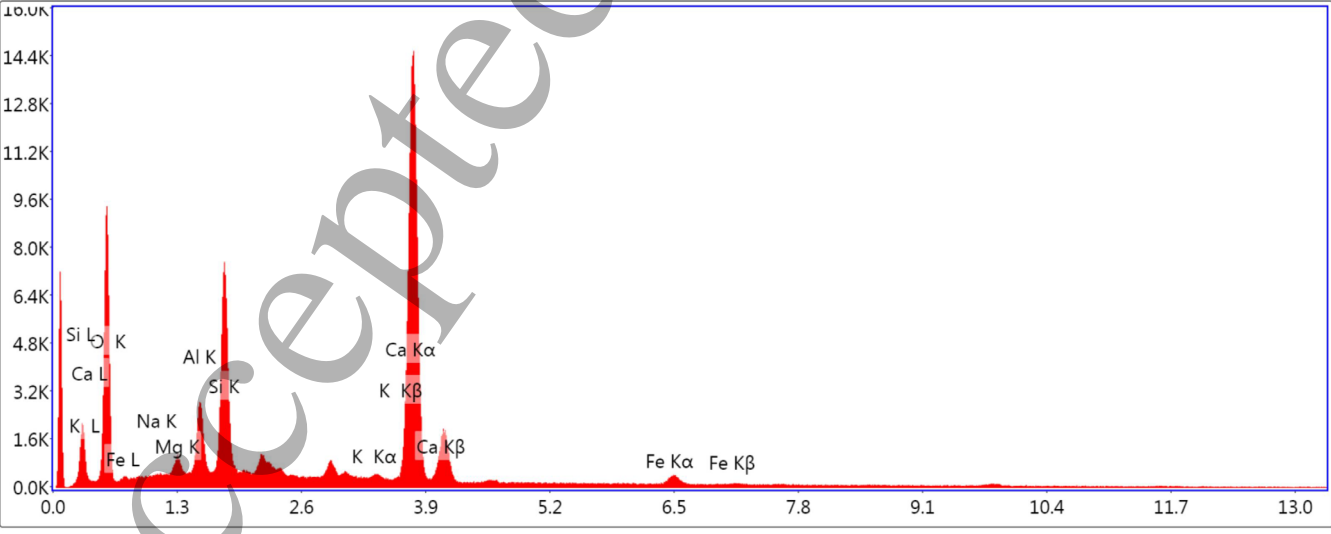
Lsec: 200.0 0 Cnts 0.000 keV Det: Octane Plus Det

(C)



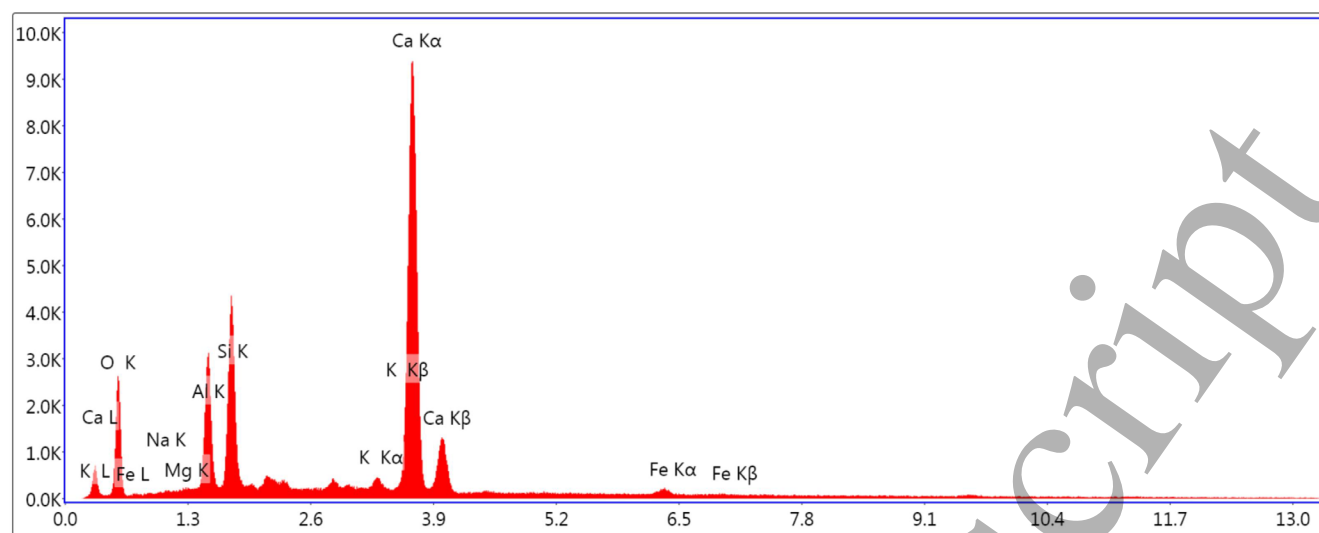
Lsec: 200.0 0 Cnts 0.000 keV Det: Octane Plus Det

(D)



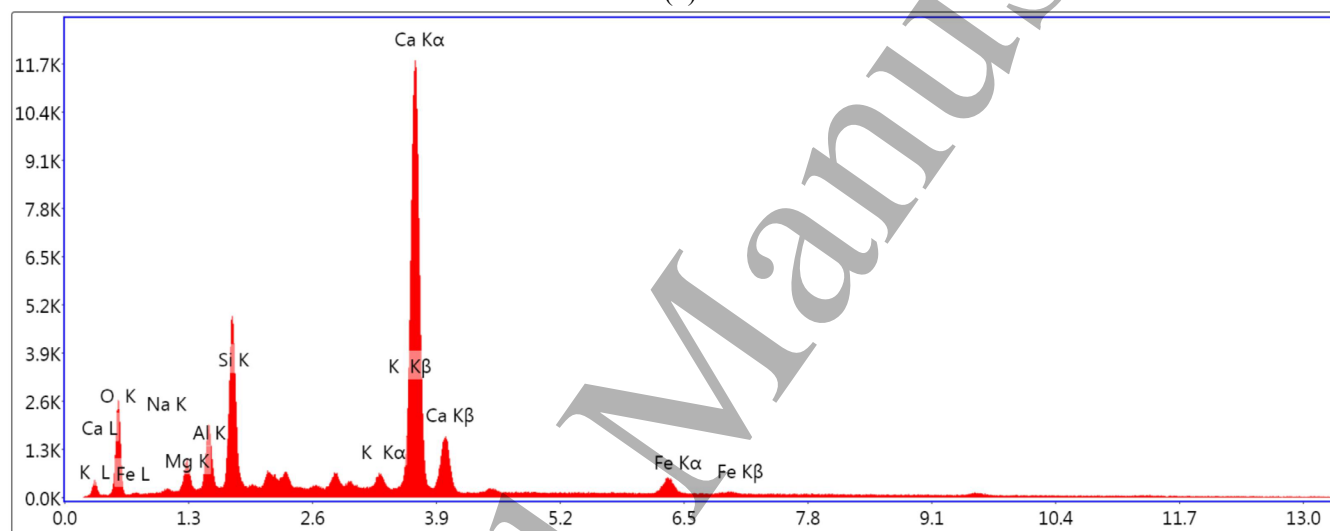
Lsec: 200.0 0 Cnts 0.000 keV Det: Octane Plus Det

(E)



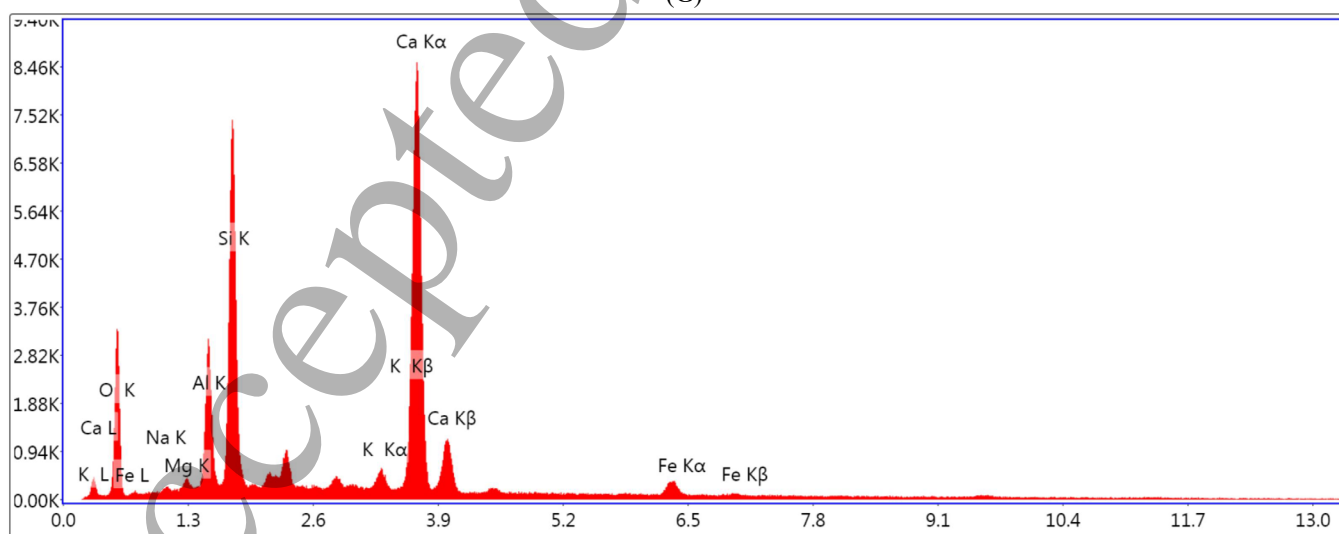
Lsec: 200.0 0 Cnts 0.000 keV Det: Octane Plus Det

(F)



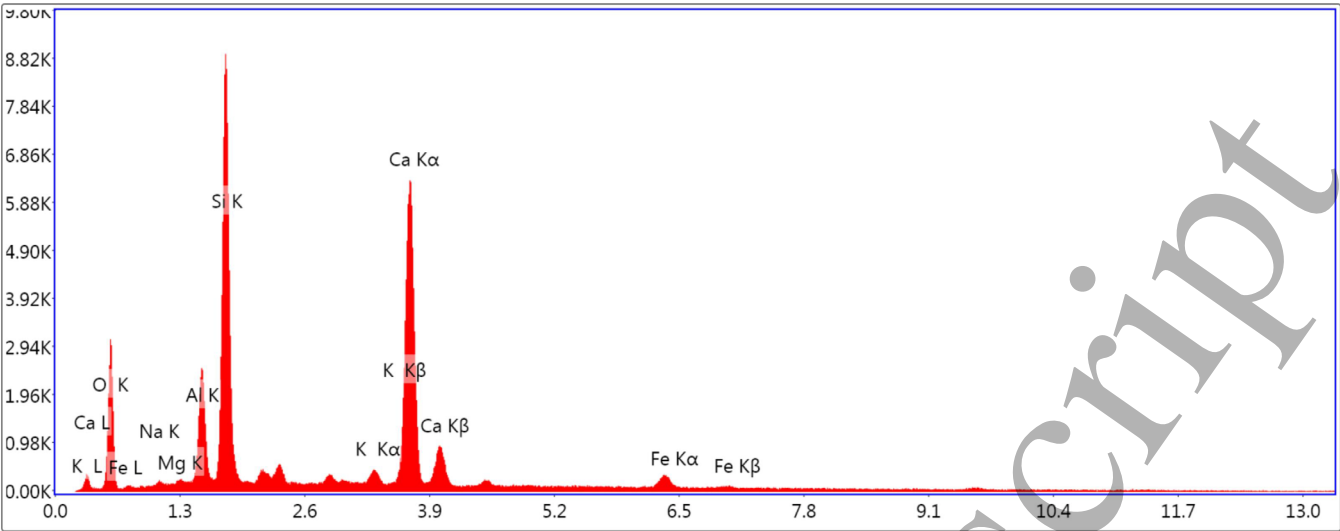
Lsec: 200.0 0 Cnts 0.000 keV Det: Octane Plus Det

(G)



Lsec: 200.0 0 Cnts 0.000 keV Det: Octane Plus Det

(H)



(I)

**Figure 12** EDAX Analysis for Different Concrete Mixtures (A) for M1, (B) for M2, (C) for M3, (D) for M4, (E) for M5, (F) for M6, (G) for M7, (H) for M8, and (I) for M9

The results of the Energy Dispersive X-ray Analysis (EDAX) reveal the presence of several elemental components within the tested samples, including Calcium (Ca), Oxygen (O), Silicon (Si), Aluminum (Al), Iron (Fe), Potassium (K), Sodium (Na), and Magnesium (Mg). These elements constitute the chemical composition of the examined materials. The quantitative proportions of these elements, as determined through EDAX analysis, are succinctly presented and organized in Table 10, providing valuable insights into the elemental makeup of the studied samples. The results are consistent with the previous study [52].

**Table 10** Percentage Weight of Various Elements for Different Mixtures

Mixture ID	(%)	Ca-K	O-K	Si-K	Al-K	Fe-K	K-K	Na-K	Mg-K
M1	Weight	28	48.21	10.49	5.27	4.12	1.43	1.04	1.46
	Atomic	15.54	67.03	8.3	4.34	1.64	0.81	1.01	1.33
M2	Weight	46.76	48.8	1.6	0.58	0.64	0.42	0.92	0.29
	Atomic	26.71	69.81	1.31	0.49	0.26	0.24	0.91	0.27
M3	Weight	38.7	41.3	9.22	4.26	4.59	0.87	0.64	0.43
	Atomic	23.08	61.71	7.84	3.77	1.97	0.53	0.67	0.42
M4	Weight	36.12	52.23	4.08	2.25	1.57	0.92	1.88	0.94
	Atomic	19.74	71.48	3.18	1.82	0.62	0.52	1.79	0.85
M5	Weight	30.81	52.78	6.99	3.25	1.97	0.75	1.82	1.64
	Atomic	16.57	71.13	5.36	2.6	0.76	0.41	1.71	1.45
M6	Weight	64.84	21.29	3.01	1.28	9.15	0.32	0.09	0.02
	Atomic	49.33	40.57	3.27	1.45	5.0	0.25	0.12	0.02
M7	Weight	45.84	33.15	8.68	3.44	5.44	1.46	0.12	1.88
	Atomic	29.56	53.55	7.99	3.29	2.52	0.96	0.14	2.0
M8	Weight	35.4	36.93	14.33	5.92	5.07	1.72	0.26	0.36
	Atomic	21.64	56.55	12.5	5.37	2.22	1.08	0.28	0.36
M9	Weight	30.92	36.84	19.62	5.61	5.11	1.44	0.33	0.13
	Atomic	18.69	55.78	16.92	5.03	2.22	0.89	0.34	0.13
Ca-Calcium, O-Oxygen, Si-Silicon, Al-Aluminum, Fe-Iron, K-Potassium, Na-Sodium, and Mg-Magnesium									

## 6. CONCLUSIONS

This paper contributes to the development of sustainable concrete by utilizing surface-modified recycled coarse aggregates (RCA) from construction and demolition waste. This approach promotes eco-friendly practices, waste reduction, and a circular economy. Substituting natural coarse aggregates with treated RCA significantly impacts hardened concrete properties. Conclusions from the study include:

- Abrasion and cement slurry treatment enhance interfacial connection between RCA and cement paste, with abrasion treatment proving more effective for removing attached mortar and enhancing recycled aggregates.
- After abrasion treatment, RCA can partially replace coarse aggregates up to 50% without major compressive strength loss. Cement slurry-treated RCA also maintains strength when used up to 50% replacement.
- Strength improvements are evident with 50% surface-modified RCA replacement, paralleling concrete with 100% natural aggregates.
- Compared to reference concrete, abrasion-treated RCA slightly reduces compressive, flexural, and split tensile strengths (10.89%, 10.42%, and 09.92%), while cement slurry-treated RCA shows greater reduction (14.80%, 13.27%, and 12.76%).
- Concrete workability matches conventional concrete with admixture. Theoretical modulus of elasticity aligns with ACI and IS code.
- Surface modification creates denser ITZ in SEM studies, suggesting potential for sustainable concrete solutions.

Using abrasion and cement slurry modification, this study counters concrete strength loss in RCA, creating a robust link to cement and allowing up to 50% RCA replacement. This approach curbs waste generation, conserves resources, and supports greener construction practices.

### Statements and Declarations:

#### Author's Credit Statement

**Harish:** Conceptualization, Methodology, Formal Analysis, Investigation, Resources, Data Curation, Writing, Original Draft Preparation, Writing Review and Editing, and Visualization.

**Awadhesh Kumar:** Conceptualization, Methodology, Validation, Resources, Review and Editing, Supervision, Project Administration, and Funding Acquisition.

#### Data Availability

All the data, materials, and methodology adopted during the research have been mentioned in this article in the form of Figures and Tables.

#### Ethics Approval

Each of the authors confirms that this manuscript has not been previously published and is not currently under consideration by any other journal.

### Consent to Participate

The authors declare consent to participate in the research work.

### Consent to Publication

The authors of the research paper agree with the publication.

### Funding

The authors declare that they did not receive any money from grants or any other sources to help them with the preparation of this publication.

### Acknowledgments:

The personnel of Delhi Technological University, Shahbad Daultpur Village, Rohini, New Delhi, Department of Civil Engineering, are gratefully acknowledged by the authors for their assistance during the research.

### References

- [1] S. Kaza, L. Yao, P. Bhada-Tata, and F. Van Woerden, "What a Waste 2.0 Introduction -"Snapshot of Solid Waste Management to 2050." Overview booklet," *Urban Dev. Ser.*, pp. 1–38, 2018, [Online]. Available: <https://openknowledge.worldbank.org/handle/10986/30317>.
- [2] E. Amasuomo and J. Baird, "The Concept of Waste and Waste Management," *J. Manag. Sustain.*, vol. 6, no. 4, p. 88, 2016, doi: 10.5539/jms.v6n4p88.
- [3] W. Z. Taffese, "Suitability Investigation of Recycled Concrete Aggregates for Concrete Production: An Experimental Case Study," *Adv. Civ. Eng.*, vol. 2018, pp. 1–11, 2018, doi: 10.1155/2018/8368351.
- [4] S. Bansal and S. K. Singh, "A Sustainable Approach towards the Construction and Demolition Waste," *Int. J. Innov. Res. Sci. Eng. Technol. (An ISO)*, vol. 3297, no. 2, pp. 2319–8753, 2007, [Online]. Available: [www.ijirset.com](http://www.ijirset.com).
- [5] J. de Brito, C. S. Poon, and B. Zhan, "Special issue: New trends in recycled aggregate concrete," *Appl. Sci.*, vol. 9, no. 11, 2019, doi: 10.3390/app9112324.
- [6] R. Sharma, "Laboratory Study on Effect of Construction Wastes and Admixtures on Compressive Strength of Concrete," *Arab. J. Sci. Eng.*, vol. 42, no. 9, pp. 3945–3962, Sep. 2017, doi: 10.1007/s13369-017-2540-0.
- [7] S. Bansal and S. K. Singh, "Sustainable Handling of Construction and Demolition (C & D) Waste," *Int. J. Sustain. Energy Environ. Res.*, vol. 4, no. 2, pp. 22–48, 2015, doi: 10.18488/journal.13/2015.4.2/13.2.22.48.
- [8] S. Bansal, S. K. Singh, and J. Kurian, "Construction and Demolition (C&D) waste recycling in New Delhi," *4th Int. fib Congr. 2014 Improv. Perform. Concr. Struct. FIB 2014 - Proc.*, no. June 2015, pp. 286–289, 2014, doi: 10.13140/RG.2.1.1022.9922.
- [9] M. S. de Juan and P. A. Gutiérrez, "Study on the influence of attached mortar content on the properties of recycled concrete aggregate," *Constr. Build. Mater.*, vol. 23, no. 2, pp. 872–877, 2009, doi: 10.1016/j.conbuildmat.2008.04.012.
- [10] M. Quattrone, S. C. Angulo, and V. M. John, "Energy and CO2 from high performance recycled aggregate production," *Resour. Conserv. Recycl.*, vol. 90, pp. 21–33, 2014, doi: 10.1016/j.resconrec.2014.06.003.
- [11] C. Shi, Y. Li, J. Zhang, W. Li, L. Chong, and Z. Xie, "Performance enhancement of recycled concrete aggregate - A review," *J. Clean. Prod.*, vol. 112, pp. 466–472, 2016, doi: 10.1016/j.jclepro.2015.08.057.
- [12] M. Sabbrojjaman and R. Mia, "A Study on Recycled Coarse Aggregate as a Full or Partial Replacement of Coarse Aggregate in Concrete Production using high FM fine aggregates Experimental Investigations on Recycled Coarse Aggregate as a Full or Partial Replacement of Coarse Aggregate in," no. February, 2019, [Online]. Available: <https://www.researchgate.net/publication/332725854>.
- [13] K. Pandurangan, A. Dayanithy, and S. Om Prakash, "Influence of treatment methods on the bond strength of

- recycled aggregate concrete,” *Constr. Build. Mater.*, vol. 120, pp. 212–221, 2016, doi: 10.1016/j.conbuildmat.2016.05.093.
- [14] S. Han, S. Zhao, and D. Lu, “Performance Improvement of Recycled Concrete Aggregates and Their Potential Applications in Infrastructure : A Review,” *Buildings*, vol. 13, no. 1411, 2023.
- [15] R. Wang, N. Yu, and Y. Li, “Methods for improving the microstructure of recycled concrete aggregate: A review,” *Constr. Build. Mater.*, vol. 242, p. 118164, 2020, doi: 10.1016/j.conbuildmat.2020.118164.
- [16] V. W. Y. Tam, X. F. Gao, and C. M. Tam, “Microstructural analysis of recycled aggregate concrete produced from two-stage mixing approach,” *Cem. Concr. Res.*, vol. 35, no. 6, pp. 1195–1203, 2005, doi: 10.1016/j.cemconres.2004.10.025.
- [17] V. W. Y. Tam, C. M. Tam, and K. N. Le, “Removal of cement mortar remains from recycled aggregate using pre-soaking approaches,” *Resour. Conserv. Recycl.*, vol. 50, no. 1, pp. 82–101, 2007, doi: 10.1016/j.resconrec.2006.05.012.
- [18] J. Li, H. Xiao, and Y. Zhou, “Influence of coating recycled aggregate surface with pozzolanic powder on properties of recycled aggregate concrete,” *Constr. Build. Mater.*, vol. 23, no. 3, pp. 1287–1291, 2009, doi: 10.1016/j.conbuildmat.2008.07.019.
- [19] D. Kong, T. Lei, J. Zheng, C. Ma, J. Jiang, and J. Jiang, “Effect and mechanism of surface-coating pozzalanics materials around aggregate on properties and ITZ microstructure of recycled aggregate concrete,” *Constr. Build. Mater.*, vol. 24, no. 5, pp. 701–708, 2010, doi: 10.1016/j.conbuildmat.2009.10.038.
- [20] S. C. Kou and C. S. Poon, “Enhancing the durability properties of concrete prepared with coarse recycled aggregate,” *Constr. Build. Mater.*, vol. 35, pp. 69–76, 2012, doi: 10.1016/j.conbuildmat.2012.02.032.
- [21] V. Spaeth and A. Djerbi Tegguer, “Improvement of recycled concrete aggregate properties by polymer treatments,” *Int. J. Sustain. Built Environ.*, vol. 2, no. 2, pp. 143–152, 2013, doi: 10.1016/j.ijse.2014.03.003.
- [22] S. Ismail and M. Ramli, “Mechanical strength and drying shrinkage properties of concrete containing treated coarse recycled concrete aggregates,” *Constr. Build. Mater.*, vol. 68, pp. 726–739, 2014, doi: 10.1016/j.conbuildmat.2014.06.058.
- [23] J. Qiu, D. Q. S. Tng, and E. H. Yang, “Surface treatment of recycled concrete aggregates through microbial carbonate precipitation,” *Constr. Build. Mater.*, vol. 57, pp. 144–150, 2014, doi: 10.1016/j.conbuildmat.2014.01.085.
- [24] J. Zhang, C. Shi, Y. Li, X. Pan, C.-S. Poon, and Z. Xie, “Performance Enhancement of Recycled Concrete Aggregates through Carbonation,” *J. Mater. Civ. Eng.*, vol. 27, no. 11, 2015, doi: 10.1061/(asce)mt.1943-5533.0001296.
- [25] L. Wang, J. Wang, X. Qian, P. Chen, Y. Xu, and J. Guo, “An environmentally friendly method to improve the quality of recycled concrete aggregates,” *Constr. Build. Mater.*, vol. 144, pp. 432–441, 2017, doi: 10.1016/j.conbuildmat.2017.03.191.
- [26] C. Shi, Z. Wu, Z. Cao, T. C. Ling, and J. Zheng, “Performance of mortar prepared with recycled concrete aggregate enhanced by CO<sub>2</sub> and pozzolan slurry,” *Cem. Concr. Compos.*, vol. 86, pp. 130–138, 2018, doi: 10.1016/j.cemconcomp.2017.10.013.
- [27] G. Dimitriou, P. Savva, and M. F. Petrou, “Enhancing mechanical and durability properties of recycled aggregate concrete,” *Constr. Build. Mater.*, vol. 158, pp. 228–235, 2018, doi: 10.1016/j.conbuildmat.2017.09.137.
- [28] B. J. Zhan, D. X. Xuan, and C. S. Poon, “Enhancement of recycled aggregate properties by accelerated CO<sub>2</sub> curing coupled with limewater soaking process,” *Cem. Concr. Compos.*, vol. 89, pp. 230–237, 2018, doi: 10.1016/j.cemconcomp.2018.03.011.
- [29] Y. Li, S. Zhang, R. Wang, Y. Zhao, and C. Men, “Effects of carbonation treatment on the crushing characteristics of recycled coarse aggregates,” *Constr. Build. Mater.*, vol. 201, pp. 408–420, 2019, doi: 10.1016/j.conbuildmat.2018.12.158.
- [30] B. J. Zhan, D. X. Xuan, W. Zeng, and C. S. Poon, “Carbonation treatment of recycled concrete aggregate: Effect on transport properties and steel corrosion of recycled aggregate concrete,” *Cem. Concr. Compos.*, vol. 104, no. July, p. 103360, 2019, doi: 10.1016/j.cemconcomp.2019.103360.
- [31] F. Kazemian, H. Rooholamini, and A. Hassani, “Mechanical and fracture properties of concrete containing treated and untreated recycled concrete aggregates,” *Constr. Build. Mater.*, vol. 209, pp. 690–700, 2019, doi: 10.1016/j.conbuildmat.2019.03.179.
- [32] L. Li *et al.*, “Roles of recycled fine aggregate and carbonated recycled fine aggregate in alkali-activated slag and



- glass powder mortar,” *Constr. Build. Mater.*, vol. 364, no. September 2022, p. 129876, 2023, doi: 10.1016/j.conbuildmat.2022.129876.
- [33] D. Gómez-Cano, Y. P. Arias-Jaramillo, R. Beraal-Conea, and J. I. Tobón, “Effect of enhancement treatments applied to recycled concrete aggregates on concrete durability: A review,” *Mater. Constr.*, vol. 73, no. 349, 2023, doi: 10.3989/mc.2023.296522.
- [34] B. L. Chauhan and G. J. Singh, “Sustainable development of recycled concrete aggregate through optimized acid-mechanical treatment : A simplified approach,” *Constr. Build. Mater.*, vol. 399, no. April, p. 132559, 2023, doi: 10.1016/j.conbuildmat.2023.132559.
- [35] A. Al-Mansour, C. L. Chow, L. Feo, R. Penna, and D. Lau, “Green concrete: By-products utilization and advanced approaches,” *Sustain.*, vol. 11, no. 19, pp. 1–30, 2019, doi: 10.3390/su11195145.
- [36] G. L. Golewski, “Combined Effect of Coal Fly Ash (CFA) and Nanosilica (nS) on the Strength Parameters and Microstructural Properties of Eco-Friendly Concrete,” *Energies*, vol. 16, no. 1, 2023, doi: 10.3390/en16010452.
- [37] Y. Chen, R. Xian, J. Wang, Z. Hu, and W. Wang, “Synergetic Effect of Superabsorbent Polymer and CaO-Based Expansive Agent on Mitigating Autogenous Shrinkage of UHPC Matrix,” *Materials (Basel)*, vol. 16, no. 7, 2023, doi: 10.3390/ma16072814.
- [38] L. Wang, P. Zhang, G. Golewski, and J. Guan, “Editorial: Fabrication and properties of concrete containing industrial waste,” *Front. Mater.*, vol. 10, no. March, pp. 2022–2023, 2023, doi: 10.3389/fmats.2023.1169715.
- [39] W. He, B. Li, X. Meng, and Q. Shen, “Compound Effects of Sodium Chloride and Gypsum on the Compressive Strength and Sulfate Resistance of Slag-Based Geopolymer Concrete,” *Buildings*, vol. 13, no. 3, 2023, doi: 10.3390/buildings13030675.
- [40] A. John, S. Kumar Mittal, and N. Dhapekar, “Applicability of Construction and Demolition Waste Concrete in Construction Sector-Review,” 2017. [Online]. Available: <http://www.ripublication.com>.
- [41] A. Kumar and G. J. Singh, “Improving the physical and mechanical properties of recycled concrete aggregate: A state-of-the-art review,” *Eng. Res. Express*, vol. 5, no. 1, 2023, doi: 10.1088/2631-8695/acc3df.
- [42] IS:269, “IS: 269-2015, ‘ Ordinary Portland Cement’ Indian Standard, New Delhi,” vol. 41, no. December 2015, 2017.
- [43] 2016 IS:383, “Coarse and Fine Aggregate for Concrete — Specification,” *Bur. Indian Stand. New Delhi, India*, vol. 19, no. January, 2016.
- [44] 1963 IS 2386- Part III, “Methods of test for Aggregates for Concrete,” *Bur. Indian Stand. New Delhi, India*, no. Reaffirmed 2021, 1974.
- [45] 1963 IS : 2386 (Part IV ), “Methods of Test for Aggregates for Concrete,” *Bur. Indian Stand. New Delhi*, vol. 2366, no. Reaffirmed 2021, 2002.
- [46] IS 516, “Method of Tests for Strength of Concrete,” *Bur. Indian Stand.*, pp. 1–30, 1959.
- [47] IS 1199, “Methods of sampling and analysis of concrete,” *Bur. Indian Stand.*, pp. 1–49, 1959.
- [48] M. Malešev, V. Radonjanin, and S. Marinković, “Recycled concrete as aggregate for structural concrete production,” *Sustainability*, vol. 02, no. 05, pp. 1204–1225, 2010, doi: 10.3390/su2051204.
- [49] M. Surya, K. R. VVL, and P. Lakshmy, “Recycled Aggregate Concrete for Transportation Infrastructure,” *Procedia - Soc. Behav. Sci.*, vol. 104, pp. 1158–1167, Dec. 2013, doi: 10.1016/j.sbspro.2013.11.212.
- [50] O. Kessal, L. Belagraa, A. Noui, and N. Maafi, “Performance Study of Eco-Concrete Based on Waste Demolition as Recycled Aggregates,” *Mater. Int.*, vol. 2, no. 2, pp. 123–130, Apr. 2020, doi: 10.33263/materials22.123130.
- [51] A. M. Wagih, H. Z. El-Karmoty, M. Ebid, and S. H. Okba, “Recycled construction and demolition concrete waste as aggregate for structural concrete,” *HBRC J.*, vol. 9, no. 3, pp. 193–200, Dec. 2013, doi: 10.1016/j.hbrj.2013.08.007.
- [52] H. S. Joseph, T. Pachiappan, S. Avudaiappan, and E. I. S. Flores, “A Study on Mechanical and Microstructural Characteristics of Concrete Using Recycled Aggregate,” *Materials (Basel)*, vol. 15, no. 21, 2022, doi: 10.3390/ma15217535.

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

# End-to-End Historical Handwritten Ethiopic Text Recognition using Deep Learning

Ruchika Malhotra<sup>3</sup> and Maru Tesfaye Addis<sup>1,2</sup> 

<sup>1</sup>Department of Computer Science, Debre Tabor University, Amhara, Ethiopia (e-mail: [marutes@dtu.edu.et](mailto:marutes@dtu.edu.et))

<sup>2</sup>Department of Computer Science and Engineering, Delhi Technological University, Delhi, India

<sup>3</sup>Department of Software Engineering, Delhi Technological University, Delhi, India

Corresponding author: Maru Tesfaye Addis (e-mail: [marutcomp@gmail.com](mailto:marutcomp@gmail.com)).

**ABSTRACT** Recognizing handwritten text is a challenging task, especially for scripts with numerous alphabets and symbols. The Ethiopic script has a vast character set and is used for historical documents in typewritten, handwritten, and hand-printed forms. However, despite its importance as an ancient script, optical character recognition research has not given enough attention to Ethiopic text recognition. In recent years, deep learning (DL) has emerged as a powerful technique for recognizing patterns. In this study, a DL approach is used to recognize historical Ethiopic handwritten texts. The recognition model uses an end-to-end strategy that enables sequential feature extraction and efficient recognition. An attention mechanism coupled with a connectionist temporal classification architecture is the core of this recognition model architecture. In addition, there are seven convolutional neural networks and two recurrent neural networks. We increase the training data using data augmentation techniques to address the data scarcity common in deep learning applications. The experiments include an original training dataset of 79,684 historical handwritten images and an augmented dataset of 10,000 images containing Ethiopic texts. The model used for recognition showed promising results. For "Test Set I" which had 6,150 samples, the character error rate (CER) was 17.95%, and for "Test Set II" which had 15,935 samples, the CER was 29.95%. These outcomes indicate that this approach has the potential to improve the recognition of historical handwritten Ethiopic text.

**INDEX TERMS** Deep Learning, End-to-End Learning, Ethiopic Script, Handwritten Text Recognition, Pattern Recognition

## I. INTRODUCTION

Ethiopia is incredibly diverse linguistically, with a wide range of languages spoken across its ten states. Each state has its own set of languages, contributing to the country's rich linguistic heritage. The country's linguistic heritage is a testament to its remarkable culture and plays a unifying role among its diverse population. The Ethiopic script is commonly used for writing federal languages such as Amharic and several regional languages including Tigrigna, Awnji, Guragigna, and more. This script plays a significant role in strengthening the cultural and linguistic diversity that thrives in the country.

The Ethiopic script is an old writing system that has been used in Ethiopia and Eritrea. It originated from the South Arabian alphabet [1, 2]. With a history spanning over two thousand years, the Ethiopic script is still in use today, making it one of the few writing systems that has stood the test of time. This script has been used for a variety of purposes, producing historical documents in typewritten, handwritten, and hand-printed formats. For many years, the Ethiopic script has been crucial in safeguarding the linguistic and cultural traditions of Ethiopia and Eritrea. It goes beyond just being a means of communication, as it also serves as a repository of historical records and cultural artifacts. The diversity of languages that use the Ethiopic script adds to its importance, as it showcases the abundant linguistic variety of the region.

In today's world, many government and private organizations are trying to reduce paper usage and move towards digital workflows [3]. However, some institutions such as post offices, banks, and medical institutes often come across handwritten documents in local languages. It's necessary to change these handwritten documents into digital text and customizable formats. There is an increasing demand for efficient optical character recognition (OCR) technology to fulfill this need, which is a crucial component of any text recognition framework [4].

OCR is a system used to transform printed or handwritten text from physical documents into computer understandable formats [3, 5, 6]. Its goal is to transform these documents into electronic versions that can be easily processed and analyzed by computers. Implementing OCR can help organizations improve their document management processes, increase data accessibility, and facilitate efficient information retrieval.

Recognizing handwritten Ethiopic text is challenging due to its vast character set, complex shape, variations in handwriting styles, incomplete strokes, and noise in scanned images. Despite research in pattern recognition for popular scripts, Ethiopic text recognition has not received comparable attention in OCR research [7, 8]. Previous works [9-12] on Ethiopic script OCR have primarily been based on printed texts, and the recognition of handwritten Ethiopic scripts has remained relatively unexplored [8, 13] due to the scarcity of public research datasets [13]. It is rare to find

publicly available historical handwritten datasets, except Belay et al.'s dataset [14], which is typically required for deep learning algorithms [3]. Consequently, recognizing historical handwritten Ethiopic text from image documents poses a unique challenge in OCR. Traditional methods of transcription and analysis can be time-consuming and error-prone, often requiring extensive manual effort. Therefore, there is a need for automated and efficient techniques that can accurately transcribe and interpret the Ethiopic script, facilitating the preservation and understanding of historical documents.

Recent advances in artificial intelligence, specifically the use of deep learning (DL) techniques, have greatly improved pattern recognition. They have proven to be more effective than traditional machine learning (ML) methods [15]. However, these DL techniques require a significant amount of labeled data to work efficiently. Obtaining such data can be difficult and expensive in many cases. To overcome this limitation, image augmentation has emerged as a powerful approach. In computer vision and deep learning, this technique is used to increase the range and size of the training datasets. It enhances the model's ability to generalize and become more resilient. By applying various transformations to the original images, image augmentation replicates real-world scenarios, allowing the model to recognize and handle different image variations more effectively [3].

This advancement in historical handwritten text recognition for the Ethiopic script opens up exciting opportunities for various fields such as linguistic research, historical preservation, and cultural studies. It enables the automated analysis and understanding of historical handwritten texts, which were previously challenging to interpret. Importantly, this study is pioneering in Ethiopic historical handwritten text recognition, being the first of its kind. The study makes significant contributions in the following points:

- 1) Increasing the dataset: This is done through an augmentation technique applied to the original dataset found in [14]. We have expanded the dataset by adding 10,000 handwritten Ethiopic samples. In order to augment the dataset, we carefully diversified it to ensure that it included a broad range of variations and styles in historical handwritten texts. This enriched dataset has empowered the model with a more extensive and diverse training set, enabling it to learn and recognize historical handwritten texts with enhanced robustness and accuracy.
- 2) Comprehensive recognition model development: The proposed recognition model combines the strengths of CNN layers for automatic feature extraction, bidirectional LSTM (BLSTM) for sequencing, and CTC loss functions. Through this comprehensive framework, the model is able to automatically extract relevant features, to focus on important sections of the text, and to leverage the temporal value of the text. Furthermore, it facilitates end-to-end training without requiring explicit alignment between the images and labels [16].

3) Experimental evaluation and promising results: Leveraging the augmented dataset, we conducted extensive experiments, including careful hyperparameter selection, to evaluate the performance of the proposed historical handwritten Ethiopic text recognition. Our proposed approach demonstrated potential effectiveness in historical handwritten image text recognition in preliminary results.

The rest of this paper is structured as follows: In Section II, we'll take a look at related research on Ethiopic script recognition. Section III will cover our suggested methodology, including the deep learning architecture and the training process. We'll then move on to the experimental setup and evaluation results in Section IV. Lastly, in Section V, we'll wrap up the paper and highlight potential areas for further research and development in this field.

## II. Literature Review

Recognition of handwritten Ethiopic text is an area that has yet to be fully explored in both conventional ML and DL. This is particularly true for historical handwritten text recognition. This section discusses the various methods used to recognize handwritten text. More studies have utilized traditional machine learning approaches, while recent studies have employed deep learning and ensemble approaches.

### A. Machine Learning Techniques

When studying OCR, machine learning is essential for tasks such as image preprocessing, feature extraction and recognition. Traditional OCR pipelines using machine learning involve various image preprocessing techniques including image denoising, thresholding, and morphological operations. These techniques improve image quality and eliminate noise or artifacts in preparation for further analysis [17, 18].

After preprocessing the images, various feature extraction algorithms are utilized to capture distinctive information from them. These algorithms may involve techniques such as Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT), or Local Binary Patterns (LBP) [19]. The purpose of these algorithms is to extract relevant features that encode the texture, shape, or structural characteristics of the text in the image [20] [21].

Once the features have been extracted, recognition is carried out using a range of well-known machine learning algorithms such as random forests, support vector machines, k-nearest neighbors, or multilayer perceptrons (MLPs). These algorithms use the extracted features to train models that can accurately recognize text in images.

For optimal OCR results, choose preprocessing, feature extraction, and ML algorithms that fit the task and dataset characteristics. By using ML techniques, researchers can automate text extraction from images, which enables a range of applications, including document digitization, text translation, and information retrieval. These advancements

have opened new doors for linguistic research, historical preservation, and cultural studies, as they allow for automated analysis and comprehension of historical handwritten texts in various scripts, such as the Ethiopic script.

### B. END-TO-END LEARNING

End-to-End (E2E) learning is an approach to machine learning that aims to simplify the process of solving a task by combining all the necessary stages into a single model [22]. E2E learning eliminates the need for handcrafted features or intermediate representations. Instead, the model can learn directly from raw input data to produce the desired output. Recently, there has been a surge in the use of deep learning and neural networks [23]. These models, including CNN and RNN, have shown great success [24] in fields such as computer vision, natural language processing, speech recognition, and pattern recognition [25]. One major benefit of E2E learning is its simplicity in design. It eliminates the need for complicated feature engineering, making the overall system design more straightforward. Moreover, E2E learning models are adaptable and can conform to various domains and tasks. They can learn pertinent features and representations from raw data, capturing complex patterns and dependencies effectively [26].

It's essential to recognize that E2E learning models have specific limitations. These models require a significant amount of labeled training data to perform effectively in different situations, which can be an expensive and time-consuming process. Moreover, the integrated nature of E2E models can make it difficult to repurpose or adjust specific components for various tasks [27]. E2E learning is a powerful technique that helps models generate desired outputs from raw data. It requires labeled data and may be difficult to interpret or modularize.

### C. Handwritten Ethiopic Text Recognition

Recognizing handwritten Ethiopic text is challenging due to the intricate nature of the Ethiopic script and the limited availability of annotated datasets. Despite this, some studies in recent years have suggested both conventional ML and DL techniques to recognize handwritten Ethiopic text. Here are some examples:

Alemu and Fuchs [2] published a research paper outlining their approach to recognizing handwritten Amharic bank checks. This study is significant as it is the first of its kind to tackle this challenge. The main obstacle in recognizing handwritten Amharic text is that different writers may use different writing styles for the same numerals. HMRF can be used to extract relevant features from handwritten characters in order to address this issue. The HMRF algorithm can model context-dependent entities based on the concept of Markov Random Field (MRF) theory, which considers both local and global interactions. HMRF is not typically used in handwritten recognition because of its extensive computational demands. However, the authors considered the possibility of acceptable

time and space consumption in their approach, making a valuable contribution to handwriting recognition, particularly in the Amharic language. Their recognition model was evaluated using training images consisting of 7,240 characters and a testing image set consisting of 713 characters. They incorporated contextual information in their approach, leading to a significant increase in accuracy from 89.06% to an impressive 99.44%. It is worth noting that their study did not consider the influence of prior probability, which could have an impact on the obtained results.

A study conducted by Assabie and Begun [28] focused on recognizing offline handwritten Amharic words. The goal was to tackle the challenges posed by the large number of character classes and the complexity of Amharic script. To achieve this, they used a hidden Markov model (HMM) for recognizing unconstrained handwritten Amharic words. The authors' recognition pipeline involved feature extraction and recognition steps. They used direction field image computation and segmentation techniques at the text line, word, and pseudo-character levels to capture the unique structural characteristics of the Amharic script. According to their experiment, the recognition rate was 76% for good-quality image categories and 53% for poor-quality images. However, the authors acknowledged that further improvements could be achieved by enhancing the extraction of structural features and incorporating language models into the HMM framework. Future work could focus on refining the extraction of structural features, such as curves, loops, and junctions, and integrating language models, such as n-grams or recurrent neural networks, into the HMM-based recognition system. By addressing these aspects, the recognition accuracy of Amharic handwritten words can be further improved, which contributes to the development of more robust and accurate systems for Amharic language processing.

In their research, Tamir [29] explored handwriting recognition of Amharic characters using CNN. The goal of this study was to overcome the challenges faced in managing ancient handwritten documents in Amharic, which are prone to deterioration over time. An automatic approach to handwritten text recognition was proposed, utilizing a CNN design consisting of two convolutional layers to classify handwritten Amharic characters. Batch Normalization and Activation layers were sequentially implemented after each convolutional layer. In addition, Max-pooling was executed after the activation layer of the second convolutional layer. Finally, the output was flattened to be fully connected to the final layer, which was responsible for character classification. To conduct the research, data was collected from about 130 individuals, resulting in a dataset of 30,446 characters. Of this dataset, 27,413 characters were used for model training, while the remaining 3,033 were reserved for testing and evaluating the model's performance. Although the author provided information on the training accuracy and loss, it is important to also have validation accuracy and loss metrics to determine the model's ability to generalize. These metrics allow for

assessing how well the model performs on new data and provide insights into its overall effectiveness. It is recommended that the author includes both validation accuracy and validation loss in their report for better research. This will help to evaluate the model's performance comprehensively and determine its capacity to identify new and unseen handwritten Amharic characters. Furthermore, it would be beneficial to analyze and interpret the results to highlight the strengths and limitations of the proposed method and identify areas for future research and improvement.

In a study by Agegnehu et al. [30], deep learning was employed to identify Amharic punctuation marks and handwritten digits using a CNN architecture. The dataset comprised of 5,800 images sourced from 100 handwriting samples, with a testing accuracy of 70.04% achieved. Additionally, research has also investigated Ge'ez digits in contexts other than printed and handwritten forms.

Most researchers have primarily concentrated on recognizing English numerals, and the lack of openly accessible Ge'ez digit datasets has hindered extensive research in this field. Nonetheless, Nur et al. [13] have made progress in Ge'ez digit recognition by creating a recognition model with better accuracy and an experimental dataset consisting of 51,952-digit images written by 524 people. Their proposed CNN model has achieved a recognition accuracy of 96.21%. The authors have suggested that future research should explore various deep-learning approaches for multi-digit recognition.

Our research is focused on recognizing historical handwritten Ethiopic text in images by utilizing E2E approach. To achieve this, we acknowledge the importance of deep convolutional recurrent neural networks (CRNNs) and connectionist temporal classification (CTC) for the efficient application of end-to-end learning. CNNs have proven crucial in image processing and pattern recognition, especially in tasks involving automatic feature extraction from images [10].

### III. Methods

This section presents the E2E historical handwritten Ethiopic image text recognition using DL. This study aims to develop a recognition model that accurately recognizes handwritten text from an image. An explanation of the model's architecture and the dataset used for training and evaluating the model is provided.

#### A. THE DATASET PREPARATION

During this research, we realized that having a large dataset is crucial for developing and testing deep learning models. To overcome the challenge of limited data, we needed a dataset that was specially designed for our deep learning problem.

The dataset used in this study called "HHD-Ethiopic," which was created by Belay et al. [14]. This dataset contained 79,684 images of handwritten historical documents that featured 306 unique Ethiopic characters. We divided the data into training, validation, and testing subsets. The training set



had 57,374 samples, while the remaining samples were kept for testing.

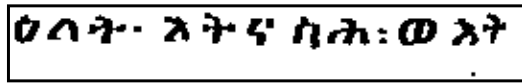
To ensure accurate assessment, we created a validation dataset using 10% of the training dataset. The testing dataset was divided into two parts: "Test Set I" and "Test Set II." Test Set I had 6,375 images randomly chosen from the training set while Test Set II contained 15,395 images from 18th-century manuscripts. The original dataset is available on the GitHub repository (<https://github.com/bdu-birhanu/HHD-Ethiopic/tree/main/Dataset>). We faced the challenge of having a small dataset for deep learning tasks. To increase its size, we used data augmentation techniques. Despite resource limitations, we were able to add around 10,000 augmented images. Sample augmented image is shown in **FIGURE 1**. By using image augmentation, deep learning models can learn effectively from limited labeled data. This leads to better performance and more reliable predictions. This technique has become an essential tool in overcoming data scarcity and maximizing the potential of DL algorithms. These algorithms excel in various applications across different domains.

Image augmentation techniques involve applying operations like rotations, shifting, zooming, shearing, flipping, and noise injection. These transformations create new training

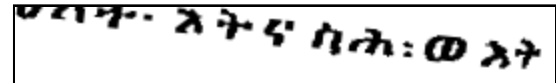
examples that are variations of the original images while preserving their semantic content. By introducing such variations, the model becomes more robust to changes in factors that can be encountered during inference on real-world data.

During the training process, augmented images are commonly utilized to offer a wider range of examples for the model to learn from. This technique helps to prevent overfitting and enhances the model's capacity to perform well on unseen data. Image augmentation is particularly useful when the available training dataset is limited, as it enables the creation of additional training samples without the need for manual data collection or annotation.

The primary objective of expanding the dataset through data augmentation was to enhance the model's performance and generalization abilities. By providing a larger and more diverse set of examples for the model to learn from, we aimed to improve its ability to recognize and handle various patterns and variations in the data. To illustrate this, sample augmented images can be seen in **FIGURE 1**, which depicts how the dataset was enriched to facilitate better learning and adaptation in our deep learning approach.



(a)



(b)

**FIGURE 1.** Sample images from the HHD-Ethiopic database (a) represents the original image and (b) represents the augmented image.

## B. ARCHITECTURE OF THE PROPOSED MODEL

This study utilizes the advantages of several deep learning algorithms by combining them, namely CNN, BLSTM, Attention mechanisms, and CTC. Each algorithm has a crucial role in different recognition stages. By bringing these algorithms together, we create a comprehensive and effective approach to recognizing handwritten text. Our careful selection and use of these algorithms contribute to the success of this study and enable more accurate and reliable recognition. Below is a detailed explanation of each algorithm.

### 1) CNN LAYERS

Our study aimed to extract deeper features from handwritten text images using CNNs, which are known for their exceptional ability to identify patterns and significant information within images [31]. Previous research has successfully applied CNNs to various tasks [32], and we utilized them to accurately recognize the distinctive characteristics of handwritten text. The CNN's parameters were carefully selected and fine-tuned to ensure the extracted features were precise.

To represent the mathematical effectiveness of CNNs in feature extraction, let's consider the input handwritten text

image  $I$  with pixel values denoted by  $I(x, y)$ , where  $x$  and  $y$  are the spatial coordinates of the image. A convolutional layer applies a set of learnable filters (also known as kernels) to the input image. The output of the convolutional layer can be computed in "(1)".

$$O(i, j) = \sum_{m=1}^M \sum_{n=1}^N I(i+m, j+n) \cdot W(m, n) \quad (1)$$

Where:

- $O(i, j)$  represents the output feature map at the spatial location  $(i, j)$ ,  $W(m, n)$  denotes the learnable weights of the filter at the position  $(m, n)$ ,  $M$  and  $N$  are the dimensions of the filter.

After the convolution operation, an activation function such as Rectified Linear Unit (ReLU) is applied element-wise to introduce non-linearity using "(2)".

$$F(i, j) = \text{ReLU}(O(i, j)) \quad (2)$$

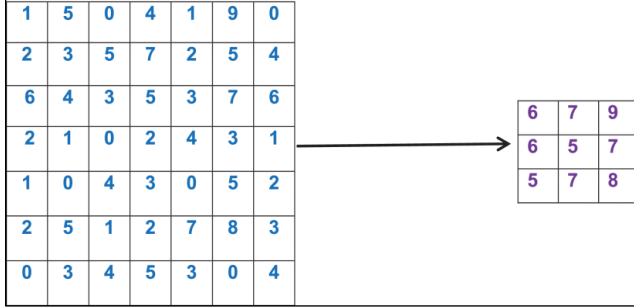
The process of pooling (commonly max-pooling) reduces the spatial dimensions of the feature maps, which helps in



reducing computational complexity and controlling overfitting. The pooled output is calculated in “(3)”.

$$P(i, j) = \max(F(s \cdot i, s \cdot j)) \quad (3)$$

Where  $P(i, j)$  represents the pooled output, and  $s$  is the stride of the pooling operation. The max pooling results with a 3x3 filter with strides 2 are shown in **FIGURE 2**.



**FIGURE 2.** Max pool with 3 × 3 filter and 2 strides [31].

The mathematical representations and fine-tuning of CNN parameters play a crucial role in this feature extraction process, making our study a significant contribution to the field of image recognition and analysis.

## 2) BLSTM LAYERS

In this study, we leverage the power of BLSTM networks to model the difficult sequential dependencies and temporal dynamics inherent in handwritten text. The inclusion of BLSTM layers empowers the network to efficiently capture contextual information [32], thereby significantly enhancing the accuracy of recognition [4].

“Equation (4)-(7)” is used to compute the mathematical formulation of a BLSTM unit. At each time step  $t$ , the BLSTM unit takes as input the hidden state  $h_{t-1}$  from the previous time step, the current input  $x_t$  (which can be a vector representation of a character or a pixel in the handwritten image), and computes the following intermediate values.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (4)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (5)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (6)$$

$$g_t = \tanh(W_g \cdot [h_{t-1}, x_t] + b_g) \quad (7)$$

Where:

- $\sigma$  represents the sigmoid activation function,  $\tanh$  denotes the hyperbolic tangent activation function,  $W_f, W_i, W_o$ , and  $W_g$  are the learnable weight matrices,  $b_f, b_i, b_o, b_g$  are the learnable bias vectors,

$[h_{t-1}, x_t]$  denotes the concatenation of  $h_{t-1}$  and  $x_t$  along the feature dimension. The intermediate values  $f_t, i_t, o_t$ , and  $g_t$  are used to update the cell state  $c_t$  and the hidden state  $h_t$  of the BLSTM unit as in “(8)” and “(9)”.

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (8)$$

$$h_t = o_t \odot \tanh(c_t) \quad (9)$$

Where  $\odot$  denotes the element-wise multiplication.

The output of the BLSTM layer is then used for further processing, such as classification in the case of handwritten text recognition.

By incorporating BLSTM layers into our model, we equip it with the ability to comprehend the underlying structure and context of the handwritten text comprehensively. This deep understanding of sequential information facilitates improved accuracy in recognition tasks, making our study a valuable contribution to the field of handwritten text analysis and understanding.

## 3) ATTENTION LAYERS

Attention mechanisms have proven to be effective in enhancing the model's ability to focus on relevant regions within the handwritten text. By dynamically weighting different parts of the sequence during recognition, the model can allocate its attention to the most salient features and characters, thereby improving recognition accuracy, especially when dealing with inter-class similarity and structural complexity challenges.

The attention mechanism can be mathematically described as follows:

Let  $H = \{h_1, h_2, \dots, h_T\}$  be the set of hidden states produced by the encoder, where  $T$  is the length of the input sequence. These hidden states capture valuable information about the handwritten text at each time step. The attention mechanism generates a set of context vectors  $C = \{c_1, c_2, \dots, c_T\}$  that represent the weighted combinations of the encoder's hidden states. The context vector  $c_t$  at the time step  $t$  is computed as a weighted sum of all hidden states  $h_i$  with attention weights  $a_{t,i}$  in “(12)”.

$$c_t = \sum_{i=1}^T a_{t,i} \cdot h_i \quad (12)$$

The attention weights  $a_{t,i}$  are typically calculated using a scoring function that measures the relevance of the input at the time step  $i$  to the output at the time step  $t$ . One commonly used scoring function is the dot product as represented in “(13)”.

$$e_{t,i} = h_t^T \cdot h_i \quad (13)$$

Where  $e_{t,i}$  represents the score between the hidden state at the time step  $t$  and the hidden state at the time step  $i$ .

The scores  $e_{t,i}$  are then normalized using the SoftMax function to obtain the attention weights calculated in "(14)".

$$a_{t,i} = \frac{\exp(e_{t,i})}{\sum_{j=1}^T \exp(e_{t,j})} \quad (14)$$

By calculating the attention weights for each time step, the model can effectively focus on the most relevant regions within the handwritten text during recognition. This adaptability allows the model to emphasize distinctive features and characters, thereby overcoming challenges posed by inter-class similarity and structural complexity.

The context vectors  $C$  are then combined with the decoder's hidden states to generate the final output sequence during the decoding (inference) process.

In summary, the integration of attention mechanisms enables the model to dynamically allocate its focus on salient regions within the handwritten text, leading to improved recognition accuracy, particularly in scenarios where class similarities and structural intricacies are prominent. This mathematical formulation of attention mechanisms highlights their significant impact on the performance of the model in handwritten text recognition tasks.

#### 4) CTC LAYER

CTC is a valuable technique used in sequence recognition tasks, particularly for applications like handwritten text recognition. CTC enables the model to learn from sequences of variable length without requiring explicit alignment between input images and their corresponding labels. This makes it a powerful tool for end-to-end training and decoding.

The CTC loss function is employed to train the model. Given an input sequence of feature vectors (representing the handwritten text image) denoted by  $X = \{x_1, x_2, \dots, x_T\}$ , and the corresponding label sequence denoted by  $Y = \{y_1, y_2, \dots, y_U\}$  where  $T$  and  $U$  are the lengths of the input sequence and label sequence, respectively, the CTC loss is computed using "(15)".

Let's define  $S$  as the set of all possible labels, including a special "blank" symbol denoted by  $\emptyset$ . The CTC loss function  $L(X, Y)$  is the negative log-likelihood of the correct label sequence given the input sequence  $X$ .

$$L(X, Y) = -\log p(Y | X) \quad (15)$$

The probability  $p(X | Y)$  is computed by summing over all valid alignments between the input sequence  $X$  and the label sequence  $Y$  in "(16)".

$$p(Y | X) = \sum_{align \in A(X, Y)} \prod_{t=1}^T p(y_{align(t)} | x_t) \quad (16)$$

Where  $A(X, Y)$  is the set of valid alignments, considering the blank symbol and allowed repetitions and removals of labels during the alignment process.

During training, the CTC loss is minimized using gradient-based optimization techniques like stochastic gradient descent (SGD) or Adam to update the model's parameters, allowing it to learn to recognize handwritten text effectively.

Furthermore, during decoding (inference), CTC enables the model to produce the most probable label sequence directly from the input sequence without requiring explicit alignment. This is achieved through a decoding algorithm, such as the beam search algorithm, which explores the possible label sequences and selects the most likely output sequence.

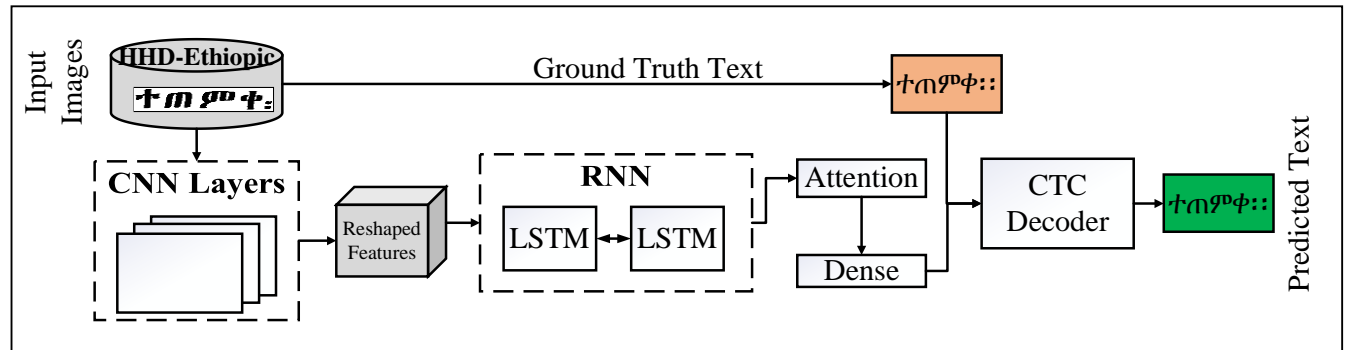


FIGURE 3. The general architecture of the proposed historical handwritten Ethiopic text recognition.

The general architecture for historical handwritten Ethiopic text recognition is illustrated in FIGURE 3. The diagram depicts the flow of the recognition model, starting with the input image obtained from the "HHD-Ethiopic" image database. The input image dimensions, in our case, 48 by 368

pixels, are then passed through a series of CNN layers. The model architecture consists of seven convolutional layers along with ReLU activation function for each, three max-pooling, and two batch normalization, as depicted in TABLE I. The extracted features from the CNN layers are then passed

to the next two BLSTM layers for sequence modeling. Each LSTM layer is constructed with 128 hidden units. Next, the attention mechanism is applied by concatenating the LSTM layers. A fully connected or dense layer with Soft Max activation processes features and prepares them for classification. The number of units in the dense layer is determined based on the number of classes or categories in the recognition task. A total of 306 alphabet symbols are included

in this study. Therefore, there are 307 classes, including the blank CTC space. The model is compiled with the Adam optimizer and CTC loss. Finally, the CTC decoder generates the predicted result by cross-checking the ground truth labels with the alphabet or character set. The CTC decoder ensures that the recognition model can handle sequences of variable lengths and directly learn from input-output alignment without explicit alignment.

**TABLE I**

THE PROPOSED HISTORICAL HANDWRITTEN ETHIOPIC TEXT RECOGNITION LAYERS AND THEIR HYPER-PARAMETER VALUES

S. No	Layers	Configuration
1.	Input	Rows:48, Columns: 368, Channels:1
2.	Conv1	Feature:32, kernel:(3,3), activation: ReLU, padding: same
3.	MaxPooling1	Pooling size: (2,1), strides:2
4.	Conv2	Feature:32, kernel:(3,3), activation: ReLU, padding: same
5.	MaxPooling2	Pooling size: (2,2)
6.	Conv3	Feature:32, kernel:(3,3), activation: ReLU, padding: same
7.	Conv4	Feature:32, kernel:(3,3), activation: ReLU, padding: same
8.	MaxPooling3	Pooling size: (2,2)
9.	Normalize	Batch Normalization
10.	Conv5	Feature:32, kernel:(3,3), activation: ReLU, padding: same
11.	Conv6	Feature:32, kernel:(3,3), activation: ReLU, padding: same
12.	Normalize	Batch Normalization
13.	Conv7	Feature:128, kernel:(2,2), activation: ReLU
14.	BLSTM1	RNN Units: 128, droout:0.25, return sequence: True
15.	BLSTM2	RNN Units: 128, droout:0.25, return sequence: True
16.	Attention	Feature: 91, activation: tanh, SoftMax
17.	Dense	Units: no_class+1, activation: SoftMax
18.	Output + CTC	max_len:46, input length: Label length:

### C. Performance Evaluation Metrics

Our proposed model's performance is assessed using character error rate (CER). The CER is calculated using equation (1).

$$CER = \left( \frac{(I + D + S)}{GT} \right) * 100, \quad (15)$$

Here,  $I$ ,  $D$ , and  $S$  represent the number of character insertions, deletions, and substitutions respectively.  $GT$  denotes the total number of characters in the ground truth text. The CER provide valuable insights into the error rates of the recognition model. By evaluating these metrics, we can measure the effectiveness of our proposed model in accurately transcribing characters, and identify areas where improvements may be required.

We used additional evaluation metrics, alongside our primary CER metric, to provide a comprehensive analysis of the model's performance. These metrics, including precision, recall, and F1-score, focus on evaluating the model's

classification accuracy, specifically in identifying characters and their positions in the sequence.

To better understand these metrics, let's explain the parameters used in generating the classification report. TP (true positive) refers to the number of positive instances correctly predicted, while FP (false positive) refers to the number of positive instances predicted incorrectly. Similarly, TN (true negative) represents the number of negative instances correctly predicted, and FN (false negative) signifies the number of negative instances predicted incorrectly. "Equations (16)", "(17)", and "(18)" are used to calculate precision, recall, and F1-score respectively.

$$P = \frac{TP}{(TP + FP)}, \quad (16)$$

$$R = \frac{TP}{(TP + FN)}, \quad (17)$$

$$F = \frac{2PR}{R+P}, \quad (18)$$

Here, the precision value is represented by P, recall value is represented by R, and F-score is represented by F.

#### IV. Experimental Setup, Results, and Discussion

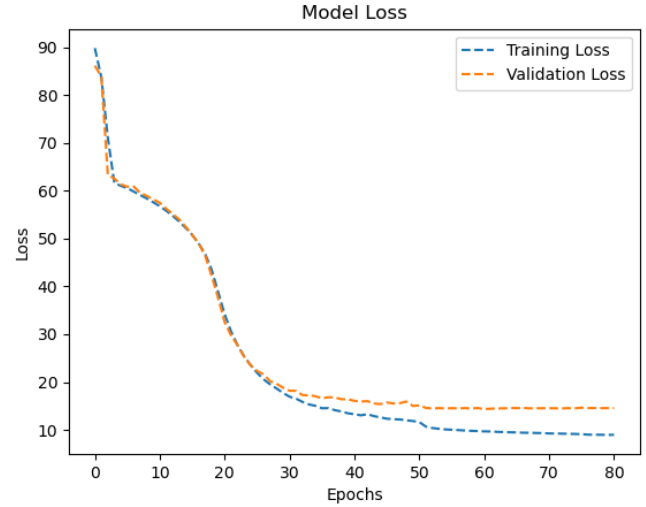
We conducted extensive experiments utilizing deep learning techniques to demonstrate the efficacy of the proposed historical handwritten text recognition model. In this section, we present experimental setup and a detailed overview of the performance evaluation measures employed, followed by a description of the experimental setup. We then present the results and engage in a thorough discussion of the experiments conducted using two testing datasets. Furthermore, we discuss the techniques employed for selecting optimal parameters for the designed architectural model and provide the results obtained from these optimal parameter settings. Finally, we perform an in-depth error analysis of the results.

##### A. EXPERIMENTAL SETUP

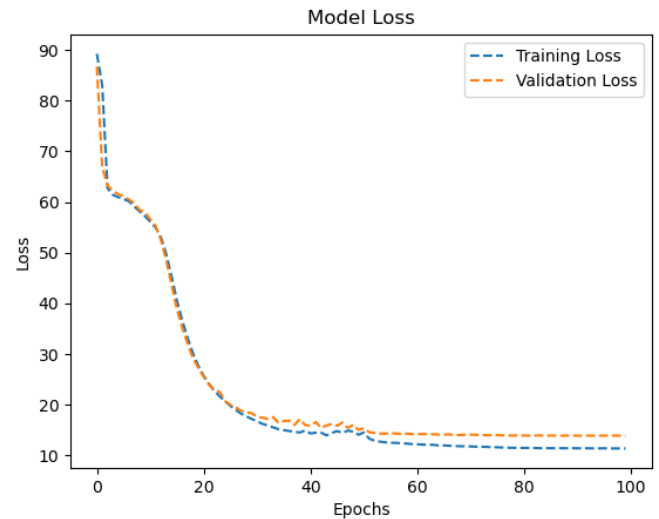
The proposed model was implemented in Python using the Keras framework with TensorFlow as the backend. This combination allowed for the efficient development and training of the model. Powerful GPUs were utilized to expedite the training process and take advantage of accelerated computation. This was made possible through Kaggle notebooks, eliminating expensive hardware. Kaggle, a cloud-based service, offers remote code execution from any location with an internet connection. Its user-friendly interface and comprehensive support for popular data science libraries make it an excellent platform for deep learning projects. We streamlined the development process by utilizing Kaggle's capabilities and efficiently training the proposed model.

##### B. RESULTS

In this research, we created two recognition models with 100 epochs and a batch size of 64, each selected numerically. The first model utilized the original historical handwritten images, while the second model incorporated augmentation by adding 10,000 images generated from the original dataset. In the following results section, we present the outcomes achieved by both historical handwritten text recognition models for the Ethiopic script. **FIGURE 4** and **FIGURE 5** showcase the training and validation losses plotted against the number of epochs for the first and second models, respectively. These graphs provide valuable insights into each model's learning progress and performance throughout the training process.



**FIGURE 4.** Training and validation loss vs epoch graph for the first model



**FIGURE 5.** Training and validation loss vs epoch graph for the second model.

**TABLE II** and **TABLE III** displays the CER values of the first and the second model respectively. The number of samples in each testing set also described.

**TABLE II**

CER TEST RESULTS OF THE FIRST MODEL

Test Set Type	Samples	CER
Test Set I	6,150	18.33%
Test Set II	15,935	30.69%

**TABLE III**

CER TEST RESULTS OF THE SECOND MODEL

Test Set Type	Samples	CER
Test Set I	6,150	17.95%
Test Set II	15,935	29.95%

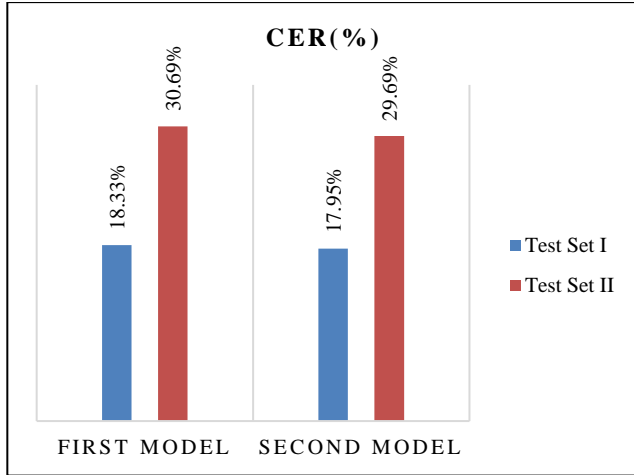


FIGURE 6. CER comparison of the first and the second models.

TABLE IV

PRECISION, RECALL, F-SCORE RESULTS OF THE BETTER MODEL.

Test set type	Precision	Recall	F-score	Support
Test set I	0.9014	0.8780	0.8895	89780
Test set II	0.8488	0.8011	0.8243	280917

## 5) ERROR RESULTS

In FIGURE 7, we present the error results from the 18<sup>th</sup> century testing dataset for the second recognition model. The text on the left shows the actual ground-truth texts, while the text on the right displays the predicted texts generated by our recognition model. Characters that were incorrectly recognized by the recognition model are highlighted in green, while characters that were present in the ground-truth texts but not identified as characters in the predictions are shown in blue. In the predicted text, characters that are not recognized correctly are colored in red.

## 6) CORRECT RESULTS

FIGURE 8, displays sample images which are recognized by correctly by the proposed models.

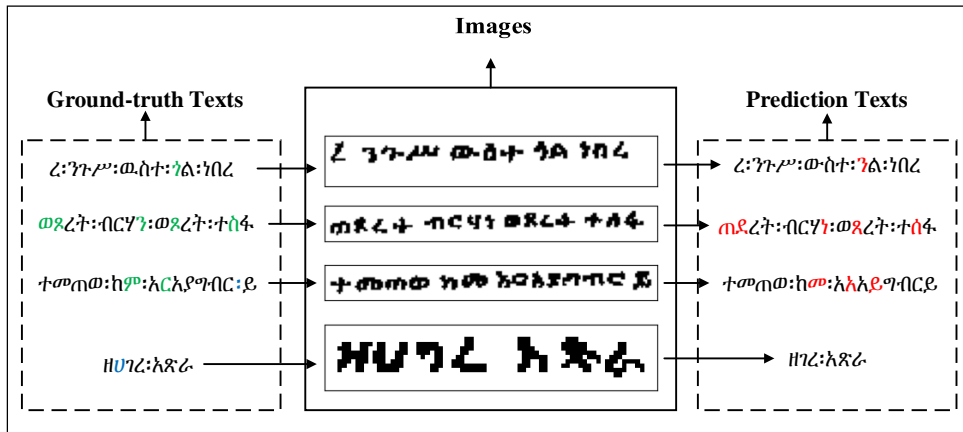


FIGURE 7. Sample error results from the second model using "Test Set II".

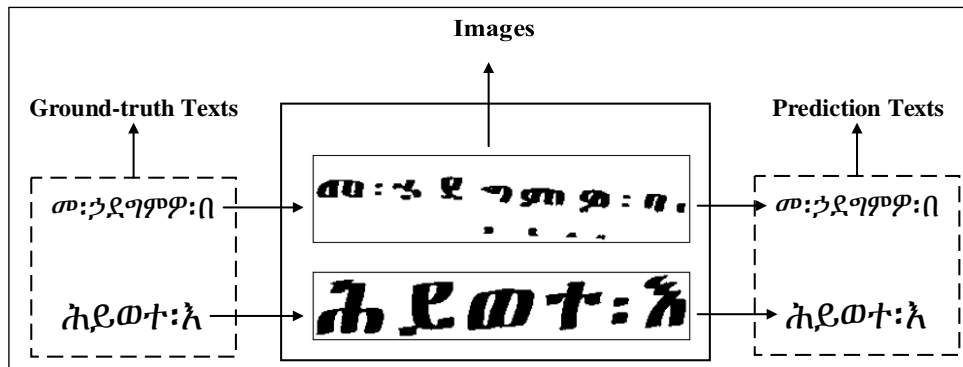


FIGURE 8. Sample correctly recognized results in both models.



### C. Discussion

This study compares two models for recognizing handwritten Ethiopic text. The first model uses the "HDD-Ethiopic" dataset, while the second model uses augmented images. The results show that the model trained with augmented images performs better in terms of CER on both testing datasets, "test set I" and "test set II", as shown in **TABLE II** and **TABLE III**. The first model achieved a CER of 18.33% and 30.69% for "test set I" and "test set II", respectively. The second model achieved a CER of 17.95% and 29.95% for "test set I" and "test set II", respectively. The reason for the second model's better performance is the use of augmented images during training. This expands the training data, improves model generalization, and enhances the ability to recognize diverse variations in handwritten Ethiopic text. As a result, the model trained with augmented images achieves superior results.

We analyzed the superior model or second model and presented its precision, recall, and f-score results in **TABLE IV** for both testing datasets. The second model performed better in all metrics for test set I, confirming its superiority. However, test set II had more instances of highly confused characters, despite having a larger support.

The character “:” is frequently used in this dataset. Because this character is used in every word to separate. So, it has better recognition results for precision (0.9973), recall (0.9955), and F-score (0.9964).

Ethiopic characters often share similar or approximate shapes, which can lead to confusion during recognition. For example, as we observed in **FIGURE 7**, the character “?” is recognized as “?”. When we see these two characters almost they have similar shape on the top part. The character “w” is recognized as “m”. Here from the ground truth text “w” is connected but the predicted character “m” doesn’t have connection at the bottom of the character. This is the minor difference in terms of shape. Other characters also misrecognized as “z” to “x”, “y” to “v” and “h” to “u”.

Some texts are fully recognized correctly with in the sequence. We can observe from **FIGURE 8**. Here all the characters inside the sequence are fully recognized.

### V. Conclusion

In summary, the recognition of Ethiopic OCR is an important area of research with numerous applications. Therefore, a recognition model is necessary for this purpose. This study presents an end-to-end learning model that is specifically designed to recognize historical handwritten Ethiopic text. With the use of deep learning techniques, our model automatically extracts relevant features from input images and effectively captures the sequential nature of the text. Additionally, the integration of attention mechanisms and a CTC-based loss function enables the model to concentrate on crucial regions of the input, allowing for end-to-end training without requiring explicit alignment between images and

labels. Our proposed approach was found to be effective, as demonstrated by experimental results.

To improve our recognition system, we can expand the dataset, explore alternative deep learning architectures, and test it in other languages. Using a more diverse dataset will enhance performance and reduce errors. With ongoing research, we can achieve more accurate recognition, benefiting various applications.

### ACKNOWLEDGMENT

We would like to express our sincere gratitude to the faculty members, staff, and students of Delhi Technological University for their invaluable guidance and support throughout the completion of this manuscript. Their continuous assistance and expertise have played a crucial role in the success of this research project. We are grateful for their unwavering commitment to fostering academic excellence and their contributions to our growth as researchers. Additionally, “ChatGPT (GPT-3.5)” and “Word tune” have been extremely helpful in the development of this study.

### REFERENCES

- [1] R. Meyer, "The Ethiopic script: linguistic features and socio-cultural connotations," *Multilingual Ethiopia: Linguistic Challenges and Capacity Building Efforts*, vol. 8, no. 1, pp. 137–172, 2016.
- [2] W. Alemu and S. Fuchs, "Handwritten Amharic Bank Check Recognition Using Hidden Markov Random Field," in *2003 Conference on Computer Vision and Pattern Recognition Workshop*, 2003, vol. 3, pp. 28–28.
- [3] P. Goel and A. Garatra, "Handwritten Gujarati Numerals Classification Based on Deep Convolution Neural Networks Using Transfer Learning Scenarios," *IEEE Access*, vol. 11, pp. 20202–20215, 2023.
- [4] Y. S. Chernyshova, A. V. Sheshkus, and V. V. Arlazarov, "Two-Step CNN Framework for Text Line Recognition in Camera-Captured Images," *IEEE Access*, vol. 8, pp. 32587–32600, 2020.
- [5] J. Memon, M. Sami, R. A. Khan, and M. Uddin, "Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)," *IEEE Access*, vol. 8, pp. 142642–142668, 2020.
- [6] T. Nasir, M. K. Malik, and K. Shahzad, "MMU-OCR-21: Towards End-to-End Urdu Text Recognition Using Deep Learning," *IEEE Access*, vol. 9, pp. 124945–124962, 2021.
- [7] E. Y. Obsie, H. QU, and Q. Huang, "Amharic Character Recognition Based on Features Extracted by CNN and Auto-Encoder Models," in *The 13th International Conference on Computer Modeling and Simulation*, Melbourne VIC, Australia, 2021, vol. 21, pp. 58–66: ACM.
- [8] B. Belay, T. Habtegebrial, M. Liwicki, G. Belay, and D. Stricker, "Factored Convolutional Neural Network for Amharic Character Image Recognition," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 2906–2910.
- [9] M. Meshesha and C. V. Jawahar, "Recognition of printed Amharic documents," presented at the Eighth International Conference on Document Analysis and Recognition (ICDAR'05), Seoul, Korea (South), 31 Aug.–1 Sept., 2005. Available: <https://ieeexplore.ieee.org/ielx5/10526/33307/01575652.pdf?tp=&arnumber=1575652&isnumber=33307&ref=>
- [10] B. H. Belay, T. A. Habtegebrial, and D. Stricker, "Amharic Character Image Recognition," presented at the 18th IEEE International Conference on Communication Technology, Chongqing, China, 08–11 October 2018.



- [11] R. Malhotra and M. T. Addis, "Ethiopic Base Characters Image Recognition using LSTM," presented at the 2021 2nd International Conference on Computational Methods in Science & Technology (ICCMST), Mohali, India, 2021.
- [12] D. Addis, C. D. Liu, and V. D. Ta, "Printed Ethiopic Script Recognition by Using LSTM Networks," presented at the 2018 International Conference on System Science and Engineering (ICSSE), New Taipei, Taiwan, 28-30 June, 2018. Available: <https://ieeexplore.ieee.org/ielx7/8500054/8519965/8519972.pdf?tp=&arnumber=8519972&isnumber=8519965&ref=>
- [13] M. A. Nur, M. Abebe, and R. S. Rajendran, "Handwritten Geez Digit Recognition Using Deep Learning," *Applied Computational Intelligence and Soft Computing*, Open access vol. 2022, pp. 1-12, 2022.
- [14] B. H. Belay *et al.*, "HHD-Ethiopic: A Historical Handwritten Dataset for Ethiopic OCR with Baseline Models and Human-level Performance (Revision 50c1e04)," ed. Hugging Face, 2023.
- [15] S. Aly and A. Mohamed, "Unknown-Length Handwritten Numeral String Recognition Using Cascade of PCA-SVMNet Classifiers," *IEEE Access*, vol. 7, pp. 52024 - 52034, 2019.
- [16] Y. Zhu, Z. Xie, L. Jin, X. Chen, Y. Huang, and M. Zhang, "SCUT-EPT: New Dataset and Benchmark for Offline Chinese Text Recognition in Examination Paper," *IEEE Access*, vol. 7, pp. 370-382, 2019.
- [17] M. Sonka, V. Hlavac, and R. Boyle, *Sonka, M., Hlavac, V., & Boyle, R. (2014). United States of America: Global Engineering: Timothy L. Anderson, 2014, p. 930.*
- [18] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital Image Processing Using MATLAB. Prentice Hall*. Prentice Hall, 2004, p. 302.
- [19] D. B. Honnaraju, M. Meghana, D. S. Sanjana, N. S. Nisarga, and H. R. Nikhil, "SIGN LANGUAGE RECOGNITION USING DEEP LEARNING (CNN) AND SVM " *International Research Journal of Modernization in Engineering Technology and Science*, vol. 05, no. 05, pp. 6479-6483, 2023.
- [20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, 2005, vol. 1, pp. 886-893: IEEE.
- [21] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 1, 2004.
- [22] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436-444, 2015.
- [23] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* Cambridge, Massachusetts London, England MIT press, 2016
- [24] M. G. Gurmu, "Offline Handwritten Text Recognition of Historical Ge'ez Manuscripts Using Deep Learning Techniques," MSc, Information Science, Jimma University, Jimma, Ethiopia, 2021.
- [25] N. Mungoli, "Adaptive Feature Fusion: Enhancing Generalization in Deep Learning Models," *International Journal of Computer Science and Mobile Applications*, vol. 11, no. 3, pp. 1-11, 2023.
- [26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," presented at the Proceedings of the IEE, 1998.
- [27] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," presented at the 2nd International Conference on Learning Representations, ICLR 2014, Banff, Canada, 14-16 April 2014.
- [28] Y. Assabie and J. Bigun, "HMM-Based Handwritten Amharic Word Recognition with Feature Concatenation," in *2009 10th International Conference on Document Analysis and Recognition*, 2009, pp. 961-965.
- [29] K. Tamir, "Handwritten Amharic characters Recognition Using CNN," presented at the 2019 IEEE AFRICON, Accra, Ghana, 2019.
- [30] M. Agegnehu, G. Tigistu, and M. Samuel, "Offline Handwritten Amharic Digit and Punctuation Mark Script Recognition using Deep learning," presented at the 2nd Deep Learning Indaba-X Ethiopia Conference 2021, Adama, Ethiopia, January 27 – 29, 2022.
- [31] H. T. Weldegebriel, H. Liu, A. U. Haq, E. Bugingo, and D. Zhang, "A New Hybrid Convolutional Neural Network and eXtreme Gradient Boosting Classifier for Recognizing Handwritten Ethiopian Characters," *IEEE Access*, vol. 8, pp. 17804-17818, 2020.
- [32] B. Shi, X. Bai, and C. Yao, "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition," *IEEE Trans Pattern Anal Mach Intell*, vol. 39, no. 11, pp. 2298-2304, Nov 2017.



**Prof. Dr. Ruchika Malhotra** is a Professor in the Department of Software Engineering at Delhi Technological University, Delhi, India. She received her master's and doctorate degrees in software engineering from the University School of Information Technology, Guru Gobind Singh Indraprastha University, Delhi, India. Her research interests are in software testing, improving software quality, statistical and adaptive prediction models, software metrics, neural nets modeling, and the definition and validation of software metrics. She has published more than 200 research papers in international journals and conferences.



**Maru Tesfaye Addis** is a Ph.D. student in Computer Science and Engineering at Delhi Technological University in India. He holds a Bachelor's degree in Computer Science and Information Technology, which he obtained in 2010 from Arba Minch University in Ethiopia. Additionally, he holds a Master's degree in Computer Science from Bahir Dar University in Ethiopia, which he completed in 2017. Currently, Maru serves as a lecturer in the Department of Computer Science at Debre Tabor University. Maru's research interests revolve around the fields of pattern recognition, deep learning, artificial intelligence, and image processing. He is dedicated to advancing knowledge and contributing to the development of innovative solutions in these areas.

RESEARCH ARTICLE | SEPTEMBER 05 2023

## Energy analysis of the multi-stage vapour compression refrigeration system using eight low GWP refrigerants **FREE**

Manjit Singh ✉; Akhilesh Arora



AIP Conf. Proc. 2863, 020014 (2023)

<https://doi.org/10.1063/5.0155351>



Export  
Citation

CrossMark

### Articles You May Be Interested In

Organic Rankine Cycle design as optimization of dry-steam system geothermal power plant

*AIP Conference Proceedings* (April 2020)

Development of geothermal heat pumps by using environment friendly refrigerants

*J. Renewable Sustainable Energy* (August 2018)

Experimental analysis of a vapour compression refrigeration system by using nano refrigerant (R290/R600a/Al<sub>2</sub>O<sub>3</sub>)

*AIP Conference Proceedings* (July 2019)

500 kHz or 8.5 GHz?  
And all the ranges in between.

Lock-in Amplifiers for your periodic signal measurements



Find out more



# Energy Analysis of the Multi-Stage Vapour Compression Refrigeration System Using Eight Low GWP Refrigerants

Manjit Singh <sup>a)</sup> and Akhilesh Arora <sup>b)</sup>

*Department of Mechanical Engineering, Delhi Technological University, Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India*

<sup>a)</sup> Corresponding author: manjitsingh\_2k20the12@dtu.ac.in

<sup>b)</sup> akhilesharora@dce.ac.in

**Abstract.** This paper aims to optimize a multi-stage vapour compression refrigeration system using flash inter-cooling for different refrigerants. The coefficient of performance (COP) for the system is optimized based on the evaporation temperature, condensation temperature, sub-cooling parameter, and de-superheating parameter. Eight low global warming potential and zero ozone depletion potential refrigerants (R717, R32, R152a, R290, R41, R600a, R134a, and R1234ze(E)) is analysed for optimization at different operating conditions. Modelling of the system is accomplished using EES software. The conjugate direction method, which is generally known as the direct search method, is used to optimize the COP of the system. Research suggests that increasing the sub-cooling parameter increases the system's COP. R717 performs better than other refrigerants with a maximum COP of 6.199, followed by R152a with a maximum COP of 6.155. When the de-superheating parameter increases, the performance of the refrigerants R717, R32, and R152a increases, R1234ze(E) and R600a show a negligible change in COP, and the rest of the refrigerants show adverse effect, i.e., COP decreases with an increase in the de-superheating parameter.

**Keywords:** Energy; Coefficient of Performance; Sub-cooling; De-superheating; Optimization; Flash Chamber

## Nomenclature

### Abbreviation

COP	Coefficient of performance (—)
HFC	Hydrofluorocarbon
VCRS	Vapour-compression refrigeration system
LPC	Low-pressure compressor
ODP	Ozone depletion potential
EV	Expansion valve
GWP	Global warming potential
HPC	High-pressure compressor
NBP	Normal boiling point [°C]

### Symbols

1,2,3 ...	State points
evap	Evaporator
cond	Condenser
el	Electrical

in	Inlet
out	Outlet

### *Subscripts*

T	Temperature [K]
a	Sub-cooling parameter [%]
P	Intermediate pressure [kPa]
h	Specific enthalpy [kJ/Kg]
de	De-superheating parameter [%]
W	Work input [kW]
s	Specific entropy [KJ/Kg K]
m	Mass flow rate [kg/sec]
Q	Heat transfer rate [kW]

## INTRODUCTION

Heating, ventilation, and air conditioning technology has improved in many areas, including temperature manipulation for human comfort and environmental issues. In most industrial applications, vapour compression refrigeration technology is widely used. The evaporator's low-pressure vapour refrigerant is compressed and transported to the high-pressure condenser in a single step inside the compressor. In some applications, because the vapour refrigerant is already at a low temperature, the compressor's desired compression ratio is quite high, decreasing capacity. When the pressure ratio across the compressor exceeds 4 or 5, the compressor's power need rises, lowering the system performance. Multi-staging is preferable to single-stage compression for overcoming this difficulty.

Numerous theoretical and experimental research on two-stage vapour compression refrigeration systems with various configurations has been published to increase the system's overall efficiency. Nasution et al. [1] carried out a numerical investigation of VCRS under the different number of stages using R32 as a working refrigerant and compare the performance of the system with several stages their result shows that an increase in the number of stages leads to an increase in COP of the system. Jatinder and Jagdev [2] tried to find an alternate option for R134a refrigerant due to its high GWP value therefore they conduct an experimental analysis on VCRS and considered R134a/LPG and R134a as the working refrigerants, their result show that LPG/R134a performed better in terms of COP by 15.1-17.82% under various operating conditions and conclude that R134a/LPG can be a better option in place of R134a for long-term. Mosaffa and Farshi [3] introduced phase change materials to absorb the heat from the latent heat thermal energy storage which cools the air and the absorbed heat by the phase change material is then extracted with the help of a refrigeration system. Their results show the COP is more in the case when the cooling load is more and also concluded that the time is taken by the phase change materials to solidify increases with an increase in air inlet temperature.

The refrigerating effect of a system with sub-cooling after refrigerant condensation increases, and the system's performance improves as a result. Torrella et al. [4], based on the concept of increasing refrigerating effect, presented an expression of COP for inter-stage compression systems based on sub-cooling and de-superheating and compared the results of R717 and 404A refrigerants for various configurations of an inter-stage system. Their study shows that R717 performs better than R404A and that the COP of R404A decreases as the de-superheating parameter is increased.

De Paula et al. [5] carried out the 4-E analysis of VCRS using R290, R600a, and R1234yf as working fluids and evaluate all the three costs associated with the modelling of the system and concluded that the cost rate associated with the operation of the system contributes 73% of the total cost of the system, while the penalty cost rate associated with the emission of  $CO_2$  shows only 2.6% of contributions to the total cost and also suggested that R290 can replace R134a due to its better performance. Bahaa et al. [6] analysed single and multi-stage compression systems using lower boiling temperature refrigerants, their result showed that the model used with the sub-cooling system has a good positive impact on the energetic performance of the system and among other refrigerants used, the R22 refrigerant system performed better with maximum COP of 5.49, followed by R134a with a COP of 5.44.

Nikolaidis and Probert [7] performed an exergy-method analysis of a two-stage vapour compression refrigeration to evaluate the plant performance. Esfahani et al. [8] performed thermodynamic and economic analysis on multi-effect evaporation – absorption heat pump incorporated with a vapour compression system. Roy and Mandal [9] performed a full 4-E analysis of a 50kW cascade refrigeration system, a comprehensive analysis was conducted utilising four refrigerant pairs, and the results revealed that the system performed better with R41-R161 and R170-R161 refrigerant pairs in terms of minimum plant cost than with the R41-R404A pair with COP and exergetic efficiency enhancement of 4.9-7.1% and 4.5-6.6%, respectively. Akhilesh et al. [10] presented a comprehensive thermodynamic analysis on VCRS and compared the performance of R22, R407C, and R410A. Their analysis shows that the performance of R22 is better than R407C and R410A by 3-6% at evaporator and condenser temperatures of -38°C and 40°C, respectively. Akhilesh and Kaushik [11] considered HFC22, R410A, and R717 refrigerant and carried out optimum intermediate pressure for each refrigerant separately, and their result show that the optimum intermediate pressure increased by 2% when there is a drop in isentropic efficiency by 10% for HFC22 and R410A, but R717 showed negligible effect on intermediate pressure with change in isentropic efficiency and concluded that R717 perform better than HFC22 and R410A in terms of exergetic efficiency.

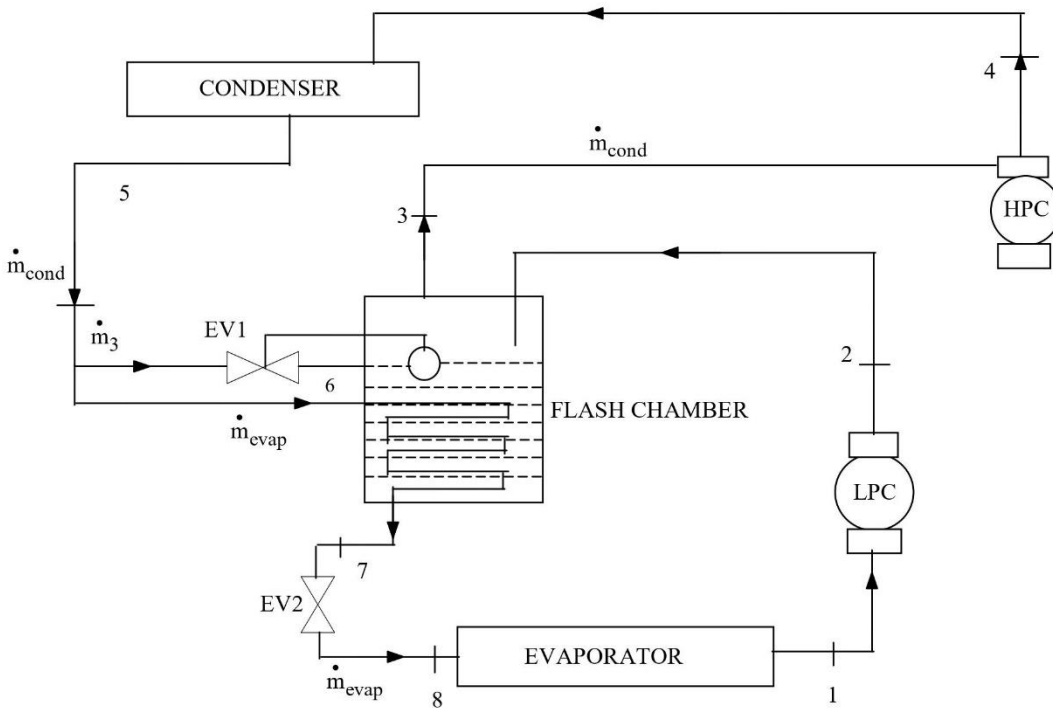
A limited fraction of research work has been published to replace R717 from low-boiling refrigerants in industrial applications, according to the literature review. The purpose of this study is to assess low-boiling-temperature refrigerants with a low GWP value. A multi-stage vapour compression system is modelled using EES software under various operating conditions to identify a replacement for R717 refrigerant.

## BACKGROUND

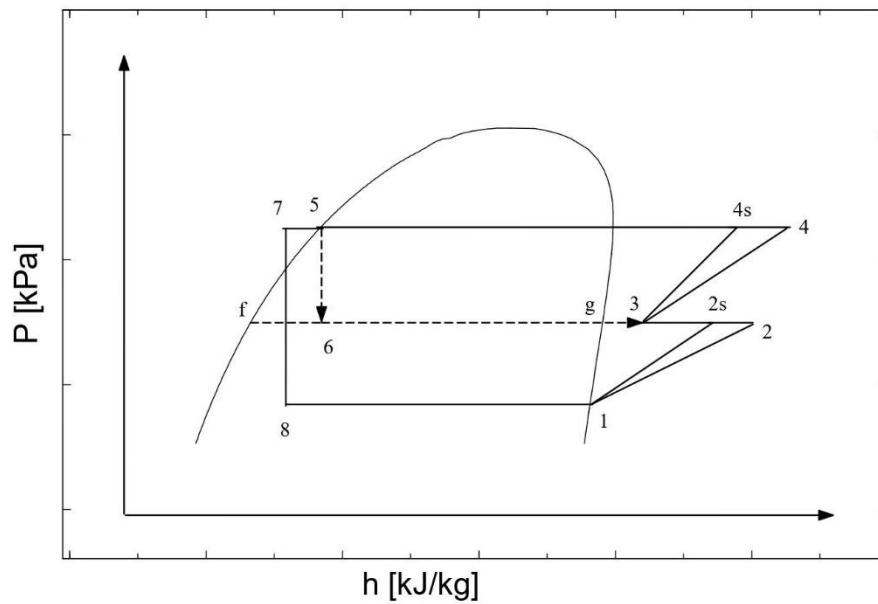
### Multistage Vapour Compression Refrigeration System

The evaporator and condenser pressure becomes more for lower boiling refrigerants due to which the pressure ratio between evaporator and condenser becomes greater than 4 or 5 which is not acceptable because the volumetric efficiency tends to zero and also the work required for the compression increases. To avoid this problem compression can be done in stages. This can be achieved by employing a flash chamber between the compression stages which can also act as a liquid sub-cooler. A schematic diagram of the multi-stage vapour compression refrigeration system with flash inter-cooling and p-h chart of the system is shown in Figures 1 and 2, respectively. In figure 1, the refrigerant at the exit of the evaporator is in a saturated vapour state at 1. The saturated vapour refrigerant then passed through the LPC where saturated vapour gets converted into superheated vapour with an increase in temperature and pressure from evaporator pressure to flash chamber pressure. The de-superheating process takes place after the first stage of compression by the evaporation of a part of liquid refrigerant from the flash chamber at 6.

At state 5 the refrigerant is in a saturated liquid state and then some of its parts are first expanded to the flash chamber pressure through the first expansion valve and the remaining part is then passed through the flash chamber where the sub-cooling process takes place by the evaporation of the liquid refrigerant in the flash chamber and then it passed through the second expansion valve to the evaporator pressure at 8. The entry of the refrigerant in HPC is at 3 due to de-superheating which results in lower compressor work. Similarly, during the sub-cooling process, the entry at the evaporator is at 8 resulting in an increase in refrigerating effect and consequently, an increase in performance.



**FIGURE 1.** System schematic



**FIGURE 2.** The pressure-enthalpy diagram



## Thermodynamic Properties of the Refrigerants Used

The properties of refrigerants that are very important to tell whether the refrigerant is a good substitute for the refrigerant presently used in the refrigeration and air-conditioning industry are minimum normal B.P, evaporator & condenser temperature, GWP, ODP, critical temperature and pressure and others. The N.B.P of the refrigerant should be minimum with freezing point temperature below that of evaporator temperature. From an environmental point of view, the value of ODP and GWP should be minimum. Table 1 consists of thermodynamic properties of refrigerants and at certain operating conditions R134a and R152a show similar properties so, R152a can become an alternate option for R134a due to its low GWP value.

CFC refrigerants is having more value of ODP and GWP and act as greenhouse gases. The best alternate for the CFC refrigerants came with the introduction of HC and HFC refrigerants because of the very less value of ODP and GWP. Table 2 consists of environmental and physical properties of the refrigerants used in this study where ammonia and R1234ze(E) is having zero value of GWP and ODP but the main concern raised with ammonia is its safety class, because of its slightly flammable behaviour it comes under the safety class of B2. The advantage of using R600a and R290 is that they have very less value of GWP and zero ODP value but they come under the category of A3 because of their flammable nature.

**TABLE 1.** Refrigerant thermodynamic properties. Prepared by author and referred [12]

Refrigerant	Molecular Weight	$t_s$ (N.B.P) °C	$t_c$ (Critical-Temperature) °C	$P_c$ (Critical-Pressure) bar	$t_f$ (Freezing point) °C
R717	17.031	-33.35	133.0	112.97	-77.7
R32	52.024	-51.75	78.41	58.3	-136.0
R134a	102.03	-26.15	101.06	40.56	-96.6
R600a	58.13	-11.73	135.0	36.45	-159.6
R290	44.1	-42.1	96.8	42.56	-187.1
R1234ze(E)	114	-18.95	109.4	36.32	-156.0
R152a	66.05	-24.15	113.3	45.2	-117.0
R41	34	-78.2	44.5	58.97	-142.0

**TABLE 2.** Physical and environmental properties. Prepared by author and referred [12]

Refrigerant	Type	GWP	ODP	Safety class	Flammability
R717	Natural	0	0	B2	Slightly-flammable
R32	HFC	675	0	A2	Low-flammability
R134a	HFC	1430	0	A1	Non-flammable
R600a	HC	3	0	A3	Flammable
R290	HC	3	0	A3	Flammable
R1234ze(E)	HFO	0	0	A2	Non-flammable
R152a	HFC	140	0	A2	Slightly-flammable
R41	HFC	92	0	B3	Flammable

## THERMODYNAMIC ANALYSIS

To create a thermodynamic model, the first and second laws of thermodynamics are extremely important. The mass, energy, and exergy equations are obtained for a multi-stage vapour compression refrigeration system.

While solving the equations, the following assumptions are considered:

1. The pressure drop inside the system components is considered constant with no losses.
2. Changes in kinetic and potential energies are neglected.
3. Constant enthalpy throughout the expansion process.
4. The refrigerant is in saturated condition at the evaporator and condenser outlets.

The following equations are used to model the system after taking the above assumptions:

$$\text{Material balance} \quad \sum \dot{m}_{in} - \dot{m}_{out} = 0 \quad (1)$$

$$\text{Energy balance} \quad \sum Q - \sum W - \sum \dot{m}h = 0 \quad (2)$$

$$\text{Exergy balance} \quad \sum_{in} X - \sum_{out} X - \sum_{heat} X - \sum_{work} X - X_{destroyed} = 0 \quad (3)$$

The EES software facilitates the calculation of refrigerant thermodynamic parameters such as enthalpy, entropy, thermal conductivity, saturation temperature, etc. It is comprised of both pure and blended substances, making EES suitable for energy/exergy analysis and optimization of any refrigerant in its database. As a result, the system is modelled to study the effect of thermodynamic laws.

## Energy Analysis

**TABLE 3.** Energy equations used in system modelling

Component/Parameter	Energy balance
Evaporator	$\dot{m}_{evap} = \frac{\dot{Q}_{evap}}{(h_{out} - h_{in})_{evap}} \quad (4)$
Compressor	$W_{el,LPC} = \frac{W_{LPC}}{\eta_{LPC}} \quad (5)$
	$W_{el,HPC} = \frac{W_{HPC}}{\eta_{HPC}} \quad (6)$
	$\eta_{comp} = 0.85 - 0.046667PR_{comp} \quad (7)$
Condenser	$\dot{Q}_{cond} = \dot{m}_{cond}(h_{out} - h_{in}) \quad (8)$
Expansion Valve	$h_{in,ev} - h_{out,ev} \quad (9)$
Flash Chamber	$P_{OPT,int} = \sqrt{\frac{P_{evap}P_{cond}T_{cond}}{T_{evap}}} \quad (10)$
	$\dot{m}_{cond} = \dot{m}_{evap} + \dot{m}_{fc} \quad (11)$
	$\dot{m}_{cond}h_5 + \dot{m}_{evap}h_2 = \dot{m}_{cond}h_3 + \dot{m}_{evap}h_7 \quad (12)$

Torrella et al. [4] defined three parameters for an inter-stage configuration:

$$\text{Sub-cooling parameter} \quad a = \frac{h_5 - h_7}{h_5 - h_f} \quad (13)$$

$$\text{De-superheating parameter} \quad de = \frac{h_2 - h_3}{h_2 - h_g} \quad (14)$$

$$\text{Mass ratio parameter} \quad r = \frac{\dot{m}_{cond}}{\dot{m}_{evap}} = \frac{(h_2 - h_5) + a(h_5 - h_f)}{(h_2 - h_5) - b(h_2 - h_g)} \quad (15)$$

Finally, the performance of the system is measured using the equation:

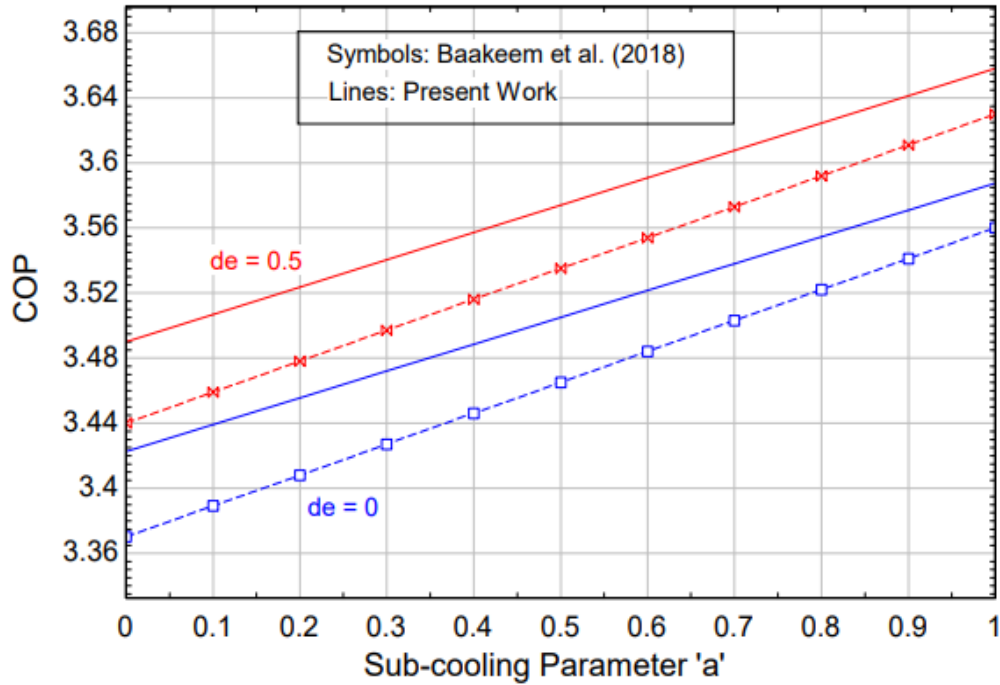
$$COP = \frac{\dot{Q}_{Evap}}{W_{el,LPC} + W_{el,HPC}} \quad (16)$$

### Model Validation

The present model is validated with the model developed by Baakeem et al. [12] in which they theoretically investigate the multi-stage compression system and optimized the model with the help of EES software. They considered eight refrigerants in their study to optimize the model taking four parameters into account. Those four parameters were  $T_e$ ,  $T_c$ ,  $a$ , and  $de$ . The direct search method is used for the optimization of the model to get maximum COP at evaporation and condensation temperature of 10°C and 40°C respectively. Table 3 presents the results obtained in the present study and reference study. There is a 0.47% difference in maximum COP of ammonia whereas R1234ze(E) refrigerant system shows the maximum difference in COP of 1.13%.

**TABLE 4.** Model validation of multi-stage compression system after optimization

Refrigerant	Referred work (Baakeem et al. 2018)	Present work	Difference (%)
	$COP_{max}$	$COP_{max}$	$COP_{max}$
R717	6.17	6.199	+0.47
R134a	6.01	6.048	+0.63
R1234ze(E)	6.01	6.078	+1.13



**FIGURE 3.** Model validation for ammonia multi-stage compression system

## Input Parameters

The thermodynamic analysis of a multi-stage compression system is carried out in the present work using EES software. Baakeem et al. [12] assumed the following input parameters listed in the table before performing operations.

**TABLE 5.** Input parameters for system modelling

Parameters	Value
Evaporation Temperature, $T_{evap}$ (°C)	0
Condensation Temperature, $T_{cond}$ (°C)	45.0
Sub-cooler Efficiency, (%)	80
LPC Efficiency, (%)	91
Cooling Load, (kW)	1

## RESULTS AND DISCUSSION

### Model Optimization

To converge the findings, EES software employs several optimization approaches. Golden search, conjugate directions, variable metric optimization, genetic approach, and Nelder-mead simplex method are some of them. When there is only one degree of freedom, the golden search method is employed to discover the minimum or maximum. Because there is more than one variable in this study, the conjugate direction approach is employed to maximise the system's performance (COP). In EES, the conjugate direction approach is also known as the direct search method, because it searches for an optimum value of,  $X_1$  while keeping,  $X_2$ ,  $X_3$ ,  $X_4$ , and so on constant. The technique is then repeated to calculate the other dimensions while maintaining one constant. To maximise the performance of the multi-stage system, four independent variables are taken into account:  $T_e$ ,  $T_c$ ,  $a$ , and  $de$ .

The system is optimised by keeping the evaporator's lower and upper temperatures at -20 and 10°C, respectively. The temperature limit for the condenser is fixed between 40 and 60°C. During the optimization, the "a" and "de" parameters vary between 0 and 1.

**TABLE 6.** Optimization Results

Refrigerant	Optimum Conditions				$COP_{max}$	$W_{el}(kW)$
	$T_{evap}$ (°C)	$T_{cond}$ (°C)	a	de		
R717	10	40	1	1	6.199	0.1613
R134a	10	40	1	1	6.048	0.1653
R32	10	40	1	1	5.871	0.1703
R1234ze(E)	10	40	1	0	6.078	0.1645
R41	10	40	1	1	4.602	0.2173
R152a	10	40	1	1	6.155	0.1625
R600a	10	40	1	1	6.123	0.1633
R290	10	40	1	1	5.993	0.1671

### *The influence of "a" and "de" parameters on the COP*

Figure 3-(a) shows that for refrigerants R717, as the "de" parameter increases with an increase in parameter "a" the performance increases for the R717 refrigerant system. At constant sub-cooler efficiency of 80% as the "de" parameter increases from 0-0.5 and 0.5-1, the increase in performance is 1.99% and 2.26%, respectively. After optimization, the COP increases by 70.959% for the R717 refrigerant system.

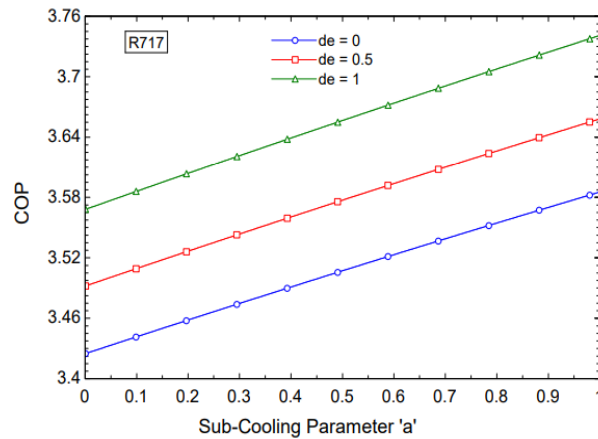
For the R32 refrigerant system, figure 3-(b) shows that the system's performance is increasing with an increase in “a” and “de” parameters. As the “de” parameter changes from 0-0.5 and 0.5-1, there is a 1.03% and 1.223% increase in the performance of the system, respectively.

It can be seen from figure 3-(c) that for refrigerant R152a, there is a slight increment of 0.109% and 0.1365% in performance of the system as “de” parameter changes from 0-0.5 and 0.5-1 respectively, at constant sub-cooler efficiency.

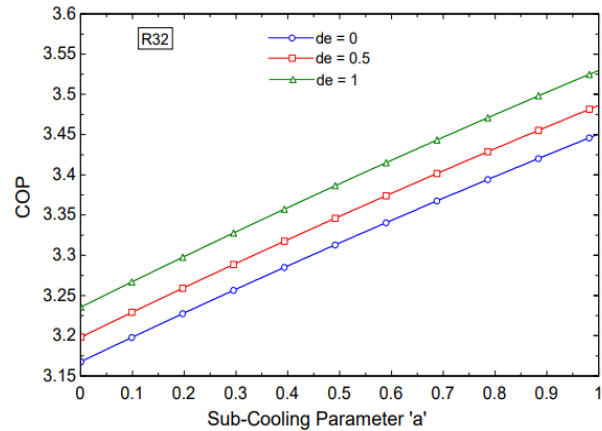
The multi-stage system shows opposite behaviour in the case of R134a, R41, and R290 refrigerants. As the de-superheating parameter increases from 0-1, the system's performance tends to decrease with 0.056% decrement for the R134a refrigerant system, 1.836-2.577% decrement in the case of the R41 refrigerant system, and 0.845% of decrement in the case of R290 refrigerant system.

When a multi-stage compression system is operating with R1234ze and R600a refrigerants, the performance of the multi-stage system is independent of the change in the de-superheating parameter and increases with an increase in the sub-cooling parameter. At constant sub-cooler efficiency, the performance of the multi-stage system is 3.609 and 3.565 for R1234ze and R600a, respectively.

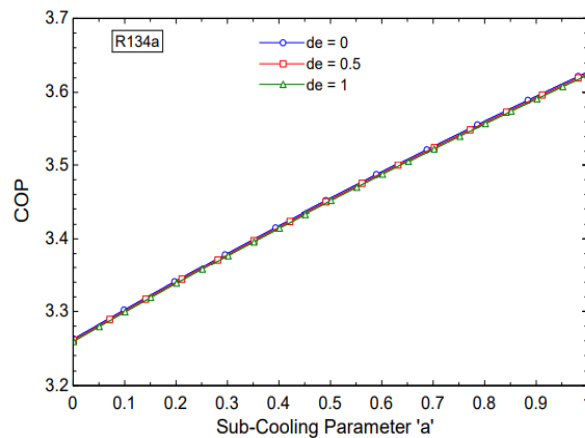
In the case of the R152a refrigerant system, the change in performance is very significant with an increase in the de-superheating parameter. Fig shows that only 0.109% and 0.1365% increment in COP when the de-superheating parameter changes between 0-0.5 and 0.5-1, respectively.



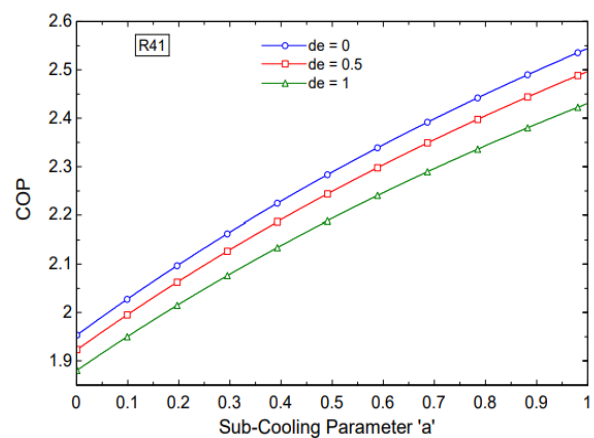
a) R717 COP Variation



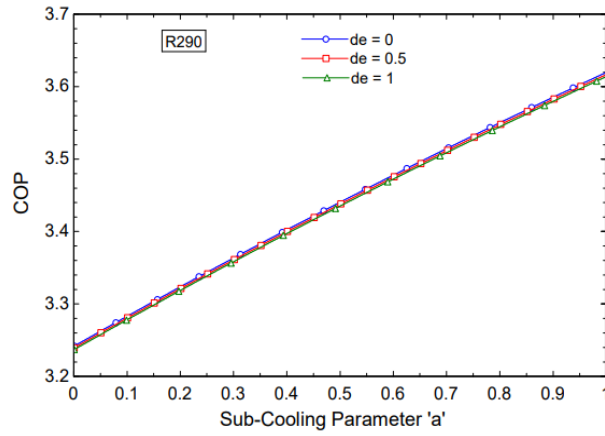
b) R32 COP Variation



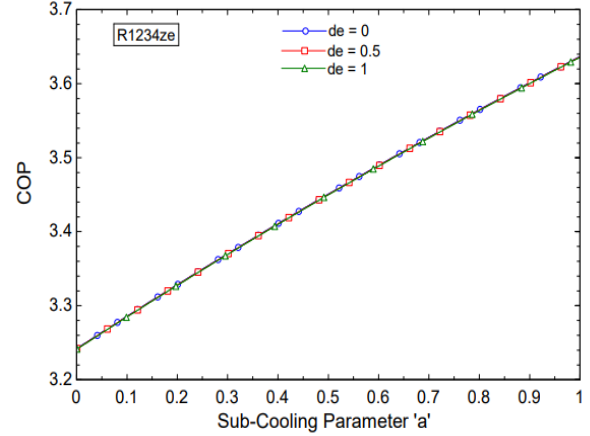
c) R134a COP Variation



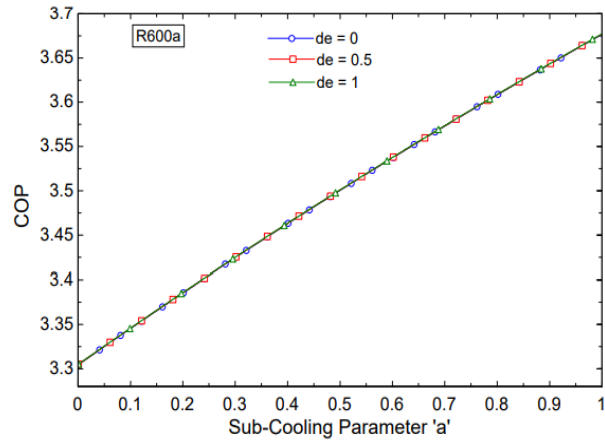
d) R41 COP Variation



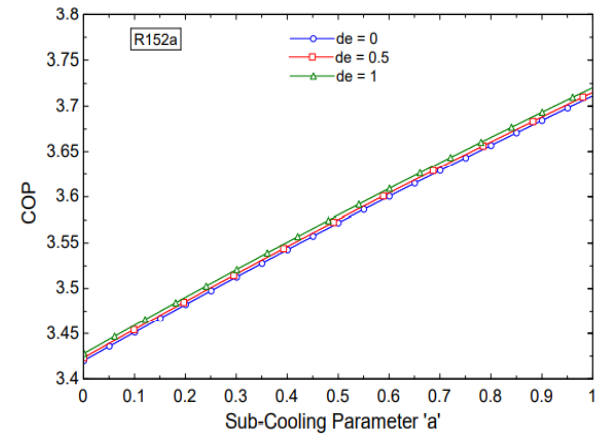
e) R290 COP Variation



f) R1234ze COP Variation



g) R600a COP Variation



h) R152a COP Variation

**FIGURE 4.** The influence of “a” and “de” parameters on the COP

## CONCLUSIONS

In the theoretical analysis of the multi-stage VCRS, the R717, R32, R290, R1234ze(E), R152a, R41, R134a, and R600a refrigerants are considered to investigate the energy analysis of the system using EES software. Based on the four parameters, i.e.,  $T_{evap}$ ,  $T_{cond}$ ,  $a$ , and  $de$ , the COP of the system is optimized.

The following conclusions can be drawn from the findings:

- R717 exceeds the other seven refrigerants in terms of COP. R717 shows a maximum COP of 3.626 before optimization and 6.199 after optimization, while the R41 refrigerant system shows a minimum COP of 2.405 before optimization and 4.602 after optimization.
- With an increase in the sub-cooling parameter the performance of the system increases. The COP of the system increases as the de-superheating parameter is increased for R717, R32, and R152a refrigerants.
- Among eight refrigerants, R134a, R290, and R41 show an adverse effect on COP when the “de” parameter is increased. The effect of increasing the “de” parameter on R1234ze and R600a refrigerants is negligible.



- After optimization, the maximum COP came out for R717 refrigerant with an increment of 70.96%, followed by R152a with an increment of 68.12% in COP.
- At sub-cooler efficiency, evaporation temperature, and condensation temperature, R41 refrigerant shows less value of COP as compared to the other seven refrigerants.

From the above conclusions, R717 performs better than the other seven refrigerants. R152a can be an alternate option for R717, and R41 refrigerant is not preferred to use because of its less performance and high flammability.

## REFERENCES

- [1] A. H. Nasution, H. Ambarita, H. V. Sihombing, E. Y. Setiawan and H. Kawai, "The effect of stage number on the performance of a vapor compression refrigeration cycle using refrigerant R32," in *IOP Conference Series: Materials Science and Engineering*, 2020.
- [2] J. Gill and J. Singh, "Energy analysis of vapor compression refrigeration system using mixture of R134a and LPG as refrigerant," *International Journal of Refrigeration*, 2017.
- [3] A. H. Mosaffa and L. G. Farshi, "Exergoeconomic and environmental analyses of an air conditioning system using thermal energy storage," *Applied Energy*, vol. 162, pp. 515-526, 2016.
- [4] E. Torrella, J. A. Larumbe, R. Cabello, R. Llopis and D. Sanchez, "A general methodology for energy comparison of intermediate configurations in two-stage vapour compression refrigeration systems," *Energy*, vol. 36, no. 7, pp. 4119-4124, 2011.
- [5] C. H. de Paula, W. M. Duarte, T. T. M. Rocha, R. N. de Oliveira, R. d. P. Mendes and A. A. Torres Maia, "Thermo-economic and environmental analysis of a small capacity vapor compression refrigeration system using R290, R1234yf, and R600a," *International Journal of Refrigeration*, vol. 118, pp. 250-260, 2020.
- [6] B. Morad, M. Gadalla and S. Ahmed, "Energetic and exergetic comparative analysis of advanced vapour compression cycles for cooling applications using alternative refrigerants," *International Journal of Exergy*, vol. 26, no. 1-2, pp. 226-246, 2018.
- [7] C. Nikolaidis and D. Probert, "Exergy-method analysis of a two-stage vapour-compression refrigeration-plants performance," *Applied Energy*, vol. 60, no. 4, pp. 241-256, 1998.
- [8] I. J. Esfahani, Y. T. Kang and C. Yoo, "A high efficient combined multi-effect evaporation-absorption heat pump and vapor-compression refrigeration part 1: Energy and economic modeling and analysis," *Energy*, vol. 75, pp. 312-326, 2014.
- [9] R. Roy and B. K. Mandal, "Thermo-economic analysis and multi-objective optimization of vapour cascade refrigeration system using different refrigerant combinations," *Journal of Thermal Analysis and Calorimetry*, vol. 139, no. 5, pp. 3247-3261, 2019.
- [10] A. Arora, B. B. Arora, B. D. Pathak and H. L. Sachdev, "Exergy analysis of a Vapour Compression Refrigeration system with R-22, R-407C and R-410A," *International Journal of Exergy*, vol. 4, no. 4, pp. 441-454, 2007.
- [11] A. Arora and S. Kaushik, "Energy and exergy analyses of a two-stage vapour compression refrigeration system," *International Journal of Energy Research*, vol. 34, pp. 907-923, 2010.
- [12] S. S. Baakeem, J. Orfi and A. Alabdulkarem, "Optimization of a multistage vapor-compression refrigeration system for various refrigerants," *Applied Thermal Engineering*, vol. 136, pp. 84-96, 2018.

# Enhancement of Power Quality of Three-Phase GC Solar Photovoltaics

Sukhbir Singh (✉ [sukhbirsinghdtu26051972@gmail.com](mailto:sukhbirsinghdtu26051972@gmail.com))

Delhi Technological University

J N Rai

Delhi Technological University

---

## Research Article

**Keywords:** Three-phase grid connected, PV, inter-harmonics, DC offset, EDRL, COA-Fuzzified PLL

**Posted Date:** August 31st, 2023

**DOI:** <https://doi.org/10.21203/rs.3.rs-3280839/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Enhancement of Power Quality of Three-Phase GC Solar Photovoltaics

<sup>1</sup>Sukhbir Singh

Department of Electrical Engineering, Delhi Technological University, New Delhi, India  
Email: sss.singh149@gmail.com

<sup>2</sup>J N Rai

Department of Electrical Engineering, Delhi Technological University, New Delhi, India  
Email: Jnraiphd1968@gmail.com

## Abstract:

The proliferation of grid-connected photovoltaic (PV) systems has generated considerable apprehension among power system operators due to worries about electricity quality, leading to the implementation of increasingly strict standards and regulations. Inter-harmonics and DC offset have emerged as prominent power quality issues in grid-connected photovoltaic (PV) systems, constituting significant obstacles. This article provides a thorough examination of the methods used to improve the performance of a three-phase grid-connected photovoltaic (PV) system, with a specific focus on mitigating inter-harmonics and DC offset. The presence of inter-harmonics and DC offset may have a substantial negative impact on the overall performance of a system, resulting in compromised power quality and diminished energy extraction capabilities. In order to address these challenges, a method known as ensembled Deep Reinforcement learning (EDRL) Maximum electricity Point Tracking (MPPT) is used to optimize the extraction of electricity from the photovoltaic (PV) array. Furthermore, the integration of a Coati Optimization Algorithm (COA) with a fuzzified Phase-Locked Loop (PLL) synchronization mechanism is used to ensure precise synchronization with the grid. The EDRL MPPT approach demonstrates a proficient ability to accurately monitor and follow the maximum power point of the photovoltaic (PV) array. This is achieved by using a reward system that is based on the lowest overall harmonic distortion in the grid current. The COA (Centralized Optimization Algorithm) is used to effectively tune the hyperparameters of the fuzzy system. The primary objective of this optimization process is to reduce the DC offset, hence ensuring a steady and precise synchronization between the fuzzy system and the grid. The efficacy of the proposed system is assessed by means of comprehensive simulations and experimental validation. The findings of this study provide evidence supporting the efficacy of the Enhanced Distributed Reactive Load Maximum Power Point Tracking (EDRL MPPT) approach in optimizing power extraction and reducing the impact of inter-harmonics. The COA-fuzzified-PLL synchronization system is designed to provide precise grid synchronization while mitigating the adverse effects of a 2.89% total harmonic distortion (THD) in grid current, particularly the influence of direct current (DC) offset. The integration of many approaches presents notable improvements in terms of power quality, energy extraction efficiency, and system stability.

**Keywords:** Three-phase grid connected, PV, inter-harmonics, DC offset, EDRL, COA-Fuzzified PLL,

## Abbreviations

MPPT	Maximum Power Point Tracking
EDRL	ensembled Deep Reinforcement learning
COA	Coati Optimization Algorithm
PV	photovoltaic
PLL	Phase-Locked Loop
DC	Direct Current
SRF	synchronous reference frame
SOGI	second-order generalized integrator
ROGI	Reduced-Order Generalized Integrator
TOGI	Third Order Generalised Integrator
P & O	perturb and observe
THD	Total Harmonic Distortion
PWM	Pulse width modulation
DQN	Distinct neural network
DDPG	Deep Deterministic Policy Gradient

rlTD3	twin-delayed deep deterministic policy gradient
PPO	Proximal Policy Optimization
LCL	Inductor–Capacitor–Inductor filter

## Introduction

The escalating global demand for sustainable and clean energy sources has led to the emergence of solar photovoltaic (PV) systems as a viable solution to address the challenges of environmental pollution and the depletion of fossil fuel reserves. The deployment of solar photovoltaic (PV) systems that are connected to the grid has attracted significant attention due to their ability to efficiently harness abundant solar energy and seamlessly incorporate it into the power grid. The grid-connected solar PV system is a promising form of a nonconventional energy resource that generates electricity without emitting carbon, thereby contributing to a cleaner environment. Nevertheless, the intermittent nature of the solar PV system precludes its direct connection to the utility grid, necessitating asynchronous coupling through the utilization of converter devices.

The utilization of a grid-tied AC/DC converter, coupled with an appropriate control methodology, facilitates the fulfilment of grid connection prerequisites for solar photovoltaic systems. These prerequisites encompass electrical power flow management, harmonic mitigation, enhanced quality of power, stability, and grid harmonization [1]. The employment of photovoltaic (PV) arrays in the design of grid systems is widely adopted and acknowledged as the predominant method. The control problem of transmitting maximal electrical energy accessible to the load regardless of the situation is commonly referred to as MPPT. The primary objective of Maximum Power Point Tracking (MPPT) is to effectively regulate the fluctuations in output voltage that arise due to variations in PV power. Furthermore, photovoltaic systems demonstrate a non-linear correlation between output current along with voltage, resulting in substantial efficiency reductions. [2]

Typically, the interface circuit for the grid is expected to carry out three primary functions, namely voltage sensing, filtering, and A/D conversion. The occurrence of direct current (DC) offset in the measured grid voltage can be ascribed to the non-linear characteristics of voltage sensors, the analog-to-digital (A/D) conversion procedure, and the thermal drift exhibited by analog components. This phenomenon may occur despite the implementation of a well-designed grid interface circuit. [3] The input sine signal's undesirable component on the resultant waveform of the PLL structure is the addressed DC offset. The reference sine signal is commonly employed in the creation of reference currents for photovoltaic or grid-tied converters. Several established standards, including IEEE 1547-200, EN61000-3-2, and IEC 61727, delineate the permissible limit of direct current injection into the grid that may be attributed to photovoltaic systems or other converters that are connected to the grid. Various techniques can be employed to eliminate the induced DC offset in the measured grid voltage. The existence of inter-harmonics within the power system has a detrimental impact on its overall performance. The connection of a significant quantity of non-linear loads to the power system results in a range of complications, including inadequate power factor, excessive heating of transformers and power cables, impaired functioning of protection devices, heightened transmission losses, and substandard voltage regulation [3]. Inter-harmonic components result in significant energy wastage. The escalation of inter-harmonics within the power system can be closely associated with non-linear power electronic loads, such as variable frequency drives, converters, and inverters. Environmental conditions, such as temperature variations and irradiance fluctuations, can affect the performance of PV modules and inverters. These variations can introduce oscillation in the output power and current harmonics, including inter-harmonics.

As a result of the previously stated DC offset in voltage and current, the waveforms exhibit asymmetry along the x-axis. Consequently, the process of interrupting asymmetrical current is considerably more challenging than interrupting symmetrical current, as per reference [3]. Moreover, as a result of the direct current offset, there is a possibility of imbalanced grid voltages. This phenomenon has the potential to result in the degradation of power quality issues. Therefore, it can be concluded that harmonics and DC offset are unnecessary in a functional power system.

### 1.1 Problem Statement and Contributions

The efficiency of solar PV systems that are associated with the grid is subject to the impact of multiple factors, including fluctuations in solar irradiance, environmental circumstances, and distortions in grid voltage due to non-linear loads and conversion processes. These distortions include inter-harmonics and DC offset [2]. Several causes can account for the fact that some observed systems' output currents have unwelcome direct current components. Non-linearities in switching devices, small errors and offset drifts in voltage and current measurement sensors

used to provide feedback signals for control systems are all examples of sources of imprecision and error in PWM signals [14].

In the context of PV inverters, it is possible that inter-harmonics are attributable to the MPPT control mechanism. Partly shedding conditions lead to irradiance fluctuations that can affect the performance of PV modules and inverters. These variations can introduce fluctuations in the output power and current harmonics, including inter-harmonics. The conventional techniques employed for achieving maximum power point tracking (MPPT) and grid synchronization are constrained by issues related to inter-harmonics. Additionally, traditional synchronization methods, such as Phase Locked Loop (PLL), may struggle to maintain accurate synchronization with the grid in the presence of voltage distortions or frequency variations, results DC offset problems.

These challenges are addressed in this work, and contributions are:

- To address these challenges, this research aims to improve the power quality of grid-tied solar PV systems by proposing the utilization of a novel ensembled Deep Reinforcement Learning (EDRL) MPPT controller, and a Coati Optimization Algorithm tuned Fuzzified-Phase Locked Loop (COA Fuzzified-PLL) based synchronizing system.
- The ensembled DRL MPPT controller leverages the power of deep learning and reinforcement learning techniques to optimize the MPPT process, enabling the mitigation of inter-harmonics and efficient power extraction with a constant reference DC voltage at the DC link.
- The objective of the synchronizing system based on COA Fuzzified-PLL is to achieve reliable synchronization with the electrical grid, even under voltage distortions and frequency variations, while also being able to reject DC-offset.

## 1.2. Article Structure

The article begins with an introduction that highlights the significance of power quality enhancement in grid-tied PV systems and outlines the existing challenges. A comprehensive literature review is provided, examining the limitations of current control strategies and exploring previous research on adaptive control techniques. The system configuration and proposed controller of the grid-tied PV system with Ensembled DRL are presented in section III. The proposed COA-Fuzzified-PLL control strategy, including the Fuzzy membership function, is then described. The article proceeds to present experimental results to validate the effectiveness of the proposed strategy. A discussion section provides an in-depth evaluation of the approach, its advantages, limitations, and implications for power quality enhancement. Finally, the paper concludes with a summary of the key findings, contributions, and potential impact of the research, along with directions for future work.

## Related Work

In current years, considerable research and development has focused on the power quality improvement of grid-connected photovoltaic scheme regarding harmonics and DC offset. Several research studies have examined the creation and impact of inter-harmonics, underscoring the necessity for efficient reduction methods. Moreover, direct current offset in photovoltaic systems results in transformer saturation and elevated losses. The scholarly community has shown interest in identifying and eliminating DC offset. Various methods for mitigating these issues have been suggested to alleviate the issues of harmonics and DC offset, such as grid synchronization, phase-locked loop (PLL) [3], synchronous reference frame (SRF-PLL) [4], and non-PLL techniques such as SOGI [9], Cascaded SOGI [10] within the context of this document. Thus, the attainment of accurate estimation and the elimination of unwanted periodic ripple in SRF-PLL [4] can be accomplished in the presence of a DC offset in the input signal. The paper [4] presents a novel phase-locked loop (PLL) that utilizes the Extended Second-Order Generalized Integrator (ESOGI) and is reconfigured by the Second-Order Generalized Integrator (SOGI) with Active Power Filter (APF). The study also suggests that the ESOGI-PLL exhibits satisfactory performance, even under conditions of elevated dc offset values and a high frequency of low-order harmonics. Furthermore, due to its uncomplicated architecture, the ESOGI-PLL suggested in this study can be readily executed on an inexpensive microcontroller. According to source [5], the second-order generalized integrator, which corresponds to the SOGI-PLL, exhibits a relatively rapid transient response, a high capacity for rejecting disturbances, and a robust performance. One of the previously published works has suggested the utilization of SOGI-FLL [6] to improve the efficacy of SOGI-PLL in the context of frequency fluctuations. This is because the performance of SOGI-PLL is deemed inadequate in scenarios with variations in frequency and DC offset in the grid voltage. The article [1] presents a proposal for a comb filter utilizing a modified sliding Goertzel discrete Fourier transform (SGDFT)-based phase-locked loop (PLL). The design incorporates the three degrees of freedom (DOFs) of second-order

fraction delay, to mitigate the effects of non-integer frequency components. The SGDF-based PLL [1] operates at a constant sampling frequency and aims to calculate the fundamental frequency, amplitude, and phase angle in order to achieve optimal synchronization. The paper [7], utilizes the Advanced Third Order Generalised Integrator (ATOGI) for controlling a two-stage three-phase photovoltaic system power quality. The ATOGI [7] is utilized for the purpose of extracting essential components from distorted grid voltages and non-linear load current. The proposed method effectively addresses the integrator delay and inter-harmonic challenges inherent in conventional SOGI techniques. Additionally, the DC-Offset Estimator component exhibits robust dc-offset rejection capabilities. The study [8] showcases a noteworthy implementation of two distinct generalized integrators: SOGI and the Reduced-Order Generalized Integrator (ROGI) controller, which are interconnected in a cascade configuration. The outcome of SOGI-ROGI yields several desirable characteristics, including the absence of phase shift, optimal filtering, minimal harmonic distortion, and favourable dynamic response. Furthermore, filtering techniques, such as passive filters and active power filters, have been explored to suppress these issues. Despite progress, challenges remain, including the need for improved detection methods and the integration of emerging technologies.

As discussed in the introduction section, the voltage disturbance of photovoltaic arrays results in power oscillations, particularly during partial shedding operation. These power oscillations contain inter harmonics. Maximum Power Point Tracking (MPPT) control techniques have been suggested to harness maximum output and increase power quality over the years. Several approaches have been proposed in the literature for maximum power point tracking (MPPT) in photovoltaic (PV) power systems. These include a modified incremental conduction MPPT algorithm with a fuzzy controller [11], a hill-climbing (HC) modified fuzzy-logic (FL) MPPT control scheme implemented in both software and hardware [12], and a hybrid MPPT control strategy that integrates a modified perturb and observe (P&O) algorithm with an enhanced particle swarm optimization (PSO) algorithm [13]. Numerous heuristic techniques have been employed in the field, including the innovative Maximum Power Point Tracking (MPPT) method, which is adaptable to applications with rapid fluctuations through the utilization of Artificial Neural Network (ANN) [14]. Most of these techniques are based on models and aim to regulate various photovoltaic (PV) systems but not predict the inter-harmonics characteristics. The acquisition of an accurate model for photovoltaic (PV) systems and their associated parameters can remedy these harmonic challenges that are associated with PV panels in various configurations. The author is compelled to seek a methodology that is independent of any specific model.

While several research studies have examined specific elements of inter-harmonics and DC offset in grid-connected solar PV systems, there is a limited availability of sophisticated approaches and comprehensive methodologies that effectively tackle both of these concerns concurrently. The research gap pertains to the lack of comprehensive solutions that can adequately address the suppression and mitigation of inter-harmonics and DC offset, taking into account their potential interactions and cumulative impact on power quality. The simultaneous management of these phenomena will greatly improve the efficiency, reliability, and performance of grid-connected solar photovoltaic (PV) systems.

## **System topology and control architecture**

### **3.1 System Topology**

The system is comprised of three primary components, namely photovoltaic (PV) panels or arrays, PV inverters, and the alternating current (ac) grid. PV inverters, such as the full-bridge inverter (IGBT), are integral in regulating the transmission of power from PV arrays to the AC grid. The Ensembled DRL maximum power point (MPP) based on total harmonic distortion (THD) Reward technique is employed for the purpose of producing a PWM signal that is utilized in a boost converter, with the aim of enhancing the voltage profile of the photovoltaic (PV) system and mitigating the inter-harmonic component. Figure 1 illustrates the conventional control configuration of 3-phase grid-connected photovoltaic (PV) inverters, A photovoltaic array with a power output of 250 kilowatts is linked to a 25 kV electrical grid through a three-phase converter. The photovoltaic (PV) array is composed of 88 parallel strings. A series connection of 7 SunPower SPR-415E modules is present in each string. Table 1 provides the relevant system parameters. To achieve optimal power extraction from photovoltaic arrays, the implementation of a maximum power point tracking (MPPT) algorithm is necessary.

Table 1 The parameter setting of PV





function parameters are efficiently and accurately tuned using COA optimization. COA-fuzzified PLL is an efficient approach to identifying and eliminating DC offset in the solar PV system connected to the grid. This technique guarantees precise synchronization with the utility grid and reduces the likelihood of power quality problems. The proposed methodology integrates the ensembled DRL-based MPPT and COA-fuzzified PLL techniques to enhance the overall performance of the 3-phase grid-connected solar PV system. The ensembled DRL optimizes the MPPT process, while the COA-fuzzified PLL eliminates DC offset and ensures precise synchronization with the utility grid. By effectively addressing inter-harmonics and DC offset, the methodology aims to improve power quality, maximize energy extraction, and enhance the overall performance of grid-connected solar PV systems.

### 3.3 Ensembled DRL-MPPT Approach for Inter-harmonic Mitigation

The present study introduces a novel approach that involves incorporating a THD reward within the framework of ensembled deep reinforcement learning (EDRL) to address the issue of maximum power point tracking (MPPT) in photovoltaic (PV) arrays. Ensembled Deep Reinforcement learning (EDRL) provides an effective solution to this problem without requiring any parametric information regarding the dynamic parameters of the model. The objective of the algorithm is to illustrate the system analogy utilizing the DQN (discrete), DDPG (continuous), rITD3 (continuous), and PPO (discrete) network framework. The ensembled DRL approach can be shown in Figure 2.

The DRL consists of four primary components. The three fundamental components of DRL are commonly referred to as the state space  $X$ , reward function  $r$ , and action space  $U$ . The primary principle of Maximum Power Point Tracking (MPPT) is to optimize the power extraction from photovoltaic (PV) modules by ensuring their operation at the voltage corresponding to the maximum power point, thereby maximizing the available power output. The acquisition of knowledge by the agent is facilitated by any form of engagement with the environment. This process involves the execution of an action  $u_t \in U$ , which triggers the evolution of the system from its current state  $x_t \in X$  to the subsequent state  $x_{t+1}$ . The agent obtains feedback in the form of a THD reward, which serves as a quantitative measure of the efficacy of the action or decision made by the agent. Consequently, the incentive serves as an indication to pinpoint the attainable objective or ideal resolution. The RL approach aims to identify an optimal policy  $\pi$  that meets a given set of criteria.

$$J^* = \max_{\pi} J_{\pi} = \max_{\pi} E_{\pi} \{r_t | x_t = x\} \quad (1)$$

The symbol  $J_{\pi}$  denotes the cumulative expected reward under a given policy  $\pi$ . Assuming a policy  $\pi$  that absolves, the value function for a given time interval, denoted as  $V^{\pi}(x)$ , or the expected cumulative reward, is a function of  $x^{\pi}$  and is defined as  $x^{\pi} = \{x_t\}_{t=1}^{t=n}$ , where the state values are represented by  $k^{\pi} = \{k_t\}_{t=1}^{t=n}$ , which are sequences of actions taken by the agent.

#### State Space

The state-space design in Maximum Power Point Tracking (MPPT) problems involves analysing the movement of the Maximum Power Point (MPP) on the Photovoltaic (PV) curve across varying environmental conditions. The methodology employed in reinforcement learning is regulated by the photovoltaic current, power, and the direct current voltage of the coupling. The state-space comprises  $X$ , which is a vector containing the variables  $X \in [V_{PV}, I_{PV}, P_{PV}, \Delta P_{PV}, \int \Delta P_{PV}, \Delta V_{DC}]$ . The selection of duty cycle within the range of  $[0,1]$  is governed by the variable  $X$ . The divergence of PV power from the intended generation capacity is denoted by  $\Delta P_{PV}$ , while the discrepancy between the reference coupling voltage and the measured  $V_{DC}$  is represented by  $\Delta V_{DC}$ .

#### Action Space

The MPPT problem is typically associated with a discrete action space. The aforementioned approach ensures a notable degree of accuracy and serves as a potent pedagogical strategy, rendering it a computationally expedient methodology. The EDRL-MPPT agent's action involves a predetermined duty cycle. The duty cycle  $D_c$  is determined within the range of  $D_c = (0,1]$  through a sequence of actions, with an incremental interval of 0.01 in case of discrete action space. Consequently, a matrix comprising of one hundred potential actions is generated.

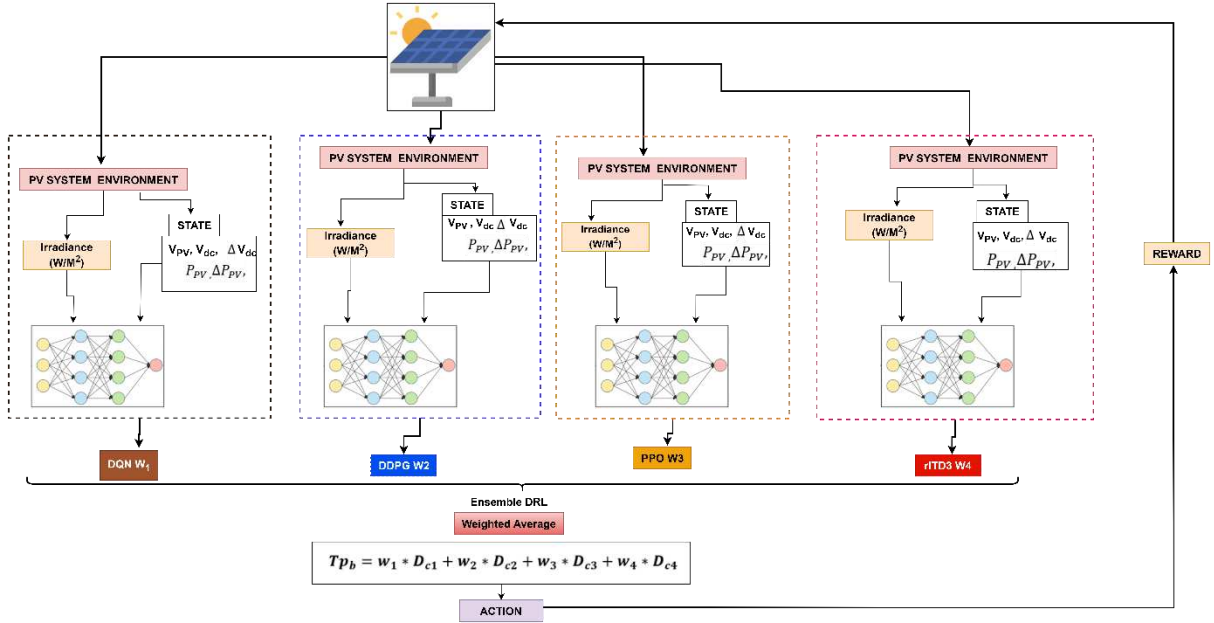


Figure 2. Ensembled DRL for MPPT control in a three-phase grid connected PV system

### Reward

The agent receives a reward contingent upon the action executed. The reward signals have been designed to achieve the objective of minimizing the total harmonic distortion in grid current. Total harmonic distortion (THD) is a crucial metric for assessing power quality as it measures the level of harmonic distortion that exists in the waveform of the grid current. Through the incorporation of Total Harmonic Distortion (THD) as a form of incentive in ensembled Deep Reinforcement Learning (EDRL), the agents are capable of acquiring knowledge to enhance the efficiency of the system's operations and control tactics. This enables the minimization of harmonic distortion and guarantees a seamless transfer of power to the grid. The utilization of the ensemble approach facilitates the agents to collaboratively explore diverse control parameters, strategies, and switching techniques with the aim of mitigating total harmonic distortion (THD). By means of iterative learning and collaborative efforts, the ensemble agents are capable of identifying optimal solutions that effectively reduce harmonic distortion and improve the overall power quality within the grid-connected photovoltaic system.

The total harmonic distortion (THD) in grid current is a measure of the harmonic distortion present in the current waveform of a three-phase grid-connected PV system. When designing a reward function for Deep Reinforcement Learning (DRL) applied in such a system, the THD can be incorporated as a component to incentivize the reduction of harmonic distortions.

The mathematical formula for calculating THD in the grid current can be expressed as follows:

$$THD = \sqrt{\left(\frac{I_{rms}}{I_1}\right)^2} \times 100 \quad (2)$$

Where, THD is the total harmonic distortion of the grid current,  $I_{rms}$  represents the rms value of the individual harmonic current components,  $I_1$  represents the rms value of the fundamental current component. When incorporating THD in the reward function for DRL, the goal would be to minimize the THD value. This can be achieved by assigning a negative reward proportional to the THD value. For instance, the THD-based reward function can be defined as:

$$Reward = -K \times THD \quad (3)$$

Where, Reward is the reward value assigned to the DRL agent, K is a scaling constant that determines the weight or importance of the THD in the overall reward calculation. By using this reward function, the DRL agent is encouraged to learn policies that result in lower THD values, thus promoting a reduction in harmonic distortions in the grid current of the three-phase grid-connected PV system.

The proposed ensembled deep reinforcement learning (EDRL) involves the construction of a collective of DRL models that cooperate in order to enhance the overall efficacy and resilience of the learning system. The weighted ensembled DRL learning process involves training multiple DRL models independently, each with its own set of weights. During the decision-making phase, the models' outputs are combined using weighted averaging or another aggregation method that considers the assigned weights. The weights can be adjusted dynamically based on the models' performance, exploration-exploitation trade-offs, or other criteria. This approach aims to leverage the diversity and expertise of individual agents while assigning different degrees of importance to their contributions based on their performance or confidence levels. Algorithm-1 presents the proposed weighted average deep reinforcement learning (DRL) approach. Four distinct neural network agents (namely DQN, DDPG, rITD3, and PPO) were utilized in this study. Each agent was trained independently, with unique sets of parameters, exploration strategies, and architectures, as outlined in the subsequent section. The aforementioned agents engage with their surroundings, obtain incentives, and revise their strategies through reinforcement learning methodologies. Upon completion of training the four Deep Reinforcement Learning (DRL) models, the resultant policy is utilized to obtain the duty cycle denoted as  $D_c$ . Therefore, the overall probability denoted as  $TP_b$  can be expressed as a weighted average of the duty cycle obtained from individual DRLs, as shown in Equation (4).

$$TP_b = w_1 * D_{c1} + w_2 * D_{c2} + w_3 * D_{c3} + w_4 * D_{c4} \quad (4)$$

The weights of DQN (discrete), DDPG (continuous), rITD3 (continuous), and PPO (discrete) are denoted as  $w_1$ ,  $w_2$ ,  $w_3$ , and  $w_4$ , respectively. Additionally, the duty cycle for each of the DRL models are represented as  $D_{c1}$ ,  $D_{c2}$ ,  $D_{c3}$ , and  $D_{c4}$ . Based on the Algorithm of DQN (discrete), DDPG (continuous), rITD3 (continuous), and PPO [23], the models are trained and saved. The approach employed to obtain output from a weighted average deep reinforcement learning (DRL) method is derived from the research paper referenced as [23]. The method is applied to each model, resulting in the acquisition of  $D_{c1}$ ,  $D_{c2}$ ,  $D_{c3}$ , and  $D_{c4}$ . The final weighted average action is obtained using Equation (4). The model structure that has been suggested is depicted in Figure 2. The algorithm

**Algorithm 1: EDRL THD Reward MPPT PV Control**

1. Establish a connection to the solar PV array SunPower SPR-415E.
2. Determine the magnitude of the current and voltage that results from a short circuit and open circuit
3. Calculate the maximum power for  $N_s = 7$  (PV's in series) and  $N_p = 88$  (PV's in parallel) using  $P_{mpp} = (N_s \times V_{mpp}) \times (N_p \times I_{mpp})$
4. Choose the DC-link voltage.
5. Set the EDRL agent's initial state, action, and reward.
6. State-space  $X = [V_{PV}, I_{PV}, P_{PV}, \Delta P_{PV}, \int \Delta P_{PV}]$
7. Action space  $U = (0,1]$
8. Update duty cycle  $D_c$
9. calculate:  $e(t)$  and  $\Delta e(t)$
10. Pass the error and  $\Delta e(t)$  through
11. Provides these values to the network DQN, PPO, TD3, and DDPG
12. Initialize / Load  $Q$ ,  $\alpha$  learning rate, and  $\gamma$  discount factor.
13. **for**  $j = 1$  **to**  $M$  **do**
14.     Get initial state  $x_0$
15.     **for**  $t = 1$  **to**  $T$  **do**
16.         Select action  $u_t$  from the set defined
17.         **Execute** the action  $u_t$
18.         Get a new state  $x_{t+1}$  and reward  $r$
19.         Store the transition  $(x_t, u_t, x_{t+1}, r)$
20.         **IF**  $|R| > N$
21.             Update the network using weighted average EDRL:
22.         **end if**
23.         Set  $x_t = x_{t+1}$
24.     **end for**
25. **end for**

The present study employs an EDRL methodology that involves assessing a weighted average of the behaviors exhibited by four distinct models. These models, which will be elaborated upon in the subsequent section, consist of two models operating within a continuous action space and two operating within a discrete action space. The average reward per episode during training of EDRL can be shown in figure 3.

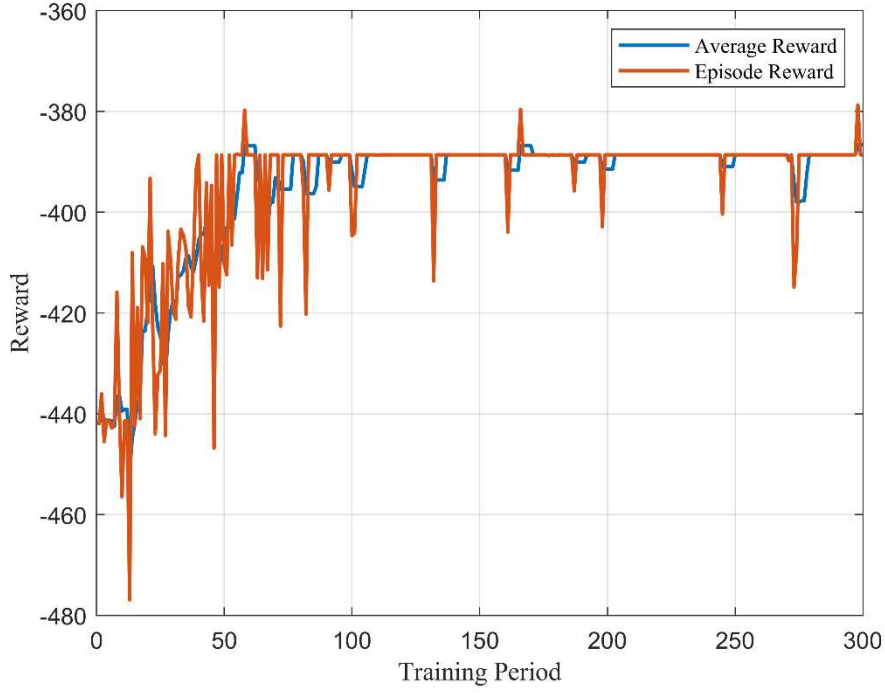


Figure 3. The average reward per episode during training of EDRL

### 3.3.1 DQN (Discrete)

DQN and its associated variant, double Q, are commonly employed to manage a discrete action space [16-19]. As per reference, the selection of a policy to implement in DQN is contingent upon identifying the action that will result in the highest attainable reward for the present state. The primary concern raised in DQN pertains to an overestimation of the Q-value in action, which poses a hindrance to the development of an optimal strategy. The utilization of a policy-based paradigm is advantageous in addressing the intricacy of DQN. The DDPG and PPO models exhibit versatility in their ability to operate effectively in both continuous and discontinuous action spaces. The present study employs a sole DQN in the context of DDPG and a discrete action space, as opposed to a continuous one. Consequently, the model computes the ensembled average of the behaviour with appropriate weighting.

DQN stores state-action pairings as  $\langle S_t, A_t, S_{t+1}, A_{t+1} \rangle$ .  $S_{t+1}, A_{t+1}$  are the states and actions at  $t + 1$ . DQN provides extremely uncorrelated data from random prior events.

Weight loss function ( $\theta$ ):

$$L_i(\theta_i) = \mathbb{E}_{S,A \sim \rho(\cdot)} [(yS_i - Q(S,A; \theta_i))^2] \quad (5)$$

Equation (5) gives the target as  $y_i$  and prediction as  $Q(S,A; \theta_i)$ . In the previous iteration, the weights were  $\theta_{i-1}$ , and the desired output from Equation (5) was

$$y_i = \mathbb{E}_{S', \sim \varepsilon} [r + \gamma \max_{A'} Q(S', A'; \theta_{i-1})] \quad (6)$$

The stochastic gradient of equation 5 with the replacement of the target  $y_i$  using equation 6 is

$$\nabla_{\theta_i} L_i(\theta_i) = \mathbb{E}_{S,A \sim \rho(\cdot)} [(r + \gamma \max_{A'} Q(S', A'; \theta_{i-1}) - Q(S,A; \theta_i)) \nabla_{\theta_i} Q(S,A; \theta_i)] \quad (7)$$

The variable  $\gamma$  denotes the discount factor, while  $r$  represents the reward. The computational procedure executes for a total of  $M$  iterations. The agent selects a set of tuples from the existing data based on the chosen action. The execution of the task is obtained by the agent via the values indicated in Equation (7). The incentive is contingent upon behaviour and can manifest as either a favourable or unfavourable outcome.

### 3.3.2 DDPG (Continuous)

The DDPG model is categorized as a policy-based approach that is applicable to action spaces that are either continuous or discrete. The present approach is proposed as a solution to address the challenges encountered in the Deep Q-Network (DQN) and to mitigate the impact of densely connected neural networks. The approach described in reference [21] has demonstrated efficacy in managing environments characterized by continuous action spaces. Therefore, it is highly appropriate for the current academic research setting.

The DDPG architecture comprises two distinct components, namely the actor and critic. The representation of the actor is denoted as  $\mu(S|\theta^\mu)$ . The representation of the critic can be denoted as  $(S, A|\theta^Q)$ .

The gradient update incorporates the deterministic policy function denoted by  $\theta^\mu$  and the Q network denoted by  $\theta^Q$ . During each iteration of the training process, the actors and critic engage in the exchange of information. In the context of soft update networks in Deep Deterministic Policy Gradient (DDPG), the actor and critic networks are denoted as  $\mu(S|\theta^{\mu'})$  and  $(S, A|\theta^{Q'})$  respectively.

The notation used in the text denotes that  $\theta^{\mu'}$  refers to the Target policy network, while  $\theta^{Q'}$  refers to the Target Q network. The policy function with direction, denoted as  $\theta^\mu$ , is commonly represented as  $J(\theta^\mu)$ , while its gradient is typically denoted as follows:

$$\frac{\partial J(\theta^\mu)}{\partial \theta^\mu} = E[\nabla_{\mu(S)} Q(S, \mu(S|\theta^\mu)|\theta^Q) \nabla_{\theta^\mu} \mu(S|\theta^\mu)] \quad (8)$$

The critic network in the Deep Deterministic Policy Gradient (DDPG) algorithm minimizes the loss function, which is the Mean Square Error (MSE), with respect to the action (A). The resulting expression for the loss function is obtained as a result.

$$(\theta^Q) = E[(Q_{target} - Q_{predict})^2] \quad (9)$$

Where,  $Q_{target} = r + \gamma Q(S_{t+1}, \mu(S_{t+1}|\theta^{\mu'})|\theta^{Q'})$  and  $Q_{predict} = Q(S, A|\theta^Q)$

In contrast to the DQN approach, the target networks undergo updates at each time and step through the utilization of soft updates.

### 3.3.3 PPO (Discrete)

The computation of PPO is an integral component of the weighted ensembled strategy employed in this context. The model in question is a policy-based approach that regulates the gradient of the policy update, as described in reference [21]. This is carried out to ensure alignment between the policies. The utilization of this can be applied to either a discrete or continuous action space. The implication of this statement is that the on-policy model in a discrete action space exhibits a limited capacity for policy modifications during implementation. The evaluation of the performance of a chosen action is conducted through the utilization of the critic network, as per the advantage function. Equation (7) provides the advantage function.

$$\hat{A}_t = \delta_t + (\lambda_\gamma) \delta_{t+1} + \dots + (\lambda_\gamma^{T-t+1}) \delta_{T-1} \quad (10)$$

$$\delta_t = r_t + \gamma V_\pi(S_{t+1}) - V_\pi(S_t) \quad (11)$$

In Equation (8), the state-value function is denoted as  $V_\pi(S)$  and serves as a representation.

$$V_\pi(S) = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | S_t = S] \quad (12)$$

The policy utilized for environmental sampling is denoted as  $\pi_{\theta_{old}}$ , while the policy subject

t to optimization is represented as  $\pi_\theta$ . PPO utilizes the clipped surrogate objective to establish the bounding constraints on the policy updates, thereby ensuring the stability of the training process. The target function utilized in PPO undergoes a transformation, as depicted in Equation (13).

$$J_\theta \approx \sum_{(S_t, A_t)} \min \left( \frac{\pi_\theta(A_t/S_t)}{\pi_{\theta_{old}}} \hat{A}_t, \text{clip} \left( \frac{\pi_\theta(A_t/S_t)}{\pi_{\theta_{old}}}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_t \right) \quad (13)$$

### 3.3.4 rTD3 twin-delayed deep deterministic policy gradient (Continuous)

TD3, an actor-critic structure, combines action as a policy function and value function on the present policy. TD3 needs continuous action space [22]. TD3, the upgraded DDPG, eliminates value function overestimation. Reducing value function overestimation improves accuracy and reduces variance. The TD3 network forms the



target network using the smallest of two critic networks. TD3's delayed actor network updates every two-time step for stability and efficiency throughout training. Clipped noise calculates targets when the action is selected. High action value makes the model robust in all scenarios (continuous action-space). DDPG overestimates the Q-function, but TD3 employs dual Q-functions, delayed policy updates to ensure stability, and smoothens the target policy. Twin delay DDPG. For continuous action-space environments.

Equation (14) represents the target action of the policy  $\mu_{\theta_{targ}}$ , which has been modified by the addition of clipped noise.

$$A'(S') = clip(\mu_{\theta_{targ}}(S') + clip(\epsilon, -c, c), A_{Low}, A_{High}), \epsilon \sim \mathcal{N}(0, \sigma) \quad (14)$$

All instances of the target action A are found to be satisfactory,  $A_{Low} \leq A \leq A_{High}$

Equation (15) presents the clipped double Q-learning function.,

$$y(r, S', d) = r + \gamma(1 - d) \min_{i=1,2} Q_{\phi_{i,targ}}(S', A'(S')) \quad (15)$$

The policy is acquired through the process of maximizing the  $Q_{\phi_1}$ :

$$\max_{\theta} E[Q_{\phi_1}(S, \mu_{\theta}(S))] \quad (16)$$

Equation (17) provides the expression for the Q-function through one-step gradient descent.

$$\nabla_{\phi_i} \frac{1}{|B|} \sum_{(S,A,r,S',d)} (Q_{\phi_1}(S, A) - y(r, S', d))^2 \quad (17)$$

for  $i = 1, 2$

Equation (18) depicts the policy update utilizing one step gradient ascent,

$$\nabla_{\phi_i} \frac{1}{|B|} \sum_{S \in B} Q_{\phi_1}(S, \mu_{\theta}(S)) \quad (18)$$

### 3.4 COA Fuzzified-PLL based Controller for DC offset Removal

The output of a PV inverter typically exhibits a direct current (DC) offset voltage component. This phenomenon arises due to various factors, such as discrepancies among power modules, asymmetry in driving pulses, and errors in current detection. According to reference [3], the presence of induced DC offset in the measured grid voltage can result in undesirable fluctuations in the estimated values of both the amplitude and frequency of the grid voltage. The amplitude of the induced ripple is contingent upon both the percentage of the DC offset value and the fundamental frequency of the power grid. Thus, the presence of DC offset in the measured grid voltage renders the estimation procedure of the grid parameters virtually infeasible. This paper presents a novel approach for mitigating induced DC offset. The proposed method employs COA Fuzzified-PLL-based controller, can be shown in Figure 3. The Fuzzified-Phase-Locked Loop (PLL) was developed through the replacement of the Proportional-Integral (PI) controller with a Fuzzy Logic Controller (FLC) within the context of the PLL. The Coati Optimization algorithms are employed to optimize the parameters of the fuzzy, including membership functions, rule base, and scaling factors. Fuzzification is applied to handle the uncertainty and imprecision associated with DC offset, transforming the input signals into fuzzy sets.

The phase-locked loop (PLL) operates by minimizing the phase discrepancy between a reference signal and a feedback signal within the control loop. The process involves modifying the phase of a voltage-controlled oscillator (VCO) until the two signals achieve phase alignment. Phase-locked loops (PLLs) are vulnerable to various factors such as noise, non-linearities, and abrupt disturbances, which can adversely impact their operational efficiency. The introduction of fuzzy logic to the PLL architecture helps overcome these limitations. Changes in angular phase angle ( $\Delta\theta$ ) can be observed subsequent to each phase angle jump. The application of phase angle change is intended to address subtle and abrupt changes. The proposed model inserts the fuzzy controller between the phase detector and the low-pass filter in a conventional PLL device.

In the conventional phase-locked loop (PLL), the three-phase voltage vector is converted from the  $abc$  natural reference frame to the  $\alpha\beta$  stationary reference frame through the utilization of Clarke's transformation. Subsequently, it is further converted to the  $dq$  rotating frame using Park's transformation, as depicted in Figure 1. The proposed modification being suggested for the PLL framework pertains to its existing control mechanism. The difference between the reference currents  $I_d$  and  $I_q$  and the measured currents  $I_d$  and  $I_q$  is utilized as the input

signal for the fuzzy logic controller. The equation depicts the modelling design of Fuzzified-PLL. The variables  $V_a, V_b$ , and  $V_c$  represent the magnitudes of the three-phase voltage.

$$\begin{bmatrix} V_a \\ V_b \\ V_c \end{bmatrix} = \begin{bmatrix} V_m \cos \theta \\ V_m \cos(\theta - 2\pi/3) \\ V_m \cos(\theta + 2\pi/3) \end{bmatrix} \quad (19)$$

The conversion of these signals into the stationary reference frame signals  $V_\alpha$  and  $V_\beta$  is achieved through the application of the Clarke transformation, while the Park transformation is utilized for the conversion to the dq frame.

$$\begin{bmatrix} V_\alpha \\ V_\beta \end{bmatrix} = \frac{2}{3} \begin{bmatrix} 1 & -1/2 & -1/2 \\ 0 & -\sqrt{3}/2 & \sqrt{3}/2 \end{bmatrix} \begin{bmatrix} V_a \\ V_b \\ V_c \end{bmatrix} \quad (20)$$

$$\begin{bmatrix} V_d \\ V_q \end{bmatrix} = \begin{bmatrix} \cos \theta^* & -\sin \theta^* \\ \sin \theta^* & \cos \theta^* \end{bmatrix} \begin{bmatrix} V_\alpha \\ V_\beta \end{bmatrix} \quad (21)$$

$$V_d = V_\alpha \cos \theta^* - V_\beta \sin \theta^* \approx V_{d,offset} + V_m \quad (22)$$

$$V_q = V_\alpha \sin \theta^* + V_\beta \cos \theta^* \approx V_{q,offset} - V_{me} \quad (23)$$

The symbols  $V, \theta$ , and  $\theta^*$  represent the, magnitude of voltage, input angle, and estimated angle, respectively.  $e$  is the error in phase angle. The main objective of the proposed controller is to eliminate the direct current (DC) offset component that exists in the synchronous q-axis and d-axis components, denoted as  $V_{d,offset}$  and  $V_{q,offset}$  respectively which is shown equation (23).

The fuzzy controller employs optimized parameters and a rule base to make decisions based on the fuzzy sets of the input signals. The controller generates control signals that adjust the PLL's parameters dynamically to cancel out the DC offset component effectively. The fundamental stages of fuzzy logic control consist of three distinct phases, which are fuzzification, decision-making, and defuzzification. The procedure of fuzzification pertains to the transformation of quantitative measurements of input variables into a precisely defined linguistic variable boundary, which is denoted by a fuzzy set. The performance of the optimized fuzzy-PLL is evaluated by assessing the accuracy of DC offset removal and the overall improvement in power quality.

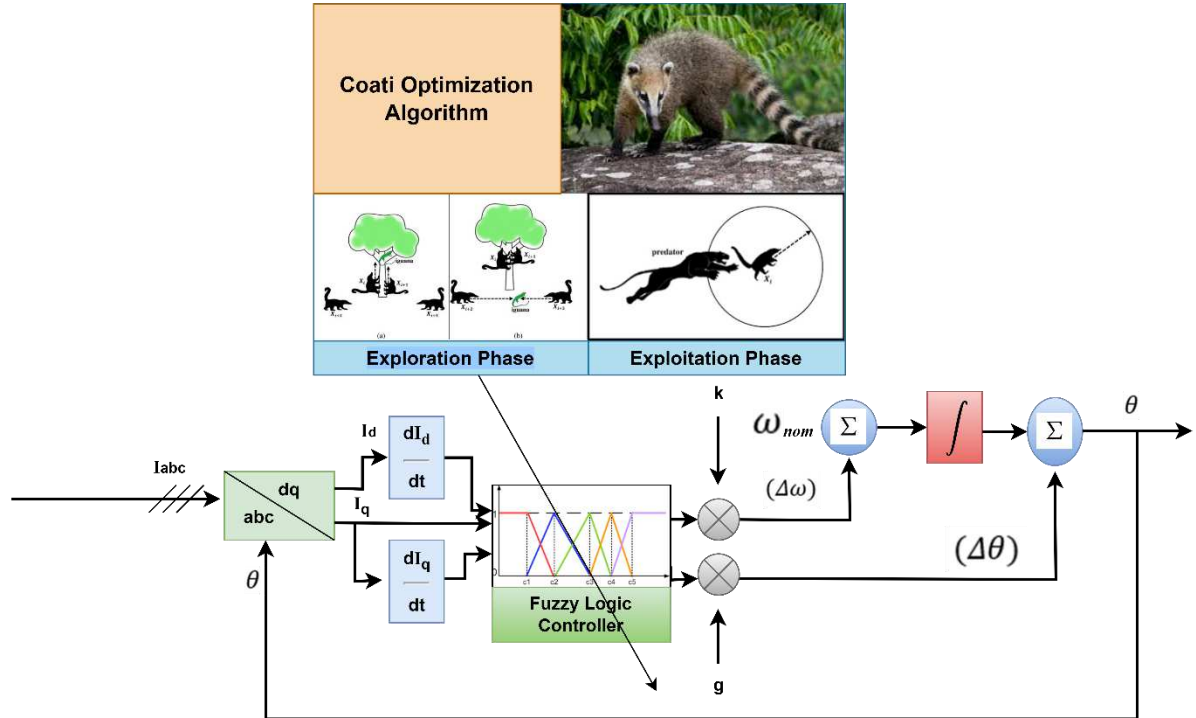


Figure 4. COA Fuzzified-PLL based Controller

The Novel Coati Optimization algorithm is used to optimize the parameter of fuzzified-PLL. Figure 4 illustrates a comprehensive model of a fuzzy phase-locked loop (PLL). The global optimization COA approach offers numerous advantages, including the absence of control parameters, the ability to solve complex high-dimensional problems across various domains, superior research and search process balancing, and effective handling of optimization applications. Given the presence of two inputs and one output, 18 positions of membership functions must be optimized by utilizing the Coati Optimization (COA) algorithm. An objective function is selected to minimize the total harmonic distortion in the optimization process. The fuzzy controller being examined employs linguistic variables, specifically negative big (NB), negative medium (NM), small negative (NS), zero (ZE), small positive (PS), medium positive (PM), and positive big (PB), to establish the error and error rate. These variables are differentiated by their corresponding memberships. The concept of memberships pertains to mathematical curves that facilitate the process of mapping each point within the input space to a corresponding membership value that falls within the range of 0 to 1.

The study involves the modification of eighteen membership function positions for each coati, followed by evaluating a fuzzified-PLL system based on the resulting error. The membership range of two inputs is defined as [-10 to 10] and [-1 to 1], and one output as [-10 to 10]. The above ranges remain constant throughout the simulation and do not undergo any modifications. The trapezoidal function has a fixed range that spans from negative infinity to positive infinity, both at the initial and final points. Additionally, it is worth noting that two points of intersection exist between each membership function and other membership functions. Table 3 provides clear evidence of this assertion.

Table 2 The range values for the input/output of fuzzy controller.

Membership functions	Range parameters
Trapezoidal	$[-\infty, -0.032, x(1), x(2)]$
Triangular	$[x(1), x(2), x(3)]$
Triangular	$[x(2), x(3), 0]$
Triangular	$[x(3), 0, x(4)]$
Triangular	$[0, x(4), x(5)]$
Triangular	$[x(4), x(5), x(6)]$
Trapezoidal	$[x(5), x(6), 0.032, \infty]$

The most common values can be readily anticipated based on the data presented in Table 2. In the case of a single variable, the tuning process involves adjusting four specific values. However, when dealing with three variables, the total number of values to be tuned increases to eighteen. There exist certain constraints that necessitate consideration when adjusting these values.

The given problem is subject to certain constraints, which is given in table 3. require that every value must adhere to the specified inequality criteria:

Table 3. Constraints for Fuzzy Controller

Input	Boundary Condition
Order	$x(1) < x(2) < x(3)$
Error $E$	-0.032 to 0.032
Change in Error $\Delta E$	-10 to 10
Phase angle output $\Delta\theta$	-1 to 1

The membership function values of the fuzzy controller are modified and the model is then executed using these new values. The total harmonic distortion (THD) will serve as the target to be minimized, representing the value of the objective function.

$$\text{objective function} = \min \text{THD} \quad (24)$$

Table 4 presents elaborate variations of the expressions mentioned above for alterations in phase angle. The q-axis component of the grid voltage, along with its derivative, are regarded as inputs to the FLC. In order to incorporate variables with distinct ranges, it is necessary to multiply two variables by the outputs of Fuzzy Logic Controllers (FLCs). The parameters depicted in Figure 4 are denoted by  $k$  and  $g$ . Figure 4 illustrates a comprehensive model of a fuzzy phase-locked loop (PLL).

Table 4 Fuzzy rule for output variable  $\Delta\theta$

$\Delta\theta$	NB	NM	NS	ZE	PS	PM	PB
NB	NB	NB	NB	NM	NM	NS	ZE
NM	NB	NB	NM	NM	NS	ZE	PS
NS	NB	NM	NM	NS	ZE	PS	PM
ZE	NM	NM	NS	ZE	PS	PM	PM
PS	NM	NS	ZE	PS	PM	PM	PB
PM	NS	ZE	PS	PM	PM	PB	PB
PB	ZE	PS	PM	PM	PB	PB	PB

The parameters of the fuzzy controller are tuned using coati Optimization. Following the tuning process, the values of the parameters are modified from their original values, resulting in a change in their overall shape. The figures presented in Figure 5. (a) - 5. (c) illustrate the updated fuzzy logic membership functions, which have been modified to encompass a new range. These modifications have been made to facilitate the identification of the minimum error through coati optimization algorithm.

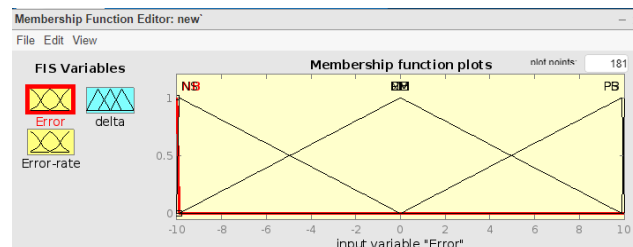


Figure 5. (a): Coati optimised membership function for input  $E$

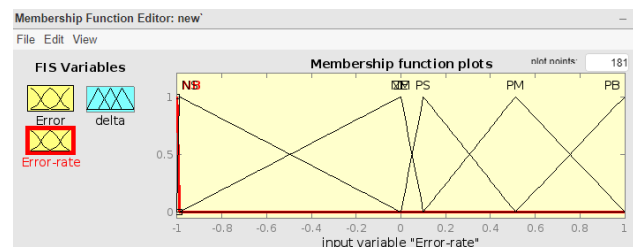


Figure 5 (b): Coati optimised membership function for input  $\Delta E$

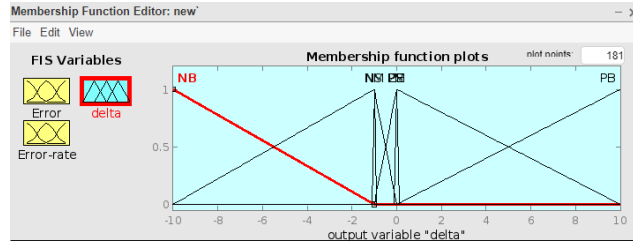


Figure 5 (c): Coati optimised membership function for output  $\Delta\theta$

## Simulation Results

The effectiveness of the proposed EDRL-MPPT technique for inter-harmonics elimination and COA-Fuzzified PLL synchronization approach, as illustrated in Figure 1, has been evaluated and simulated. The simulation has been conducted in three distinct cases as follows:

- The photovoltaic (PV) inverter functions under a consistent level of solar irradiance, specifically known as steady-state maximum power point tracking (MPPT), which corresponds to its rated power, equivalent to 250 kW. It is worth noting that during this operating condition, the emission of inter-harmonics is particularly noticeable. In the first case, this study examines the effect of inter-harmonics during (P & O) Maximum Power Point Tracking (MPPT) technique and conventional Phase-Locked Loop (PLL) synchronization technique in a three-phase grid-connected solar photovoltaic (PV) system.
- In second case, the implementation of the EDRL-MPPT technique and conventional PLL synchronization technique has been proposed for the purpose of inter-harmonics mitigation in a three-phase grid connected solar PV system.
- In the third scenario, it is suggested that the abrupt load variation results in the generation of a DC offset in the grid current. In order to address the disruption caused by this DC offset in the grid current, the EDRL-MPPT technique with COA-Fuzzified PLL synchronization approach have been proposed.

As discussed, introduction section, due to partial shedding condition, the transient response of the dc-link voltage controller is identified as one of the sources of inter-harmonics in the grid current. To mitigate inter-harmonics in the PV inverter, it is imperative to prevent perturbations in the dc-link voltage during operation. The attainment of this objective is facilitated through the utilization of an (EDRL-MPPT). The present study utilizes an ensembled deep reinforcement learning approach that incorporates a reward system based on total harmonic distortion in grid current. The objective is to effectively train the agent to execute actions that based on the minimal harmonic distortion in grid current. Consequently, the inter-harmonics can be effectively circumvented. The figure 6 illustrates a comparison between Perturb and Observe MPPT and EDRL-MPPT techniques in terms of their impact on DC coupling voltage output and inter-harmonic mitigation in waveform.

In the conventional case, to achieve optimal power extraction from photovoltaic system, an MPPT algorithm such as Perturb and Observe (P&O) is utilized to ascertain the reference DC-link voltage (i.e., PV voltage) during operation. Subsequently, the dc-link voltage ( $V_{dc}$ ) regulation is achieved by the dc-link voltage controller utilizing a PLL controller. This controller operates by controlling the grid current. The outcome denoted as  $V_{dc}$ , obtained through the employment of the normal MPPT technique with conventional phase-locked loop control, is depicted in Figure 6. The Perturb and Observe (P&O) technique was employed to evaluate the (MPPT) operation in the initial scenario. However, it is important to note that the injected grid current that is linked to this operation displays a considerably higher level of distortion, as illustrated in Figure 7. The disparity in waveform can also be observed in a magnified plot ranging from 0.1 to 0.2 seconds.

The diagram in Figure 8 displays the frequency spectrum of the grid current with total harmonic distortion of 8.90%. It is observed that the inter-harmonic level exhibits a higher absolute value when the PV inverter is functioning at its rated power. The inter-harmonic emission originating from the photovoltaic inverter holds considerable significance in this scenario. In order to mitigate the presence of harmonics, we have put forth the

EDRL-MPPT technique, which incorporates a reward and penalty system based on the distortion observed in the grid current. The subsequent section of our discussion pertained to the MPPT-EDRL system, which incorporates PLL synchronization.

The second scenario involves using EDRL-MPPT to attain maximum power output from the photovoltaic system, aiming to determine the reference DC-link voltage (i.e., PV voltage) during its operation. The direct current (DC) link voltage, denoted as  $V_{dc}$ , is regulated by utilizing a phase-locked loop (PLL) controller by the DC-link voltage controller. Figure 6 (a) illustrates the system's response to a decrease in irradiance from 1000 to 0 W/m<sup>2</sup> in the presence of a load. The EDRL-MPPT sustains a constant output of dc voltage with less waveform distortion compared to normal MPPT. At a time, interval of 0.55 seconds to 1 sec, there was a significant decrease in solar irradiance to a value of 0 W/m<sup>2</sup>.

Figure 6 (b) illustrates the grid current for the proposed (EDRL-MPPT) system. Figure 6 (c) depicts the amplitude of the primary frequency constituent of grid current. After analyzing the frequency spectrum of the output current depicted in Figure 6 (c), it becomes apparent that the dominant inter-harmonics present in the output current exhibit a significant reduction in magnitude. The total harmonic distortion is 7.069%, measured in this case. Additionally, we clearly see less damping in the angular frequency compared to the conventional MPPT in Figure 6 (d). The utilization of an EDRL MPPT in conjunction with the PLL method demonstrates reduced inter-harmonic levels. A comparative analysis has been conducted to assess the phase error between the Perturb and Observe (P&O) MPPT algorithm and the EDRL MPPT algorithm with a conventional Phase-Locked Loop in typical operational circumstances. The EDRL MPPT with conventional PLL generates precise control signals to accurately track the phase angle of the grid and achieve less phase error between the output signal and that of a reference signal which can be seen in Figure 6 (e).

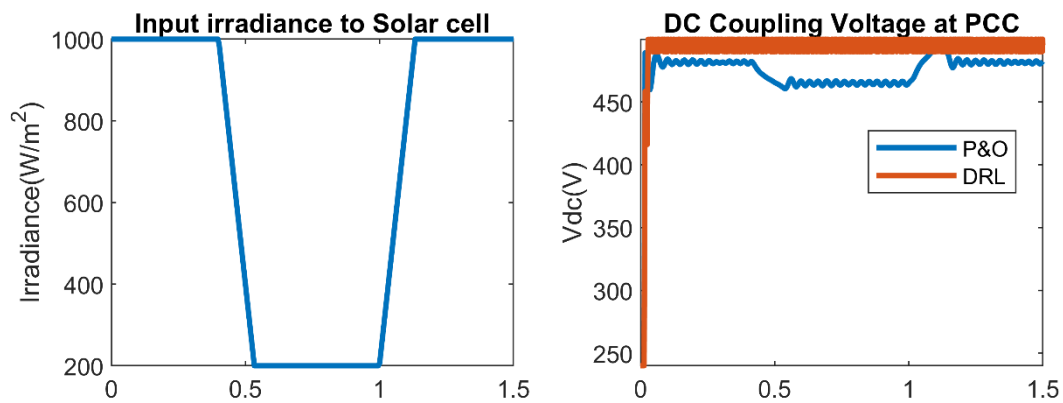


Fig 6 (a), The output of DC coupling voltage of (P&O) and EDRL



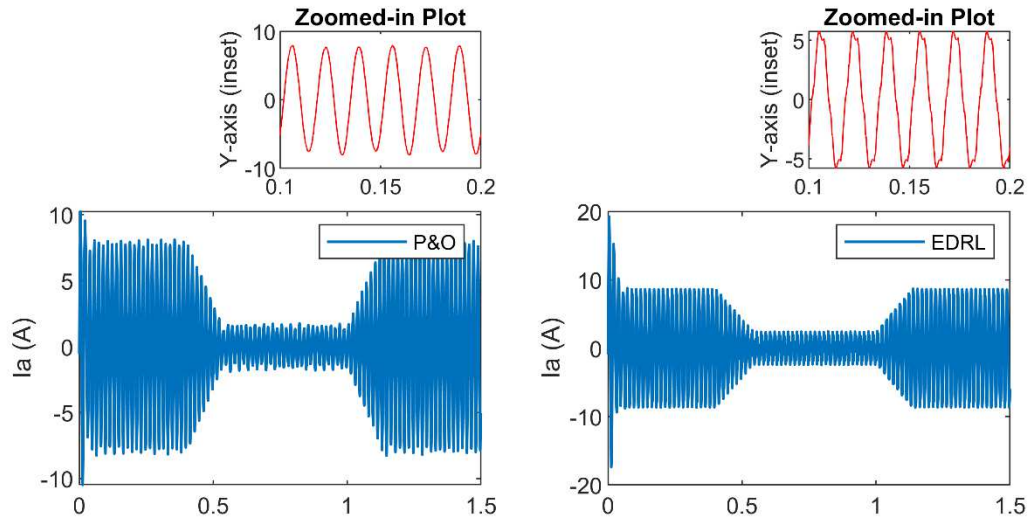


Fig 6 (b), The output of grid current during (P&O) and EDRL MPPT Technique

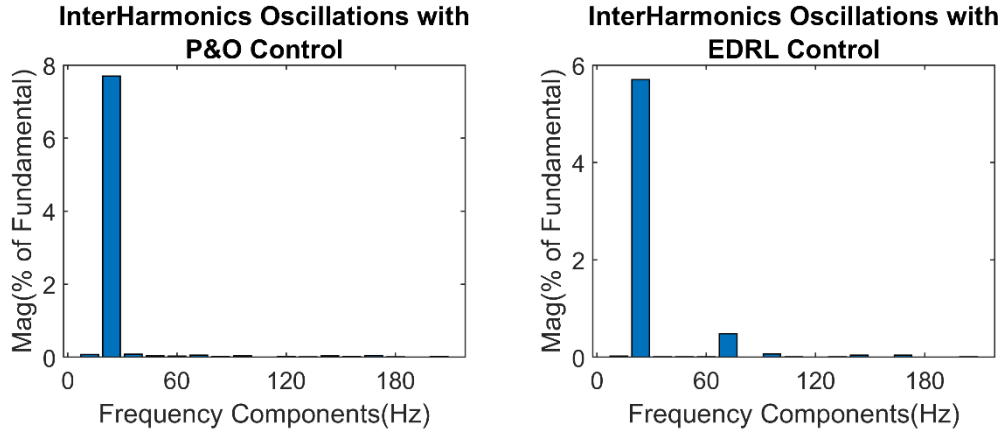


Fig 6 (c), The magnitude of fundamental frequency component during (P&O) and EDRL MPPT Technique

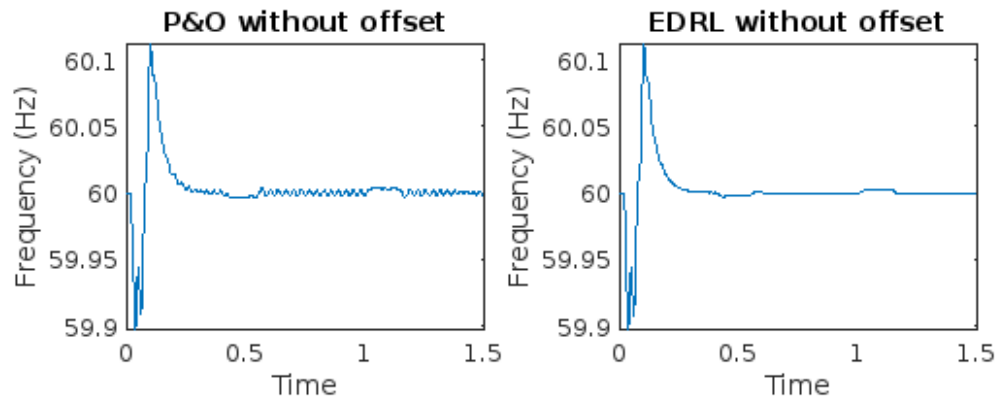


Fig 6 (d), The frequency variation during (P&O) and EDRL MPPT Technique with conventional PLL

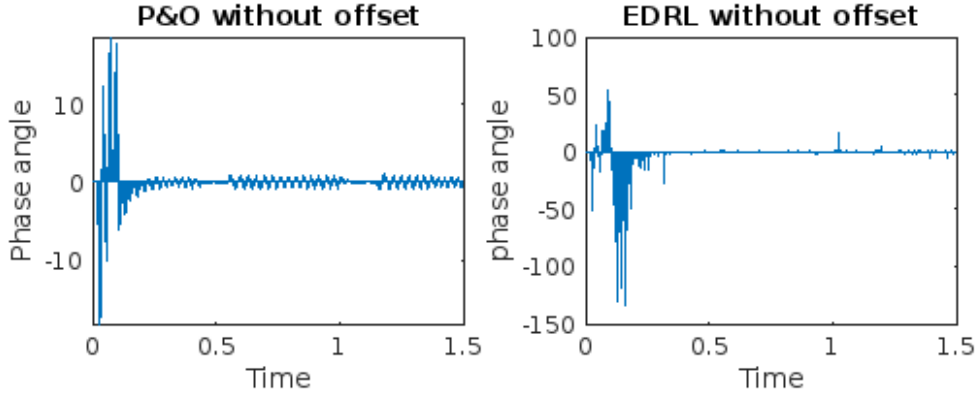


Fig 6 (e), The Phase angle error during (P&O) and EDRL MPPT Technique with conventional PLL

The third case utilizes the EDRL-MPPT and CAO-Fuzzified PLL for grid synchronization. Implementing a control algorithm in a grid-tied photovoltaic system, which aims to eliminate harmonic distortion and compensate DC offset, necessitates the calculation of both the synchronizing signal and the amplitude of the grid current. At the start of operation, the grid functions at its maximum operational capacity. The distribution of load current is divided between the grid and photovoltaic (PV) system based on the amount of solar radiation present. In the event of a sudden reduction in load, the surplus power generated by the photovoltaic (PV) system will be transmitted to the electrical grid. At full load capacity, the grid voltage and grid current exhibit a phase alignment. When there is a sudden decrease in load, it results in a phase shift. This phase shift is mitigated using fuzzy-PLL, and the hyperparameters of fuzzy are tuned using COA results mitigation of DC offset and less harmonics distortion in the grid current as seen in Figure 7 (a).

The mitigation of phase shift is achieved through the utilization of fuzzy-PLL, wherein the hyperparameters of the fuzzy system are adjusted using the COA technique. This approach effectively reduces the presence of DC offset and minimizes the distortion of harmonics in the grid current, as depicted in Figure 7 (a). The respective magnitude of the fundamental frequency component can be seen in Figure 7 (b), with less reduction in total harmonic distortion of 2.89%. The effectiveness of the proposed Fuzzified-PLL under the change in load is depicted in Figure 7 (b). The observation of Figure 7 (b) reveals that the introduction of an offset in the system within the time range of 0.28 to 0.33 seconds results in significant frequency distortion in the conventional phase-locked loop (PLL).

In contrast, the fuzzified-PLL exhibits superior performance during this period. The proposed method has superior performance in various aspects, including dc offset rejection, grid synchronization, inter-harmonic rejection, and stable  $V_{dc}$ . A comparison of the two approaches was performed to evaluate the phase error between the Perturb and Observe (P&O) technique with conventional PLL and EDRL MPPT algorithm with a COA-Fuzzified PLL under offset conditions introducing from 0.1 to 0.3 seconds. The EDRL MPPT with COA-Fuzzified PLL is capable of producing highly accurate control signals for effectively monitoring the phase angle of the grid. This results in minimal phase error between the output signal and a reference signal, as depicted in Figure 7 (c).

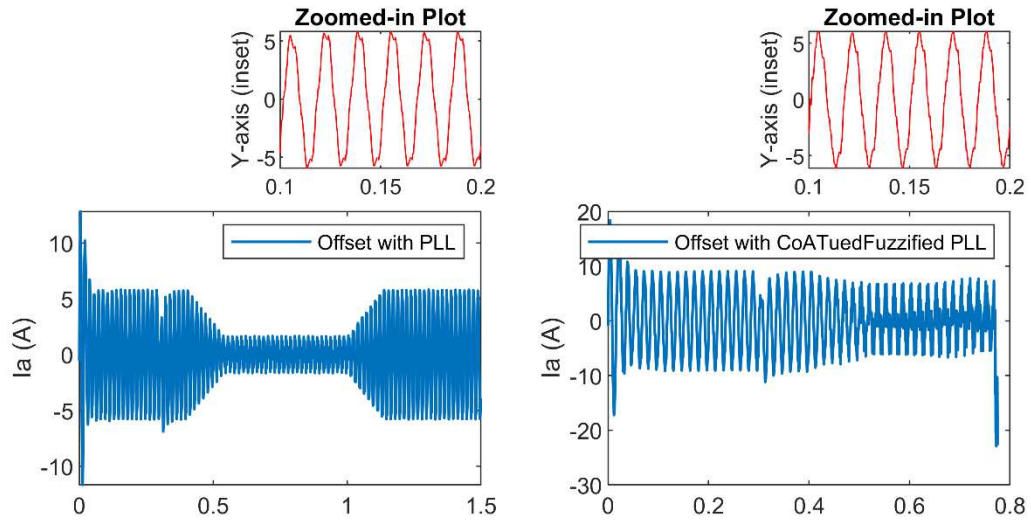


Fig 7 (a), The output of grid current during EDRL-COA Fuzzified PLL synchronization technique

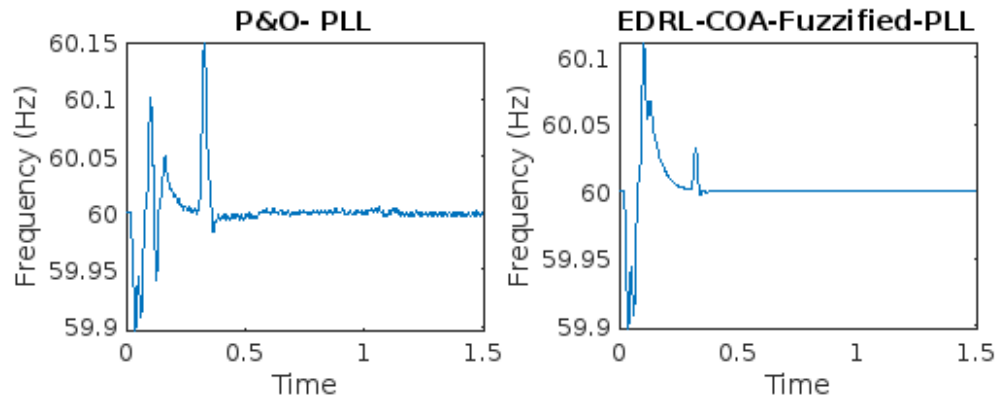


Fig 7 (b), The frequency variation during EDRL MPPT with COA-Fuzzified PLL Technique

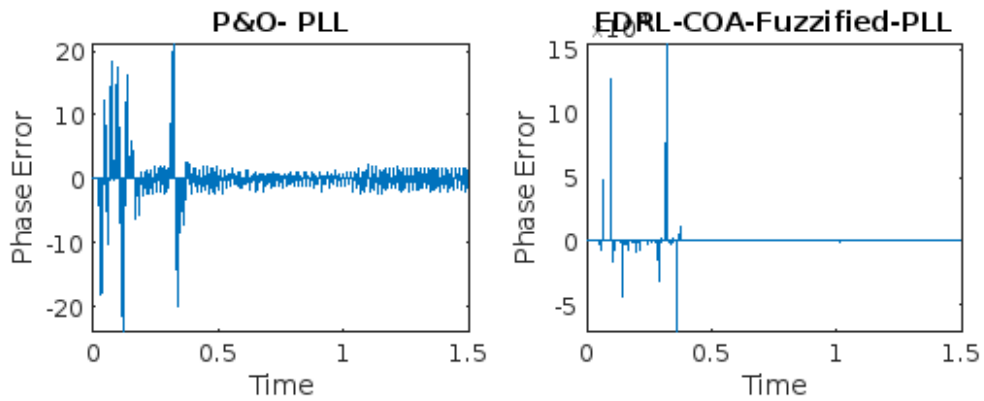


Fig 7 (c), The Phase angle error during EDRL MPPT Technique with COA-Fuzzified PLL

#### 4.1 State-of-art comparison

Table 5 displays a comparative analysis of the proposed COA-fuzzified-PLL controller alongside existing control techniques and a limited number of adaptive control techniques. The provided comparative table illustrates the efficacy of the proposed hybrid control scheme with respect to multiple attributes such as rejection of DC offset and THD of grid current. The proposed scheme demonstrates superior performance compared to existing control algorithms such as MCCF (Multiple complex coefficient filter), MCCF-SOGI (Multiple complex coefficient filter second-order generalized integrator) and fuzzy logic proportional–integrator–derivative multiple complex coefficient filter multiple second-order generalized integrator frequency-locked loop (FLPID-MCCF-MSOGI-FLL) as evidenced by the results presented in Table 5.

Traditional MPPT algorithms with conventional PLL synchronization techniques often exhibit overshoot in estimated phase angle, temporary deviations from desired values during transient conditions. The improved power quality approach employs ensemble DRL MPPT control and COA-fuzzified-PLL-based synchronizing system, which has less overshoot through adaptive control strategies and intelligent decision-making. Continuously optimizing control parameters based on real-time system dynamics, these advanced techniques minimize overshoot, resulting in enhanced power quality and stability.

Similarly, settling time measures the duration for the system's output to reach and remain within a specified range after a disturbance or change in operating conditions. Traditional methods tend to have longer settling times due to slower convergence or suboptimal control actions. In contrast, the ensemble DRL MPPT controller and COA-fuzzified-PLL-based synchronizing system reduce settling time through intelligent learning algorithms and adaptive control strategies. These advanced techniques facilitate faster convergence and more accurate tracking of desired operating points, leading to reduced settling time and improved power quality performance.

Table 5. The summary of performance of Different controller

Features	MCCF [24]	MCCF-SOGI [24]	FLPID-MCCF-MSOGI-FLL [24]	Conventional PLL	Proposed COA-Fuzzified-PLL
Oscillation	Less	Less	Less	Less	Less
DC Offset Rejection	No	No	Better	No	Good
Inter-Harmonics Removal	No	Yes	Yes	No	Yes
THD Of Grid Current	Less	Less	Better	No	Good
Steady State Performance	Good	Good	Good	Good	Good
Transient Performance	Good	Good	Good	Good	Good
Grid Synchronization	Yes	Yes	Yes	No	Yes
Overshoot	-	-	-	44.03	0
Settling Time	-	-	-	71ms	68ms

## Conclusion

In conclusion, this research study aimed to improve the power quality of three-phase grid-connected photovoltaic (PV) systems by addressing inter-harmonics and DC offset. This was achieved by employing the Enhanced Droop Regulator Control (EDRL) Maximum Power Point Tracking (MPPT) technique and the COA-fuzzified-Phase-Locked Loop (PLL) synchronization system. Extensive simulations have revealed that the integration of the EDRL MPPT technique and the COA-fuzzified-PLL synchronization system yields a substantial improvement in power quality within three-phase grid-connected photovoltaic (PV) systems. The reduction of inter-harmonics and DC offset leads to a decrease in voltage distortion, an enhancement in system stability, and an improvement in grid integration. Future research and development endeavours may prioritize the optimization of control strategies, the exploration of advanced algorithms, and the consideration of real-time implementation in order to augment the performance of the proposed techniques in practical contexts.

## Declarations

### Ethical Approval

This declaration is not applicable.

### Competing interests

Both authors hereby declare that they have no competing interests to disclose in any of the cases associated.

### Authors' contributions

Both authors had equivalent contributions in the work presented. All decisions vis-à-vis conceptual as well as diplomatic or technical aspects were taken in mutual consent and thus reflect equal contribution and accountability.

### Funding

This declaration is not applicable.

### Availability of data and materials

The data and materials can be made available on request if there are no confidentiality restrictions associated to the same.

## References

- [1]. Sridharan, K. and Babu, B.C., 2021. Accurate phase detection system using modified SGDFIT-based PLL for three-phase grid-interactive power converter during interharmonic conditions. *IEEE Transactions on Instrumentation and Measurement*, 71, pp.1-11.
- [2]. Panwar, N.L., Kaushik, S.C. and Kothari, S., 2011. Role of renewable energy sources in environmental protection: A review. *Renewable and sustainable energy reviews*, 15(3), pp.1513-1524.
- [3]. Lubura, S., Šoja, M., Lale, S.A. and Ikić, M., 2014. Single-phase phase locked loop with DC offset and noise rejection for photovoltaic inverters. *IET Power Electronics*, 7(9), pp.2288-2299.
- [4]. Liu, B., An, M., Wang, H., Chen, Y., Zhang, Z., Xu, C., Song, S. and Lv, Z., 2020. A simple approach to reject DC offset for single-phase synchronous reference frame PLL in grid-tied converters. *IEEE Access*, 8, pp.112297-112308.
- [5]. Han, Y., Luo, M., Zhao, X., Guerrero, J.M. and Xu, L., 2015. Comparative performance evaluation of orthogonal-signal-generators-based single-phase PLL algorithms—A survey. *IEEE Transactions on Power Electronics*, 31(5), pp.3932-3944.
- [6]. Saxena, H., Singh, A. and Chittora, P., 2023. Modified LMS synchronization technique for distributed energy resources with DC-offset and harmonic elimination capabilities. *ISA transactions*, 135, pp.567-574.
- [7]. Pandey, R. and Kumar, N., 2023, March. Advanced TOGI Controller for Weak Grid Integrated Solar PV System. In *2023 IEEE IAS Global Conference on Renewable Energy and Hydrogen Technologies (GlobConHT)* (pp. 1-6).
- [8]. Saxena, H., Singh, A. and Rai, J.N., 2021. Analysis of SOGI-ROGI for synchronization and shunt active filtering under distorted grid condition. *ISA transactions*, 109, pp.380-388.
- [9]. Xie, M., Huiqing wen. Canyan Zhu and Yong Yang, "DC offset rejection improvement in single-phase SOGI-PLL algorithms: methods review and experimental evaluation" *IEEE Access*, 5.
- [10]. Saxena, H., Singh, A. and Rai, J.N., 2019, November. Design and Analysis of Cascaded Generalized Integrators for Mitigation of Power Quality Problems. In *2019 International Symposium on Advanced Electrical and Communication Technologies (ISAECT)* (pp. 1-6). IEEE.
- [11]. Punitha, K., Devaraj, D. and Sakthivel, S., 2013. Development and analysis of adaptive fuzzy controllers for photovoltaic system under varying atmospheric and partial shading condition. *Applied Soft Computing*, 13(11), pp.4320-4332.
- [12]. Liu, L., Liu, C., Wang, J. and Kong, Y.G., 2015. Simulation and hardware implementation of a hill-climbing modified fuzzy-logic for mppt with direct control method using boost converter. *Journal of Vibration and Control*, 21(2), pp.335-342.

- [13]. Kermadi, M., Salam, Z., Ahmed, J. and Berkouk, E.M., 2018. An effective hybrid maximum power point tracker of photovoltaic arrays for complex partial shading conditions. *IEEE Transactions on Industrial Electronics*, 66(9), pp.6990-7000.
- [14]. Rizzo, S.A. and Scelba, G., 2015. ANN based MPPT method for rapidly variable shading conditions. *Applied Energy*, 145, pp.124-132.
- [15]. Singh, Y. and Pal, N., 2021. Reinforcement learning with fuzzified reward approach for MPPT control of PV systems. *Sustainable Energy Technologies and Assessments*, 48, p.101665.
- [16]. Xia, Z., Wu, J., Wu, L., Yuan, J., Zhang, J., Li, J., & Wu, D. (2021, July). RLCC: Practical Learning-based Congestion Control for the Internet. In *2021 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.
- [17]. Lan, Q., Pan, Y., Fyshe, A., & White, M. (2020). Maxmin q-learning: Controlling the estimation bias of q-learning. *arXiv preprint arXiv:2002.06487*.
- [18]. Anschel, O., Baram, N., & Shimkin, N. (2017, July). Averaged-dqn: Variance reduction and stabilization for deep reinforcement learning. In *International conference on machine learning* (pp. 176-185). PMLR.
- [19]. Liu, X. Y., Yang, H., Chen, Q., Zhang, R., Yang, L., Xiao, B., & Wang, C. D. (2020). FinRL: A deep reinforcement learning library for automated stock trading in quantitative finance. *arXiv preprint arXiv:2011.09607*
- [20]. Liu, Q., Cheng, L., Jia, A. L., & Liu, C. (2021). Deep reinforcement learning for communication flow control in wireless mesh networks. *IEEE Network*, 35(2), 112-119.
- [21]. Yang, H., Liu, X. Y., Zhong, S., & Walid, A. (2020, October). Deep reinforcement learning for automated stock trading: An ensemble strategy. In *Proceedings of the First ACM International Conference on AI in Finance* (pp. 1-8).
- [22]. Zhu, J., Wu, F., & Zhao, J. (2021, December). "An Overview of the Action Space for Deep Reinforcement Learning", In *2021 4th International Conference on Algorithms, Computing and Artificial Intelligence* (pp. 1-10).
- [23]. Dehghani, M., Montazeri, Z., Trojovská, E. and Trojovský, P., 2023. Coati Optimization Algorithm: A new bio-inspired metaheuristic algorithm for solving optimization problems. *Knowledge-Based Systems*, 259, p.110011.
- [24]. Babu, N., Guerrero, J.M., Siano, P., Peesapati, R. and Panda, G., 2020. An improved adaptive control strategy in grid-tied PV system with active power filter for power quality enhancement. *IEEE Systems Journal*, 15(2), pp.2859-2870.



RESEARCH ARTICLE | SEPTEMBER 05 2023

## Environomical analysis of green building having various window-to-wall ratio

Asim Ahmad ; Om Prakash; Pranav Nayan; Anil Kumar; Bharath Bhushan; Rajeshwari Chatterjee



*AIP Conf. Proc.* 2863, 020002 (2023)

<https://doi.org/10.1063/5.0155710>



CrossMark

### Articles You May Be Interested In

Environomical analysis of sensible heat storage-based greenhouse dryer

*AIP Conf. Proc.* (September 2023)

Numerical simulation of residential building on the basis of natural ventilation

*AIP Conference Proceedings* (July 2018)

Prediction of climate-based daylight metrics by simulating monthly median illuminance

*AIP Conference Proceedings* (March 2019)

500 kHz or 8.5 GHz?  
And all the ranges in between.

Lock-in Amplifiers for your periodic signal measurements



Find out more



# Environomical Analysis of Green Building Having Various Window-to-Wall Ratio

Asim Ahmad<sup>1,a)</sup>, Om Prakash<sup>2,b)</sup>, Pranav Nayan<sup>2,c)</sup>, Anil Kumar<sup>3,d)</sup>, Bharath Bhushan<sup>2,e)</sup>, Rajeshwari Chatterjee<sup>4,f)</sup>

<sup>1</sup>Faculty of Engineering and Applied Sciences, Usha Martin University, Ranchi-835103, India

<sup>2</sup>Department of Mechanical Engineering, Birla Institute of Technology, Mesra, Ranchi- 835215, India

<sup>3</sup>Department of Mechanical Engineering, Delhi Technological University, Delhi-110042, India

<sup>4</sup>Department of HMCT, Birla Institute of Technology Mesra, Ranchi-835215, India

<sup>a)</sup>Corresponding author: asimlife91@gmail.com

<sup>b)</sup>omprakash@bitmesra.ac.in

<sup>c)</sup>nayanpranav@gmail.com

<sup>d)</sup>anilkumar76@dtu.ac.in

<sup>e)</sup>bhushanbharath96@gmail.com

<sup>f)</sup>rajchmimi@gmail.com

**Abstract.** The building industry is one of the most energy-intensive in developed countries. To reduce climate change emissions across a building's entire life cycle, in addition to minimizing energy consumption during the operational phase, consideration should be given to the embodied energy and CO<sub>2</sub> emissions of the building itself. Window-to-wall ratio (WWR) is a building parameter that plays a vital role in inside lighting and temperature. Light is a significant asset with characteristics of both radiometry and photometry. In the present work, environomical analysis of the various green buildings has been analyzed.

**Key Words:** Embodied Energy, Building envelope, Window to Wall Ratio, Energy Plus and eQuest.

## INTRODUCTION

A green building uses less energy, water, and other natural resources than a traditional home, emits less waste and Green House Gases, and is better for people to live or work in (Omer et al.) [1]. Green Structure also refers to a sustainable atmosphere, clean water, and safe living. It's not just about being more productive when it comes to building green. It's all about designing structures that are friendly to the environment, use local resources, and, most notably, use less energy, water, and materials [2]. As a result, if these considerations are retained in mind, we can see that conventional architecture was, in effect, very green. Today, we have forgotten how to create natural environments and just emulate what developing countries have done. Buildings are a significant energy consumer in the economy. Buildings consume 35% to 40% of total electricity during construction [3]. Buildings use the most energy during renovation and later in lighting and air-conditioning facilities. This intake must be kept to a minimum. This could be reduced to about 80-100 watts per square meter order to have a Green Building Concept, we must consider the following [4]:

1. The most efficient use of energy.
2. Conservation of water.
3. Waste collection, treatment, and reuse of solid and liquid waste.
4. Transportation networks that are energy efficient.
5. Effective Building System Planning.

In a building, energy audit is a process which involves survey, inspection and energy analysis to achieve energy conservation. Building energy consumption has significantly increased over the past few decades as a result of development and a rise in living standards [5]. In India, building energy consumption in the year of 2020 represents about 26% of total energy and ranked third among the most CO<sub>2</sub> emitting countries. Commercial building has a total built space of 33% which is increasing at a rate of 8-10% annually. The average electricity consumption in India is around 80 KWh/m<sup>2</sup> and 160 KWh/m<sup>2</sup> for residential and commercial buildings respectively. Domestic buildings utilize 30% to 35% of energy and emit 32% of CO<sub>2</sub>. In recent years, there has been increase in energy consumption from 10% to 12% in commercial buildings due to urban development. India is focusing to bring down from 34% to 31% CO<sub>2</sub> emission by 2030. The per capita energy consumption in India is 0.64 MWh in which 37% is utilized for cooling, 29% for lighting, 12% for refrigeration and 9% for air-conditioning [6]. To assess the impact of WWR on building energy consumption, the ASHRAE recommended a three- term condition to decide the energy moving through window assembly which depends on a mixtures of physical coating properties (e.g., Single coating or twofold coating) and climatic conditions. The thermal performance of building goes through effects of WWR on building energy use. The relationship of window to wall or window to roof and vice versa is related to thermal performance. The orientation and distribution of the building are critical to determine solar heat gain, and will affect lighting. All terms depend on climate, which directly affect window performance [7]

Building parameter is characterized by ASHRAE standard 90.1, which is energy code to set up to 40% of WWR. Under different condition WWR varies distinctively based on attempt to decide ideal WWR. During the 2013 correction of ASHRAE 90.1 recommendations to decrease the most extreme WWR to 30% met some debate and were dismissed eventually [8], but WWR remains an important topic in building operation and design.

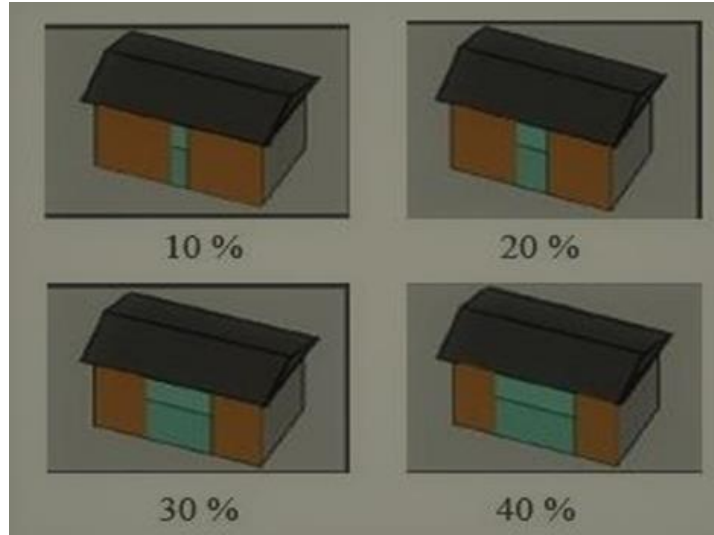
## **BUILDING STUDY**

Test buildings are located in Ranchi (23.344°N, 85.3096°E), India. India, a country in which there are six different types of climatic zone starting from montane, humid, subtropical, tropical (wet and dry), tropical wet and dry. The climatic condition of Ranchi is humid subtopic zone. In humid subtopic climate, the climate is characterized by hot and humid summers, and cold to mild winters. This type of climate normally lies on the southeast side of India.

### **Window Wall Ratio (WWR)**

Window Wall Ratio (WWR) has an impact on energy saving in buildings. It is the ratio of glazing area ( $S_g$ ) to facade area ( $S_f$ ) of the building (Wang et al.). WWRs vary from building to building and it varies from 0% to 70% (Sayadi et al.). WWR depends on the site and the orientation of the building. The high percentage of WWR provides higher thermal energy and

lightening as compared to the lower percentage and vice- versa. Three different building designs of four different WWRs (10%, 20%, 30%, and 40%) have been considered for the present study (Fig. 1).



**FIGURE 1.** Building design of 4 different WWR

## NUMERICAL ANALYSIS

### Energy payback time (EPBT)

Energy payback time is defined as the period required for a renewable energy system to generate the same amount of energy that was used to produce the system itself. It is calculated as shown in the equation below [9-10]: The emission of average CO<sub>2</sub> is approximately equivalent to 0.98 kg of CO<sub>2</sub>/kWh in the case of electricity generated by coal.

$$EPT = \frac{E_{mat} + E_{manuf} + E_{inst} + E_{trans} + E_{EOL}}{E_{agen} \times E_{aoper}}$$

### Embodied energy

The total energy required to produce any items, things or services is called embodied energy [11-12]. The embodied energy of the various materials in the fabrication is presented in Table of all buildings for different WWR [13-18]. Table 1, explains about the embodied CO<sub>2</sub> for 10% WWR. In this table material embodied carbon and inventory, area, embodied carbon, equivalent CO<sub>2</sub> and mass of the different materials are mentioned. Table 2, gives the construction embodied carbon and inventory. In this table material embodied carbon and inventory, area, embodied carbon and equivalent CO<sub>2</sub> are mentioned and Table 3, explains about the glazing embodied carbon and inventory. In this table material embodied carbon and inventory, area, embodied carbon and equivalent CO<sub>2</sub> are mentioned.

Material in the embodied energy has an indirect and direct element. The major indirect element represents the energy for extraction and transportation of material used. Raw material could be industrial waste by-products, manufactured products, natural materials, reused and recycled material. Table 4, explains about the embodied CO<sub>2</sub> for 20% WWR. In this table material embodied carbon and inventory, area, embodied carbon, equivalent CO<sub>2</sub> and mass of the different materials are mentioned. Table 5, gives the constructions embodied carbon and inventory. In this table material embodied carbon and inventory, area, embodied carbon and equivalent CO<sub>2</sub> are mentioned. Table 6, explains the glazing embodied carbon and inventory. In

this table material embodied carbon and inventory, area, embodied carbon and equivalent CO<sub>2</sub> are mentioned. Table 7, explains embodied CO<sub>2</sub> for 30% WWR. In this table material embodied carbon and inventory, area, embodied carbon, equivalent CO<sub>2</sub> and mass of the different material are mentioned. Table 8 gives the constructions embodied carbon and inventory. In this table material embodied carbon and inventory, area, embodied carbon and equivalent CO<sub>2</sub> are mentioned. Table 9, explains the glazing embodied carbon and inventory for 30% WWR. In this table material embodied carbon and inventory, area, embodied carbon and equivalent CO<sub>2</sub> are mentioned. Table 10, explains about the embodied CO<sub>2</sub> for 40% WWR. In this table material embodied carbon and inventory, area, embodied carbon, equivalent CO<sub>2</sub> and mass of the different material are mentioned. Table 11, gives the constructions embodied carbon and inventory. In this table material embodied carbon and inventory, area, embodied carbon and equivalent CO<sub>2</sub> are mentioned. Table 12, explains the glazing embodied carbon and inventory. In this table material embodied carbon and inventory, area, embodied carbon and equivalent CO<sub>2</sub> are mentioned.

**TABLE 1. Embodied CO<sub>2</sub> of Building 1 for 10% WWR**

<b>Materials Embodied Carbon and Inventory</b>	<b>Area (m<sup>2</sup>)</b>	<b>Embodied Carbon (kgCO<sub>2</sub>)</b>	<b>Equivalent CO<sub>2</sub> (kgCO<sub>2</sub>)</b>	<b>Mass (kg)</b>
Timber Flooring	221.9	1990.7	2034	4327.7
Floor/Roof Screed	221.9	2982.8	2982.8	18642.3
Plasterboard	221.9	3069.8	3231.3	8078.3
Gypsum Plastering	190.3	940.2	989.7	2474.1
Gypsum Plasterboard	93.4	252.3	273.3	2102.2
Urea Formaldehyde Foam	221.9	524.2	565.4	294.5
MW Glass Wool (rolls)	221.9	588.8	646.5	384.8
XPS Extruded Polystyrene -CO <sub>2</sub> Blowing	190.3	1525.1	5073.2	529.6
Concrete Block (Medium)	190.3	2131.6	2131.6	26644.5
Cast Concrete	221.9	3550.9	3550.9	44386.3
Brickwork Outer Leaf	190.3	7117.9	7441.4	32354
Asphalt	221.9	233	233	4660.6
Sub Total		24907.2	29153.1	144879

**TABLE 2. Construction Embodied Carbon and Inventory for 10% WWR**

<b>Materials Embodied Carbon and Inventory</b>	<b>Area (m<sup>2</sup>)</b>	<b>Embodied Carbon (kgCO<sub>2</sub>)</b>	<b>Equivalent CO<sub>2</sub> (kgCO<sub>2</sub>)</b>
Project external glazing	20.1	375.5	375.5
Sub Total	701	25282.7	29528.6

**TABLE 3. Glazing Embodied Carbon and Inventory for 10% WWR**

<b>Materials Embodied Carbon and Inventory</b>	<b>Area(m<sup>2</sup>)</b>	<b>Embodied Carbon (kgCO<sub>2</sub>)</b>	<b>Equivalent CO<sub>2</sub> (kgCO<sub>2</sub>)</b>
Area (m <sup>2</sup> )	190.3	11714.8	15635.8
Embodied Carbon (kg CO <sub>2</sub> )	46.7	252.3	273.3
Equivalent CO <sub>2</sub> (kg CO <sub>2</sub> )	221.9	3891.6	4110.9
Project ground floor	221.9	9048.6	9133.1
Sub Total	680.9	24907.21	29153.1

**TABLE 4. Embodied CO<sub>2</sub> of Building 1 for 20% WWR**

<b>Materials Embodied Carbon and Inventory</b>	<b>Area (m<sup>2</sup>)</b>	<b>Embodied Carbon (kgCO<sub>2</sub>)</b>	<b>Equivalent CO<sub>2</sub> (kgCO<sub>2</sub>)</b>	<b>Mass (kg)</b>
Timber Flooring	221.9	1990.7	2034	4327.7
Floor/Roof Screed	221.9	2982.8	2982.8	18642.3
Plasterboard	221.9	3069.8	3231.3	8078.3
Gypsum Plastering	170.2	841	885.2	2213.1
Gypsum Plasterboard	93.4	252.3	273.3	2102.2
Urea Formaldehyde Foam	221.9	524.2	565.4	294.5
MW Glass Wool (rolls)	221.9	588.8	646.5	384.8
XPS Extruded Polystyrene - CO <sub>2</sub> Blowing	170.2	1364.2	4537.9	473.7
Concrete Block (Medium)	170.2	1906.6	1906.6	23833.1
Cast Concrete	221.9	3550.9	3550.9	44386.3
Brickwork Outer Leaf	170.2	6366.8	6656.2	28940.2
Asphalt	221.9	233	233	4660.6
Sub Total		23671.1	27503.3	138337



**TABLE 5. Constructions Embodied Carbon and Inventory for 20% WWR**

<b>Materials Embodied Carbon and Inventory</b>	<b>Area (m<sup>2</sup>)</b>	<b>Embodied Carbon (kgCO<sub>2</sub>)</b>	<b>Equivalent CO<sub>2</sub> (kgCO<sub>2</sub>)</b>
Area (m <sup>2</sup> )	170.2	10478.7	13986
Embodied Carbon (kgCO <sub>2</sub> )	46.7	252.3	273.3
Equivalent CO <sub>2</sub> (kgCO <sub>2</sub> )	221.9	3891.6	4110.9
Project ground floor	221.9	9048.6	9133.1
Sub Total	660.8	23671.11	27503.26

**TABLE 6. Glazing Embodied Carbon and Inventory for 20% WWR**

<b>Materials Embodied Carbon and Inventory</b>	<b>Area (m<sup>2</sup>)</b>	<b>Embodied Carbon (kgCO<sub>2</sub>)</b>	<b>Equivalent CO<sub>2</sub> (kgCO<sub>2</sub>)</b>
Project external glazing	40.2	751.1	751.1
Sub Total	701	24422.2	28254.3

**TABLE 7. Embodied CO<sub>2</sub> of Building 1 for 30% WWR**

<b>Materials Embodied Carbon and Inventory</b>	<b>Area (m<sup>2</sup>)</b>	<b>Embodied Carbon (kgCO<sub>2</sub>)</b>	<b>Equivalent CO<sub>2</sub> (kgCO<sub>2</sub>)</b>	<b>Mass (kg)</b>
Timber Flooring	221.9	1990.7	2034	4327.7
Floor/Roof Screed	221.9	2982.8	2982.8	18642.3
Plasterboard	221.9	3069.8	3231.3	8078.3
Gypsum Plastering	150.2	741.8	780.8	1952
Gypsum Plasterboard	93.4	252.3	273.3	2102.2
Urea Formaldehyde Foam	221.9	524.2	565.4	294.5
MW Glass Wool (rolls)	221.9	588.8	646.5	384.8
XPS Extruded Polystyrene - CO <sub>2</sub> Blowing	150.2	1203.3	4002.6	417.8
Concrete Block (Medium)	150.2	1681.7	1681.7	21021.6
Cast Concrete	221.9	3550.9	3550.9	44386.3
Brickwork Outer Leaf	150.2	5615.8	5871	25526.2
Asphalt	221.9	233	233	4660.6
Sub Total		22435	25853.4	131794

**TABLE 8. Constructions Embodied Carbon and Inventory for 30% WWR**

<b>Materials Embodied Carbon and Inventory</b>	<b>Area (m<sup>2</sup>)</b>	<b>Embodied Carbon (kgCO<sub>2</sub>)</b>	<b>Equivalent CO<sub>2</sub> (kgCO<sub>2</sub>)</b>
Area (m <sup>2</sup> )	150.2	9242.5	12336.1
Embodied Carbon (kgCO <sub>2</sub> )	46.7	252.3	273.3
Equivalent CO <sub>2</sub> (kgCO <sub>2</sub> )	221.9	3891.6	4110.9
Project ground floor	221.9	9048.6	9133.1
Sub Total	640.7	22435	25853.41

**TABLE 9. Glazing Embodied Carbon and Inventory for 30% WWR**

<b>Materials Embodied Carbon and Inventory</b>	<b>Area (m<sup>2</sup>)</b>	<b>Embodied Carbon (kgCO<sub>2</sub>)</b>	<b>Equivalent CO<sub>2</sub> (kgCO<sub>2</sub>)</b>
Project external glazing	60.2	1126.6	1126.6
Sub Total	701	23561.6	26980

**TABLE 10. Embodied CO<sub>2</sub> of Building 1 for 40% WWR**

<b>Materials Embodied Carbon and Inventory</b>	<b>Area (m<sup>2</sup>)</b>	<b>Embodied Carbon (kgCO<sub>2</sub>)</b>	<b>Equivalent CO<sub>2</sub> (kgCO<sub>2</sub>)</b>	<b>Mass (kg)</b>
Timber Flooring	221.9	1990.7	2034	4327.7
Floor/Roof Screed	221.9	2982.8	2982.8	18642.3
Plasterboard	221.9	3069.8	3231.3	8078.3
Gypsum Plastering	130.1	642.6	676.4	1690.9
Gypsum Plasterboard	93.4	252.3	273.3	2102.2
Urea Formaldehyde Foam	221.9	524.2	565.4	294.5
MW Glass Wool (rolls)	221.9	588.8	646.5	384.8
XPS Extruded Polystyrene -CO <sub>2</sub> Blowing	130.1	1042.3	3467.3	361.9
Concrete Block (Medium)	130.1	1456.8	1456.8	18210.1
Cast Concrete	221.9	3550.9	3550.9	44386.3
Brickwork Outer Leaf	130.1	4864.7	5085.8	22112.3
Asphalt	221.9	233	233	4660.6
Sub Total		21198.9	24203.6	125252

**TABLE 11. Constructions Embodied Carbon and Inventory for 40% WWR**

<b>Materials Embodied Carbon and Inventory</b>	<b>Area (m<sup>2</sup>)</b>	<b>Embodied Carbon (kgCO<sub>2</sub>)</b>	<b>Equivalent CO<sub>2</sub> (kgCO<sub>2</sub>)</b>
Area (m <sup>2</sup> )	130.1	8006.4	10686.3
Embodied Carbon (kgCO <sub>2</sub> )	46.7	252.3	273.3
Equivalent CO <sub>2</sub> (kgCO <sub>2</sub> )	221.9	3891.6	4110.9
Project ground floor	221.9	9048.6	9133.1
Sub Total	620.7	21198.88	24203.55

**TABLE 12. Glazing Embodied Carbon and Inventory for 40% WWR**

<b>Materials Embodied Carbon and Inventory</b>	<b>Area (m<sup>2</sup>)</b>	<b>Embodied Carbon (kgCO<sub>2</sub>)</b>	<b>Equivalent CO<sub>2</sub> (kgCO<sub>2</sub>)</b>
Project external glazing	80.3	1502.1	1502.1
Sub Total	701	22701	25705.7

### **Illuminance factor**

Illuminance values of the test buildings are mentioned for Building 1 in Fig. 2 on 21<sup>st</sup> June and Fig. 3 on 21<sup>st</sup> December. For Building 2 in Fig. 4 on 21<sup>st</sup> June and Fig. 5 on 21<sup>st</sup> December. For Building 3 in Fig. 6 on 21<sup>st</sup> June and Fig. 7 on 21<sup>st</sup> December.

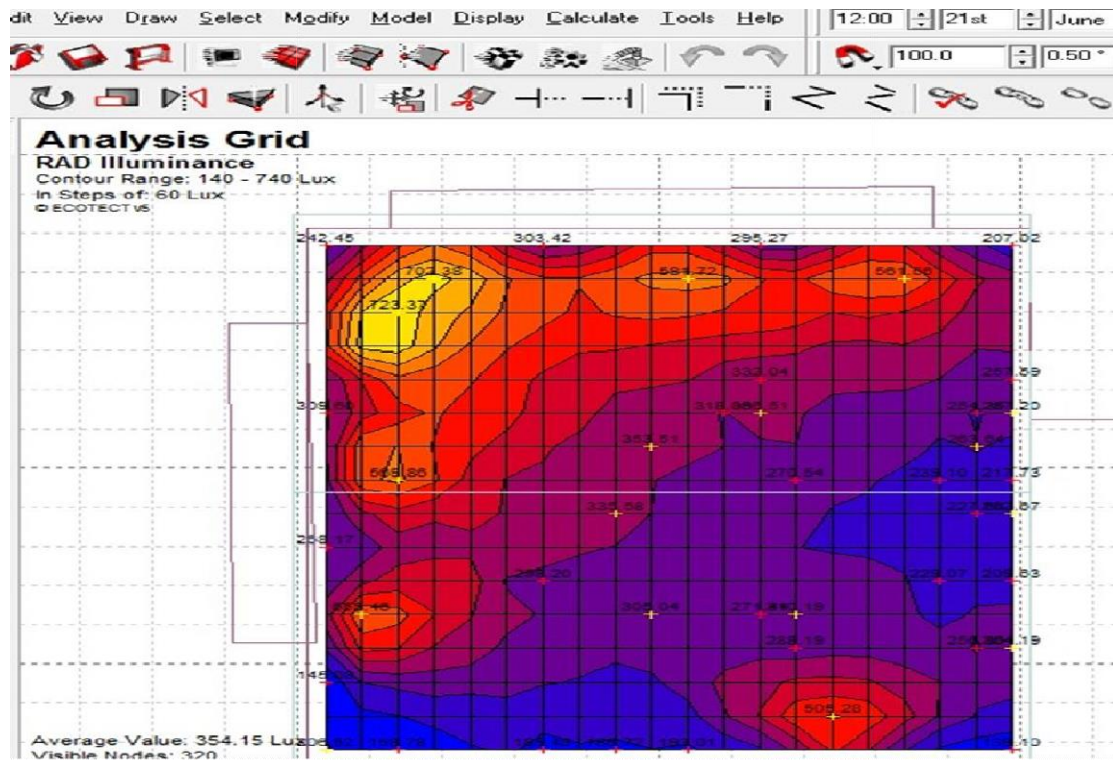


FIGURE 2. Building 1 Illuminance analysis on 21<sup>st</sup> June

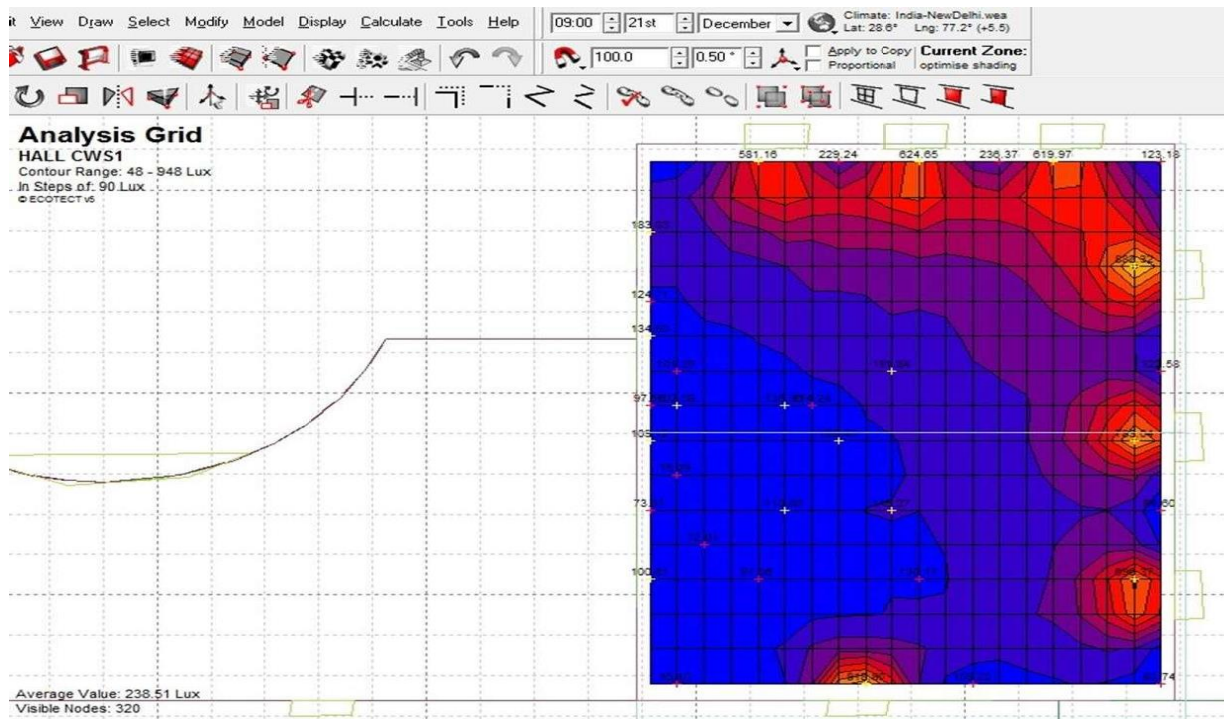


FIGURE 3. Building 1 Illuminance analysis on 21<sup>st</sup> December



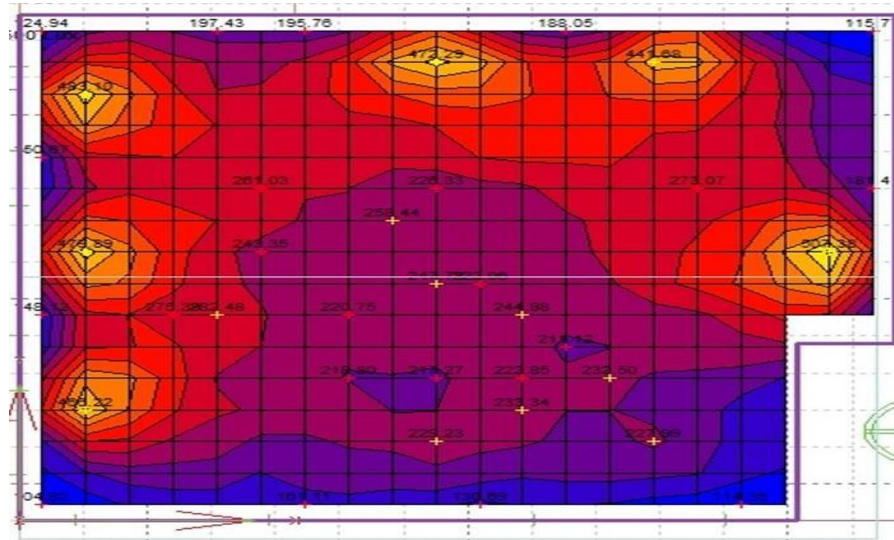


FIGURE 4. Building 2 Illuminance analysis on 21<sup>st</sup> June

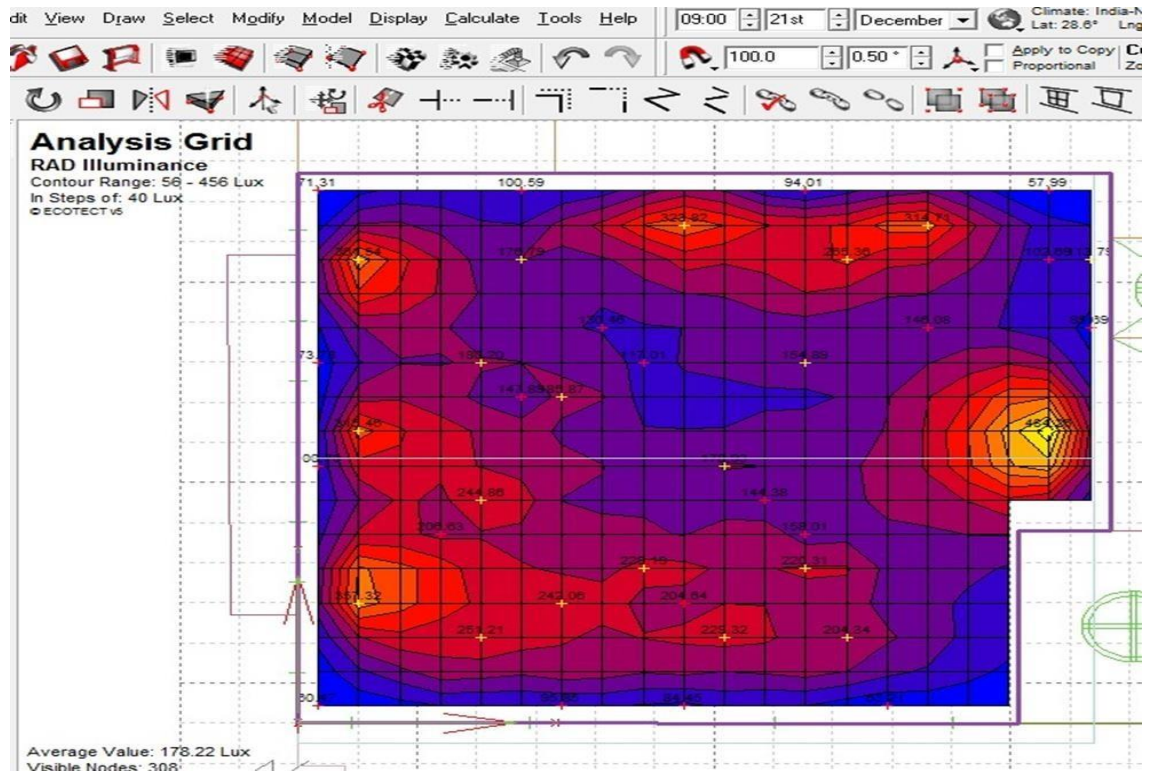


FIGURE 5. Building 2 Illuminance analysis on 21<sup>st</sup> December



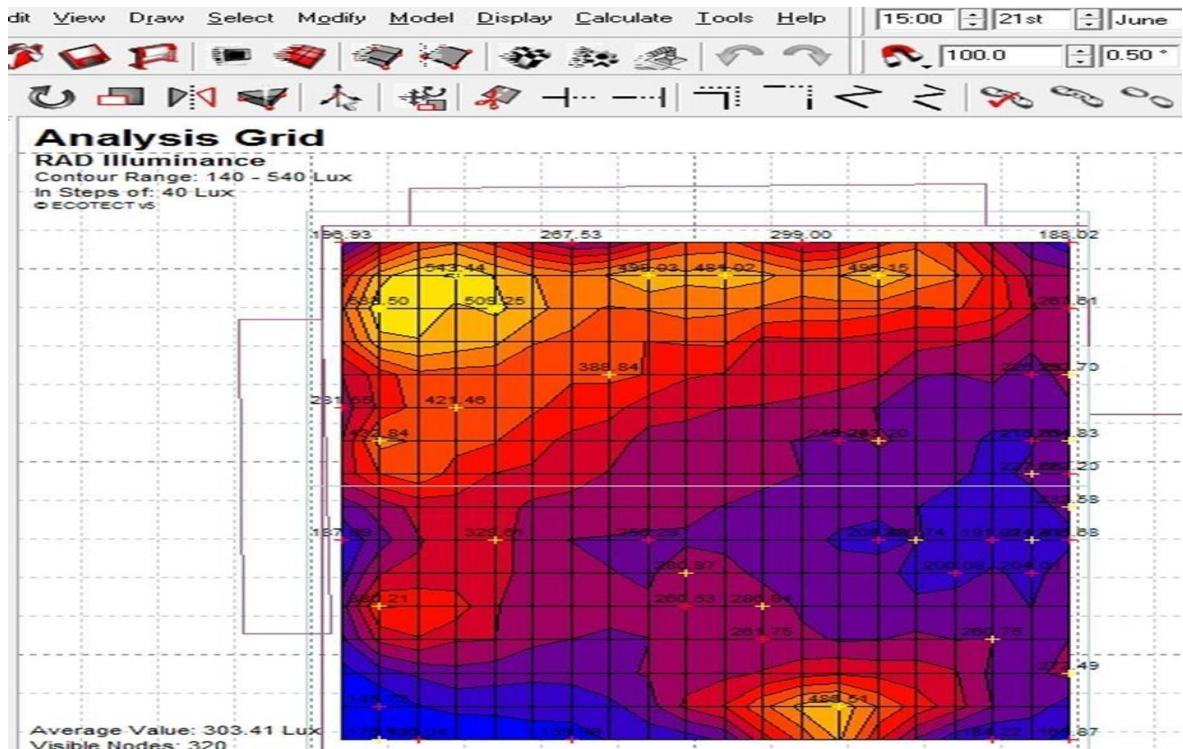


FIGURE 6. Building 3 Illuminance analysis on 21<sup>st</sup> June

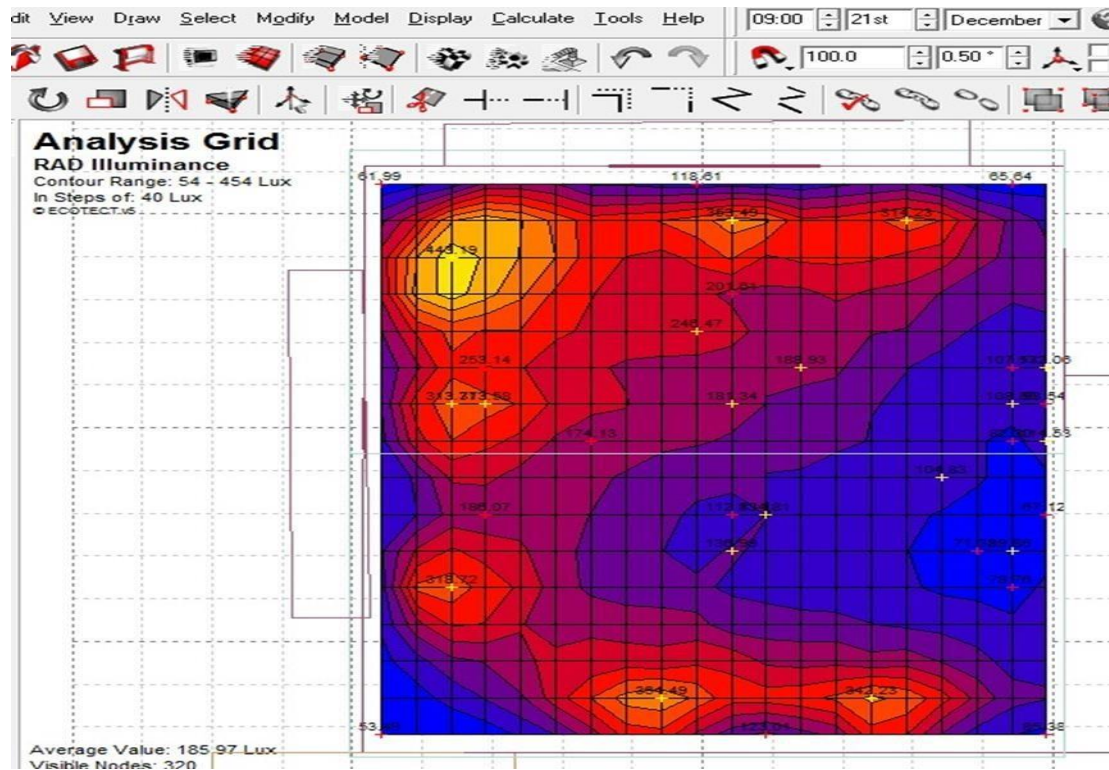


FIGURE 7. Building 3 Illuminance analysis on 21<sup>st</sup> December

## CONCLUSION

Attempts to reduce or replace conventional high-energy materials such as cement, steel, and bricks with less expensive and locally available alternatives will result in a reduction in the embodied energy in buildings. The structure of building fundamentals depends on thermal comfort and energy performance. In this study, building simulation results are utilized to study direct building envelope courses. The thermal mass and the radiative properties of the building is the primary factor when deciding on the best building material. Buildings in urban cities require a significant reduction of heating and cooling loads which are a primary challenge in the composite climate. For thermal comfort, the thermal storage wall is selected to have standard void squares brick with external wooden fiber insulation.

## REFERENCES

1. Omer A.M.,2008. Energy, environment and sustainable development. *Renewable and sustainable energy reviews*, 12(9), pp.2265-2300.
2. Kibert C.J., 2016. *Sustainable construction: green building design and delivery*. John Wiley & Sons.
3. Thewes A., Maas S., Scholzen F., Waldmann D. and Zürbes A., 2014. Field study on the energy consumption of school buildings in Luxembourg. *Energy and Buildings*, 68,pp.460-470.
4. Ding Z., Fan Z., Tam V.W., Bian Y., Li S., Illankoon, I.C.S. and Moon S.,2018.Green building evaluation system implementation. *Building and Environment*, 133, pp.32-40.
5. Prakash O., Ahmad A., Kumar A., Hasnain S.M.M, Zare A., and Verma P.,2021.Thermal performance and energy consumption analysis of retail buildings through daylighting: A numerical model with experimental validation. *Materials Science for Energy Technologies*, 4, pp.367-382. <https://doi.org/10.1016/j.mset.2021.08.008>
6. Phillips R., Troup L., Fannon D., Eckelman M. J., 2020. Triple bottom line sustainability assessment of window-to-wall ratio in US office buildings. *Building and Environment*, 182, 107057. DOI: 10.1016/j.buildenv.
7. Ahmad A., Kumar A., Prakash O., Aman A., 2020. Daylight availability assessment and application of energy simulation software-A literature review. *Material Science and Energy Technology,Elsevier*,3 ,pp.679-689, <https://doi.org/10.1016/j.mset.2020.07.002>
8. Luke T., Phillips R, Matthew J. Eckelman, Fannon D.,2019.Effect of window to wallratio on measured energy consumption in US office buildings. *Energy and Buildings*, 203, pp.109434. <https://doi.org/10.1016/j.enbuild.2019.109434>
9. Ahmad A., Prakash O., Kumar A., Hasnain S.M.M., Verma P., Zare A., Dwivedi G., Pandey A.,2022.Dynamic analysis of daylight factor, thermal comfort and energy performance under clear sky conditions for building: An experimental validation. *Material Science and Energy Technology*, 5, pp.52-65.
10. Balaban O., Oliveira D., J.A.P.,2017. Sustainable buildings for healthier cities: assessing the co-benefits of green buildings in Japan. *Journal of Cleaner Production*, 163,pp.S68-S78.
11. Prakash O., Dogra R., Salik M., Khanara A. 2019. Daylight Factor: A Simulation Approach on daylighting of a hostel building Room. National Conference on Recent Advancement in Mechanical Engineering. Organized at Department of Mechanical Engineering, BIT Mesra, Ranchi on 18- 19<sup>th</sup> Feb 2019.
12. Yonghe w., Wang R., Gaomei Li, Peng C.,2020. An investigation of optimal window-to-wall ratio based on changes in building orientations for traditional dwellings. *Solar Energy* 195, pp.64-81. <https://doi.org/10.1016/j.solener.2019.11.033>
13. Damico A., Ciulla G., Panno D., Ferrari S., 2019.Building energy demand assessment through heating degree days: The importance of a climatic dataset. *Applied energy*, 242 ,pp.1285-1306. <https://doi.org/10.1016/j.apenergy.2019.03.167>
14. Darko E., Nagrath K., Niaizi Z.,Scott A.,Varsha D. Vijaya K., 2013. Green building: case

study. *Shaping policy for Development. Overseas Development Institute, London.*

15. Meral O., 2019. Influence of glazing area on optimum thickness of insulation for different wall orientations. *Applied Thermal Engineering*, 147, pp.770-780.  
<https://doi.org/10.1016/j.applthermaleng.2018.10.089>
16. Patrick S., Tarantino S., Martin F., 2018, Parametric analysis of design stage building energy performance simulation models. *Energy and Buildings*, 172, 2018, pp.78-93.  
<https://doi.org/10.1016/j.enbuild.2018.04.045>
17. Sormunen P. and Kärki T. 2019. Recycled construction and demolition waste as a possible source of materials for composite manufacturing. *Journal of Building Engineering*, 24, pp.100742.
18. Yonghe W., Wang R., Gaomei Li, and Peng C., 2020. An investigation of optimal window-to-wall ratio based on changes in building orientations for traditional dwellings. *Solar Energy*, 195, pp.64-81. <https://doi.org/10.1016/j.solener.2019.11.033>.

RESEARCH ARTICLE | SEPTEMBER 05 2023

## Environomical analysis of sensible heat storage-based greenhouse dryer **FREE**

Asim Ahmad ✉; Om Prakash; Anil Kumar; Md Shahnawaz Hussain



AIP Conf. Proc. 2863, 020028 (2023)

<https://doi.org/10.1063/5.0155711>



CrossMark

### Articles You May Be Interested In

Environomical analysis of green building having various window-to-wall ratio

AIP Conf. Proc. (September 2023)

Thermal behavior modeling of a cabinet direct solar dryer as influenced by sensible heat storage in a fractured porous medium

AIP Conference Proceedings (May 2018)

Experimental investigations of high efficiency smart solar tunnel dryer system

AIP Conference Proceedings (October 2020)

500 kHz or 8.5 GHz?  
And all the ranges in between.

Lock-in Amplifiers for your periodic signal measurements



Find out more



# Environomical Analysis of Sensible Heat Storage-Based Greenhouse Dryer

Asim Ahmad<sup>1, a)</sup>, Om Prakash<sup>2, b)</sup>, Anil Kumar<sup>3, c)</sup> and Md Shahnawaz Hussain<sup>4, d)</sup>

<sup>1</sup> Department of Mechanical Engineering, Usha Martin University, Ranchi-835103, India

<sup>2</sup> Department of Mechanical Engineering, Birla Institute of Technology, Mesra, Ranchi-835215, India

<sup>3</sup> Department of Mechanical Engineering, Delhi Technological University, Delhi-110042, India

<sup>4</sup> Department of Electronics and Communications Engineering, Birla Institute of Technology Mesra, Ranchi-835215, India

<sup>a)</sup> Corresponding author: asimlife91@gmail.com

<sup>b)</sup> omprakash@bitmesra.ac.in

<sup>c)</sup> anilkumar76@dtu.ac.in

<sup>d)</sup> mshane87@gmail.com

**Abstract.** The demand for fossil fuels has increased to a larger extent in the past few years. So the prime concern is to move toward renewable energy. Solar energy is one of the forms of renewable energy and the solar drying method is one of the forms of use of this energy. Solar drying method not only reduces the consumption of fossil fuel but also saves the crops from post-harvest loss. In the present study, various environomical parameters have been analyzed such as energy analysis, embodied analysis, and CO<sub>2</sub> emission rate for the proposed system. The embodied analysis for the proposed setup of the given system is found to be 530.4976 kWh.

**Keywords:** Greenhouse Dryer, CO<sub>2</sub> Emission, Energy Analysis, Embodied Energy Analysis.

## INTRODUCTION

The agricultural production level of India can be increased, if the farmer utilizes more and more renewable energy technology to cope with their demands. The wastage increased to 10-40% in the countries of the Asia Pacific region because of a poor framework for processing and marketing [1]. As the daily requirements for fruits and vegetables are not sufficient and the human development index is very low.

The techniques used for food preservation are purely traditional like a process of refrigeration, freezing, smoking, drying, salting (marinate), miming such things as sugar, bolting, and packing in cane.

In developing countries, drying is the most appropriate technique which is an economical and profitable means of preservation to reduce post-harvest losses and standardize the shortage in supply [2].

Heat energy is utilized in the process of dehydration which is known as drying. By the application of electrical energy or by burning fossil fuels or by solar radiation conversion into heat, a necessary amount of heat can be produced [3]. Drying is the chief and the most applicable resource for the implementation of solar energy as it minimizes the utilization of non-renewable energy by 30-80 % [4]. It also utilizes low-temperature heating for food drying. Currently, researchers have a reasonable interest in the field of modeling, progression, and testing of drying modes like direct,



greenhouse indirect, and mixed-mode [5]. There is a no. of models and details of their progression and operating principles of a huge variety of solar drying systems are represented by some researchers. Solar dryers are primarily classified in two categories as solar active and passive dryers, and their classification depends on their organization of various components and modes of heat transfer [6-8].

The best example of a direct type of solar dryer is a greenhouse dryer in which the cultivated crops are dried in a very big quantity. For small-scale farmers, a greenhouse dryer is the most appropriate for low-temperature drying.

For the greenhouse dryer model and observations, a large no of research has been carried out. The chief ultimatum for the most demanding greenhouse dryer is to stop incident solar radiation losses from the north wall, Hence the wall must be blurred. A large no of researchers initiated plans to reduce the losses from the north wall of the greenhouse dryer [9-12].

By resolving the insulated blurred north wall in the existing greenhouse drying system, the heat losses and solar radiation stopped completely in the present study.

The most necessary parameter in the greenhouse dryer is the convective heat transfer coefficient since the temperature difference between the air and ground changes with this coefficient [13-15].

All the experimental observations have been carried out in sensible heat storage conditions such as the ground floor, concrete floor, gravel floor, and black painted gravel floor. It has been found that very few researchers have calculated the parameters like embodied and energy analysis. Hence in this paper, we analyzed the parameters of embodied and energy analysis of greenhouse dryer with an insulated north wall in no-load condition operating under passive mode.

## EXPERIMENTAL PROCEDURE & INSTRUMENTATION

The experiments were carried out in Solar Energy Lab at the Department of Mechanical Engineering, Birla Institute of Technology, Mesra, Ranchi having a latitude of 23.340 N and a longitude of 77.250 E. The set up consist of a greenhouse dryer with dimension 1m x 1.5m x 0.5m while the center height is 0.75m as shown in Fig. 1. The setup is made up of an acrylic sheet of a thickness of 1 mm and is further supported by an Aluminum structure. The experiments were carried out at different sensible heat storage conditions like the ground floor, concrete floor, gravel floor, and black painted gravel floor. The north wall of the dryer is insulated with black paint from the outside and a mirror from the inside so the maximum amount of solar energy get absorbs inside the chamber. The instruments used to measure different parameters are the solar power meter (TENMAR TM-207), HTC infrared thermometer, and HTC anemometer. These instruments were used to measure solar radiation, room and ground temperature, wind velocity, and relative humidity.



(a)



(b)

**FIGURE 1.** (a) Passive Greenhouse dryer (b) Insulated North wall



## NUMERICAL ANALYSIS

### Energy Analysis

Due to prompt rise in fuel price and raw material it becomes a major concern to focus on this issue [16-17]. Energy analysis consists of cost analysis, CO<sub>2</sub> emission per year, payback period, mitigation & credit of carbon. Energy analysis also helps us to find out the rate of solar energy received by solar heater & accessible for the dryer. Energy analysis can also be explained based on mass and energy equation in steady state [18].

Energy received by the absorber is determined by:

$$Q_{ad} = m_{ad} C_{pad} (T_{acod} - T_{acin}) \quad (1)$$

Thermal efficiency for different bed conditions can be calculated as:

$$\eta = \frac{m_{ad} C_{pad} (T_{acod} - T_{acin})}{A_c I} \quad (2)$$

### Embodied Analysis

Embodied energy is the whole quantity of energy utilized to produce a good or to carry out a task. At the source, coal energy production has an average CO<sub>2</sub> equivalent intensity of 0.98 kg/kWh [19]. CO<sub>2</sub> emissions are expressed as follows in a kilogram of CO<sub>2</sub>/year:

$$CO_2 \text{ emission per year} = \frac{E_{embodied} \times 0.98}{L.T} \quad (3)$$

If domestic losses for India are assumed to be 20% and transmission and supply losses are assumed to be 40%, then this value increases from 0.98 to 1.58 and a new equation is presented as [20]:

$$CO_2 \text{ emission per year} = \frac{E_{embodied} \times 1.58}{L.T} \quad (4)$$

Therefore, the indirect solar dryer overall CO<sub>2</sub> gas emission over its lifetime:

$$Total CO_2 \text{ gas emission} = E_m \times 1.58 \quad (5)$$

## RESULT & DISCUSSION

### Energy Analysis at different sensible heat storage

The dryer efficiency is calculated by calculating the dryer thermal energy storage. According to observations, the black painted gravel floor has the highest dryer efficiency, followed by the gravel floor, the concrete floor, and the ground floor. It has been observed that the maximum efficiency has been calculated at the time of 15 IST or 16 IST (Fig.2). It is because at that time the heat absorbed by the floor has reached to maximum due to increased time duration. It has also been observed that the lower efficiency is at 10 IST. The gravel floor with the black paint has the highest efficiency at 24.63 %, followed by the concrete floor and the ground floor at 22.73 %, 13.87 %, and 9.33 %. The ground level, concrete floor, and gravel floor, in that order, have minimum efficiencies of 20 %, 19 %, 16 %, and 14 %, respectively. This kind of sensible heat storage condition is highly advised since the black-coated gravel floor has been designed to have maximum efficiency.

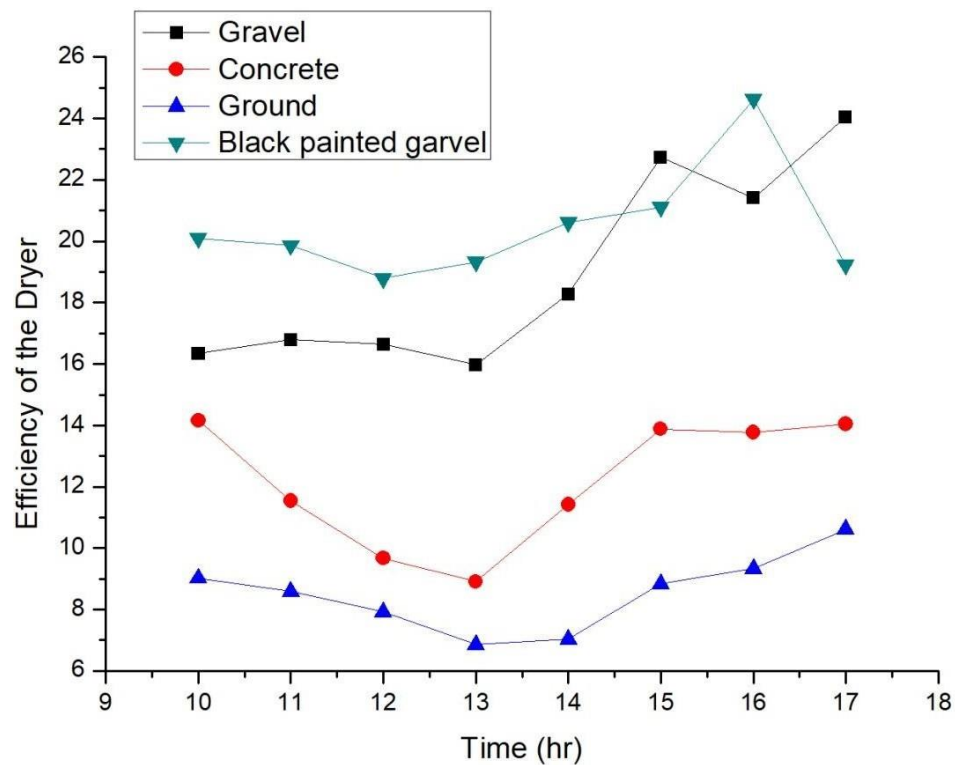


FIGURE 2. Drying efficiency of sensible heat storage-based greenhouse dryer

## CO<sub>2</sub> emission and Embodied energy Analysis

The proposed system embodied energy is measured as 530.4976 kWh. While CO<sub>2</sub> emissions per year are calculated to be 14.85 kg when power generation from coal is 0.98 kg when domestic losses and distribution losses are factored in, coal consumption is raised by 1.58 kg, and CO<sub>2</sub> emissions are calculated to be 23.94 kg. The embodied energy is found to be the same for all sensible heat storage conditions since varied sensible heat storage conditions do not differentiate the material required in the construction of the dryer. As a result, the CO<sub>2</sub> emission rate for varied sensible heat storage conditions is determined to be the same. The embodied energy of the suggested configuration is significantly lower than that of the indirect mode of the solar dryer and the active conventional dryer. Table 1 shows the calculated embodied energy.

**TABLE 1.** Embodied energy of greenhouse dryer.

S.NO.	Substance	Amount	Embodied energy coefficient (kWh/kg)	Gross (kWh)	Reference
i.	Glass	5.40 kg	7.28	39.312	[20]
ii.	Silver coating	0.75 m <sup>2</sup>	0.278	0.2085	
iii.	Polycarbonate sheet	15.60 kg	10.1974	159.079	
iv.	Black PVC sheet	0.325 kg	19.44	6.318	
v.	Wire mesh steel tray	0.70 kg	9.67	6.769	
vi.	Aluminium section				
		• 3.59 kg	55.28	198.455	
		• 0.82 kg	55.28	45.3296	
		• 0.08 kg	55.28	4.4224	
vii.	Fitting				
		• 0.20 kg	55.28	11.056	
		• 0.025 kg	55.28	1.382	
		• 0.10 kg	55.28	5.528	
		• 0.25 kg	9.67	2.4175	
viii.	Paint	2 kg	25.11	50.22	
<b>Grand total (kWh)</b>			<b>530.4976</b>		

## CONCLUSION

The dryer embodied energy was determined to be 530.4976 kWh. For 1.5 kg of coal, the dryer CO<sub>2</sub> emissions are calculated to be 23.94 kg. The maximum efficiency of the dryer is found to be 24.63% for black-painted gravel under sensible heat storage conditions. By analyzing these parameters in no-load conditions, the working of the greenhouse

dryer in load conditions and its efficiency can also be analyzed. As a result, it is possible to conclude that this sort of solar dryer is environmentally beneficial and can be built for the drying of crops and vegetables also it is useful for protecting the crop's nutrients value in comparison to open sun drying.

## ACKNOWLEDGMENTS

The first author would like to thank Usha Martin University, Ranchi, India for providing financial support, and BIT Mesra, Ranchi, India for carrying out the research work at the Solar Energy Lab.

## REFERENCES

1. Sharma L., Saxena A., Maity T., 2019. Trends in the Manufacture of Coatings in the Postharvest Conservation of Fruits and Vegetables. *Polymers for Agri-Food Applications*, pp.355-375, Springer, Cham.
2. Pangavhane D.R., Sawhney R.L. and Sarsavadia P.N., 2002. Design, development and performance testing of a new natural convection solar dryer. *Energy*, 27(6), pp.579-590.
3. Ahmad A. and Prakash O., 2019. Thermal analysis of north wall insulated greenhouse dryer at different bed conditions operating under natural convection mode. *Environmental Progress & Sustainable Energy*, 38(6), pp.3257.
4. Sacilik K., 2007. Effect of drying methods on thin-layer drying characteristics of hull-less seed pumpkin (*Cucurbita pepo* L.). *Journal of food engineering*, 79(1), pp. 23-30.
5. Swami V.M., Autee A.T., Anil T.R., 2018. Experimental analysis of solar fish dryer using phase change material. *Journal of Energy Storage*, 20, pp.310-315.
6. Abubakar S., Umaru S., Anafi, F.O., Abu-Bakr A.S., Kulla D.M., 2018. Design and performance evaluation of a mixed-mode Solar Crop Dryer. *FUOYE Journal of Engineering and Technology*, 3(1), pp. 22-26.
7. Raj Kumar P., Kulanthaisami S., Raghavan G.S.V., Gariépy Y. and Orsat V. 2007. Drying kinetics of tomato slices in vacuum assisted solar and open sun drying methods. *Drying Technology*, 25(7-8), 1349-1357.
8. Abubakar S., Anafi F.O., Kaisan, M.U., Narayan S., Umar S. and Umar U.A., 2019. Comparative analyses of experimental and simulated performance of a mixed-mode solar dryer. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, pp.0954406219893394.
9. Kareem M.W., Habib K., Ruslan M.H., Saha B.B., 2017. Thermal performance study of a multi-pass solar air heating collector system for drying of Roselle (*Hibiscus sabdariffa*). *Renewable Energy*, 113, pp.281-292.
10. Abubakar S., Umar S., Kaisan M.U., Umar U.A., Ashok B. and Nanthagopal K., 2018. Development and performance comparison of mixed-mode solar crop dryers with and without thermal storage. *Renewable Energy*, 128, pp. 285-298.
11. Ahmad A., Prakash O., Kumar A., 2021. Drying kinetics and economic analysis of bitter gourd flakes drying inside hybrid greenhouse dryer. *Environmental Science and Pollution Research*, pp.1-15.
12. Chauhan P.S., Kumar A., 2016. Performance analysis of greenhouse dryer by using insulated north-wall under natural convection mode. *Energy Reports*, 2, pp.107-116.
13. Prakash O., Kumar A., 2015. Annual Performance of a Modified Greenhouse Dryer Under Passive Mode In No-Load Conditions. *International journal of green energy*, 12, pp.1091-1099.
14. Prakash O., Kumar A., 2014. Design, development, and testing of a modified greenhouse dryer under conditions of natural convection. *Heat transfer research*, 45( 5).
15. Kumar A. and Tiwari G.N., 2007. Effect of mass on convective mass transfer coefficient during open sun and greenhouse drying of onion flakes. *Journal of food engineering*, 79(4), pp.1337-1350.
16. Tiwari G.N., Kumar S. and Prakash O., 2004. Evaluation of convective mass transfer coefficient during drying of jiggery. *Journal of food engineering*, 63(2), pp.219-227.
17. Chauhan P.S., Kumar A., 2018. Thermal analysis of insulated north-wall greenhouse with solar collector under passive mode. *International Journal of Sustainable Energy*, 37(4), pp.325-339.
18. Fudholi A., Sopian K., Yazdi, Ruslan M.H., M.H., Gabbasa M., Kazem H.A., 2014. Performance analysis of solar drying system for red chili. *Solar Energy*, 99, pp.47-54.
19. El-Sebaei A.A. and Shalaby S.M., 2013. Experimental investigation of an indirect-mode forced convection solar Dryer for drying thymus and mint. *Energy Conversion and Management*, 74, pp.109-116.

20. Ahmad, A. and Prakash, O., 2020. Performance Evaluation of a Solar Greenhouse Dryer at Different Bed Conditions Under Passive Mode. [Journal of Solar Energy Engineering](#), 142(1).



ORIGINAL RESEARCH ARTICLE

# Experimental and Numerical Analysis of Residual Stresses in Similar and Dissimilar Welds of T91 and Super304H Steel Tubes

Ranjeet Kumar, Prahlad Halder, Murugaiyan Amrithalingam, N. Yuvraj, Anand Varma, Y. Ravi Kumar, Suresh Neelakantan, and Jayant Jain

Submitted: 10 June 2023 / Revised: 26 August 2023 / Accepted: 31 August 2023

Residual stress distribution and its magnitude varies across the weldment, contributing to many catastrophic failures. Moreover, it is challenging to reliably measure residual stresses, considering a particular technique. Therefore, the present investigation aims to examine residual stresses in similar (T91-T91) and dissimilar (T91-Super304H) welds before and after post-weld heat treatments (PWHT), using non-destructive methods ( $\sin^2\psi$  and  $\cos \alpha$ ) and SYSWELDS simulations. For a similar weld, the peak tensile residual stresses near to fusion line reached  $\sim 238$  MPa (as per  $\sin^2\psi$  method) and  $\sim 258$  MPa (as per  $\cos \alpha$  method), which is  $\sim 48\%$  of yield stress (520 MPa) of T91 steel. Alternatively, for the case of dissimilar welds, peak tensile residual stresses of  $\sim 518$  MPa and peak compressive residual stresses of  $\sim 290$  MPa were observed at the fusion line of the T91 side and Super304H side, respectively. Dissimilar welds show relatively high residual stresses with significant deviation across weldment due to varying thermal coefficients of expansion/contraction resulting from dissimilar metal joints. Hence, PWHTs were performed to decrease the magnitude of peak residual stresses and their deviation across weldment to enhance the life of welded joints. For instance, the peak tensile residual stresses decreased from  $\sim 258$  to  $\sim 120$  MPa after 775 °C—30 min PWHT condition in similar welds. Similarly, for dissimilar welds, post-weld normalizing and tempering (PWNT) at 1050 °C—30 min followed by 760 °C—60 min condition was found to decrease the residual stresses from  $\sim 518$  to  $\sim 70$  MPa, which is a significant reduction achieved due to austenitizing.

**Keywords** dissimilar welds, residual stress, super304H, SYSWELDS, T91

## 1. Introduction

SA213-T91 and Super304H steels are widely used as tubes/pipes in fossil and nuclear power plants due to their low thermal expansion and high thermal conductivity (Ref 1). Welding is a potential technique for joining tubes and pipes. However, welding develops residual stress across the weldments, which can be influenced by the materials and welding processes (Ref 2-5). The residual stress and associated distortion can severely affect the integrity of welded structures. Generally, fatigue properties differ for different materials; however, compressive residual stresses extend the joints' fatigue life by delaying the crack propagation rate (Ref 6).

In contrast, tensile residual stresses are highly detrimental to the joint because they promote crack growth and contribute to failures (Ref 7). As residual stresses maintain a self-equilibrating state, both kinds (i.e., tensile and compressive) of residual stresses are present across the weldments; however, their distribution across weldment is different based on the location (Ref 8). Many researchers (Ref 9-11) have studied the residual stress behavior across the weldments and found that the tensile residual stresses are present in the heat-affected zone (HAZ) and the weld metal (WM) region. Yaghi et al. (Ref 12) examined the effect of thickness on residual stresses and found that thick-walled ( $\sim 40$  mm) tubes have a different residual stress distribution than thin-walled ( $\sim 7$  mm) tubes. Considering this, peak tensile stresses on the outer surface and inner surface of thick and thin pipes, respectively was reported. Generally, a characteristic 'M' shaped distribution of residual

This invited article is part of a special topical issue of the *Journal of Materials Engineering and Performance* on Residual Stress Analysis: Measurement, Effects, and Control. The issue was organized by Rajan Bhambroo, Tenneco, Inc.; Lesley Frame, University of Connecticut; Andrew Payzant, Oak Ridge National Laboratory; and James Pineault, Proto Manufacturing on behalf of the ASM Residual Stress Technical Committee.

**Ranjeet Kumar**, Materials Engineering Division, CSIR-National Metallurgical Laboratory, Jamshedpur 831007, India; and Department of Materials Science and Engineering, Indian Institute of Technology, Delhi, New Delhi 110016, India; **Prahlad Halder**, **Anand Varma**, and **Y. Ravi Kumar**, Advanced Materials Research Laboratory, NTPC Energy Technology Research Alliance (NETRA), NTPC Ltd, Greater Noida 201306, India; **Murugaiyan Amrithalingam**, Department of Metallurgical and Materials Engineering, Indian Institute of Technology Madras, Chennai 600036, India; **N. Yuvraj**, Department of Mechanical Engineering, Delhi Technological University, Delhi 110042, India; **Suresh Neelakantan** and **Jayant Jain**, Department of Materials Science and Engineering, Indian Institute of Technology, Delhi, New Delhi 110016, India. Contact e-mails: sureshn@iitd.ac.in and jayantj@iitd.ac.in.



**Table 1** Composition (in wt.%) of SA213-T91 and Super304H alloy steel tubes with filler materials

Material	Elements, wt.%											
	C	Cr	Mo	Mn	Si	Ni	V	Nb	Ti	N	Cu	Fe
SA213-T91	0.1	9.05	0.92	0.51	0.35	0.18	0.20	0.1	...	0.06	...	Bal
Super304H	0.08	18.4	...	0.84	0.21	8.77	...	0.53	...	0.11	2.74	Bal
ER90S-B9	0.09	8.90	0.9	0.54	0.22	0.46	0.22	0.046	...	0.044	...	Bal
E9015-B9	0.09	9.0	1.1	0.6	0.2	0.8	0.2	0.05	...	0.04	...	Bal
ERNiCr-3	0.03	21	...	3.4	0.22	71	...	2.5	0.31	...	0.15	Bal

stresses is observed across similar welds, where peak tensile residual stresses can reach up to  $\sim 475$  MPa (Ref 13) and 520 MPa (Ref 9, 14, 15), respectively for SA508 steel and API X65 steel.

On the contrary, dissimilar metal welds (DMW) exhibit asymmetric and opposite residual stress distribution on both sides of WM due to different materials (Ref 16). Most of the studies concluded that the distribution of residual stresses magnitude and nature is complex and both are important to decide the life of components during application. Residual stresses across the welds can be measured using several methods. X-ray diffraction, ultrasonic, neutron diffraction, barkhausen noise are non-destructive methods, while hole drilling, incremental hole drilling, ring core, sectioning and contour methods are destructive ones. Each method has advantages and disadvantages, depending on factors such as material type, accuracy and availability. However, the x-ray diffraction method is most commonly considered (Ref 17, 18) for residual stress measurement due to accuracy, reliability, repeatability and non-destructive nature. Considering the complexity of residual stresses, tools based on numerical simulation (e.g., ANSYS, ABAQUS and SYSWELDS) can be useful. Kumar et al. (Ref 19) studied the residual stresses in a laser-welded grade 91 plate (9 mm thick) and compared it with SYSWELDS software, reporting good agreement.

The weldment microstructure of the T91 steel welds consists of a predominantly martensitic phase, which exhibits low toughness and high-strength, hardness and residual stress (Ref 20). This combination of properties is unsuitable for power plants and may result in premature failure. As a result, PWHTs are advised for microstructure stabilization, residual stress minimization and enhancing the toughness with ductility recovery. Generally, PWHTs are performed below the  $Ac_1$  (Ref 13), referred to as post-weld direct tempering (PWDT) (Ref 21). Irrespective of the PWDT treatment, a gradient in the hardness values and a heterogeneous grain size distribution remains across the weldment (Ref 22). Hence, a new PWHT (Ref 23) has been suggested as post-weld normalizing and tempering (PWNT), where normalizing is performed at 1050 °C, followed by tempering at 760 °C. PWNT was reported as superior to PWDT in terms of uniform microstructures and related mechanical properties across the weldment (Ref 24). However, the influence of PWNT on residual stress has not been reported yet for dissimilar welds (T91-Super304H). In addition, the studies of residual stresses in tube welds of T91 are limited using SYSWELDS and experimental measurements ( $\sin^2\psi$ ,  $\cos \alpha$ ). Therefore, the current investigation aims to measure the residual stresses in T91 welds using  $\sin^2\psi$  and  $\cos \alpha$  methods and validate the

measured values using SYSWELDS (Ref 25) for the as-welded condition of similar welds of T91. Moreover, effective PWHTs have been proposed to mitigate the residual stress for similar as well as dissimilar welds.

## 2. Materials and Experimental Procedure

### 2.1 Materials and Welding Procedure

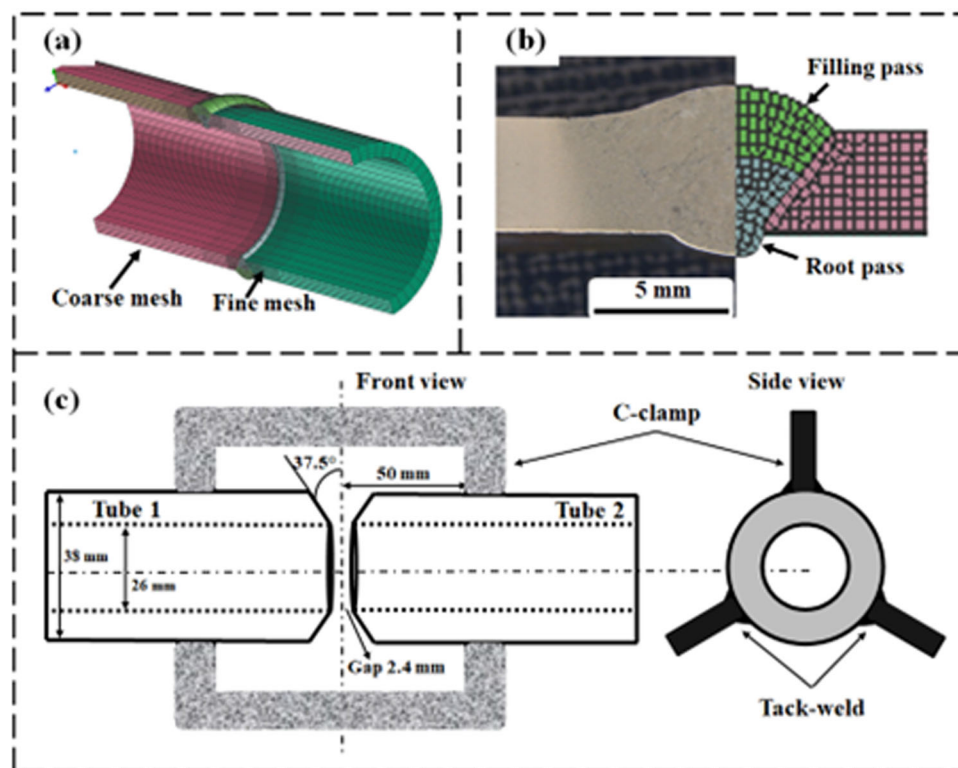
SA213-T91 and Super304H tubes of dimensions 38 mm outer diameter with 6 mm wall thickness were used for investigations. Table 1 shows the chemical composition of the initial tubes. Manual GTAW/SMAW Lincoln Electric Invertec V305-T AC/DC was used to weld 37.5° V-shaped groove tubes with a 2.4 mm root gap. C-clamps were used to assure the alignment of tubes during welding. Initially, tubes were preheated to a temperature of  $\sim 200$  °C (T91 only) using a pencil heater and monitored using thermal chalk and a pyrometer. Commercially pure argon (99.99% purity) was used for purging (inside the tubes) as well as shielding gas during root pass and filling passes. The composition of fillers for similar (ER-90S-B9 and E9015-B9 were used for root pass and filling pass, respectively), as well as dissimilar welds (ERNiCr-3), is given in Table 1. The welding parameters for similar as well as dissimilar welding are reported in Table 2, which have been chosen based on the literature (Ref 26, 27). Heat input values were calculated by considering 60% and 80% efficiency for GTAW and SMAW. To avoid martensitic transformation, 350 °C interpass temperature was maintained. After welding, samples were subjected to PWDT at 760 °C—120 min and 775 °C—30 min conditions have been considered for similar welds. For DMW, heat treatment was carried out at 760 °C for 60 min below the  $Ac_1$ , referred to as PWDT. Since DMW introduce quite higher residual stress across weldment, hence, new heat treatment was performed, austenitizing at 1050 °C for 30 min (above the  $Ac_3$ ) followed by tempering at 760 °C for 60 min, referred to as PWNT.

### 2.2 Microstructural and Mechanical Characterization

For metallography, a standard procedure, as reported in detail elsewhere (Ref 28, 29) was followed. A stereomicroscope (Leica M125C) was used to reveal the macrostructure. Field emission scanning electron microscope (FESEM) JEOL JSM 7800F was used to obtain secondary electron imaging, which was operated at 15 kV. Microhardness was measured by an instrumented microhardness tester (Zwick/Roell Z2.5) at 500 g force.

**Table 2** Welding parameters used for similar (T91-T91) and dissimilar (T91-Super304H) welds

Weld type	Weld layers	Process	Filler		Current		Voltage	Welding speed, mm/min	Heat input, kJ/mm
			Class	Dia., mm	Type	Amps.			
T91-T91	Root Pass	GTAW	ER90S-B9	2.4	DCEN	103	10-15	35	1.32
	Fill Pass	SMAW	E9015-B9	4.2	DCEP	143	21-28	50	3.36
T91-Super304H	Root Pass	GTAW	ERNiCr-3	2.4	DCEN	95	10-12	39	0.96
	Fill Pass	GTAW	ERNiCr-3	2.4	DCEN	110	12-15	72	0.74



**Fig. 1** (a) Shows elemental meshing in tubes and (b) weld bead macrostructure and corresponding meshing and (c) schematic illustrating the welding geometry with C-clamp arrangement around tubes

### 2.3 Residual Stress Measurement

Proto i-XRD and Pulstec  $\mu$ -X360, x-ray residual stress analyzer were used to measure the residual stress. Proto i-XRD is based on the  $\sin^2\psi$  and operated at 20 kV and 4 mA. Proto i-XRD was used at five tilt angles (range of  $\pm 25^\circ$ ) of incident x-ray to cover many grains (facilitate multi-exposure). On the other hand, Pulstec  $\mu$ -X360 was based on the  $\cos \alpha$  method and operated at 30 kV and 1 mA. To measure the residual stresses in T91 steel, Cr-target ( $\lambda$ : 2.291 Å) was used and lattice strain was measured at  $156^\circ$  ( $2\theta$ ) corresponding to (211) plane (BCC), while for Super304H steel, Mn target ( $\lambda$ : 2.103 Å) was used and lattice strain was measured at  $152^\circ$  ( $2\theta$ ) corresponding to (311) plane (FCC). However, due to the unavailability of the Mo target, we could not able to measure the residual stresses in WM (ERNiCr-3) of DMW. Additionally, the numerical method 'SYSWELDS' software (Ref 25) was used to simulate residual stresses for similar welds of T91. However, due to the unavailability of data for ERNiCr-3 and Super304H in SYSWELDS software, not able to simulate for DMW.

### 2.4 Finite Element Analysis: SYSWELDS

Unigraphics NX 11.0 and visual mesh 15.0 were used for modelling and meshing, respectively. The fine mesh was used in weldment and near the HAZ, while the coarse mesh was far from the HAZ region, as shown in Fig. 1(a). The total number of mesh elements and nodes were 20800 and 24846, respectively. The software package includes heat source geometry definition, moving heat source function, heat treatment and phase transformation of material (T91) with thermal, mechanical and metallurgical properties of T91. To validate the heat source, weld bead geometry was used as shown in Fig. 1(b). The boundary of the weld bead defines the temperature field around it (Ref 30), since distortion and residual stresses are more sensitive to heat and its distribution. Therefore, the most realistic and appropriate Goldak's double ellipsoidal model was considered to simulate the heat distribution (Ref 31), which is quite suitable for GTAW and SMAW types of welding (Ref 32). Boundary conditions (degree of restraint) are essential during welding as it can alter the residual stresses (Ref 33). Three C-

clamps were used at 120° apart around the tube to ensure the alignment of the tubes, clamped at 50 mm from the weld center using tack weld as shown in Fig. 1(c). The initial temperature of the tubes was considered 25 °C (ambient temperature).

### 3. Results and Discussion

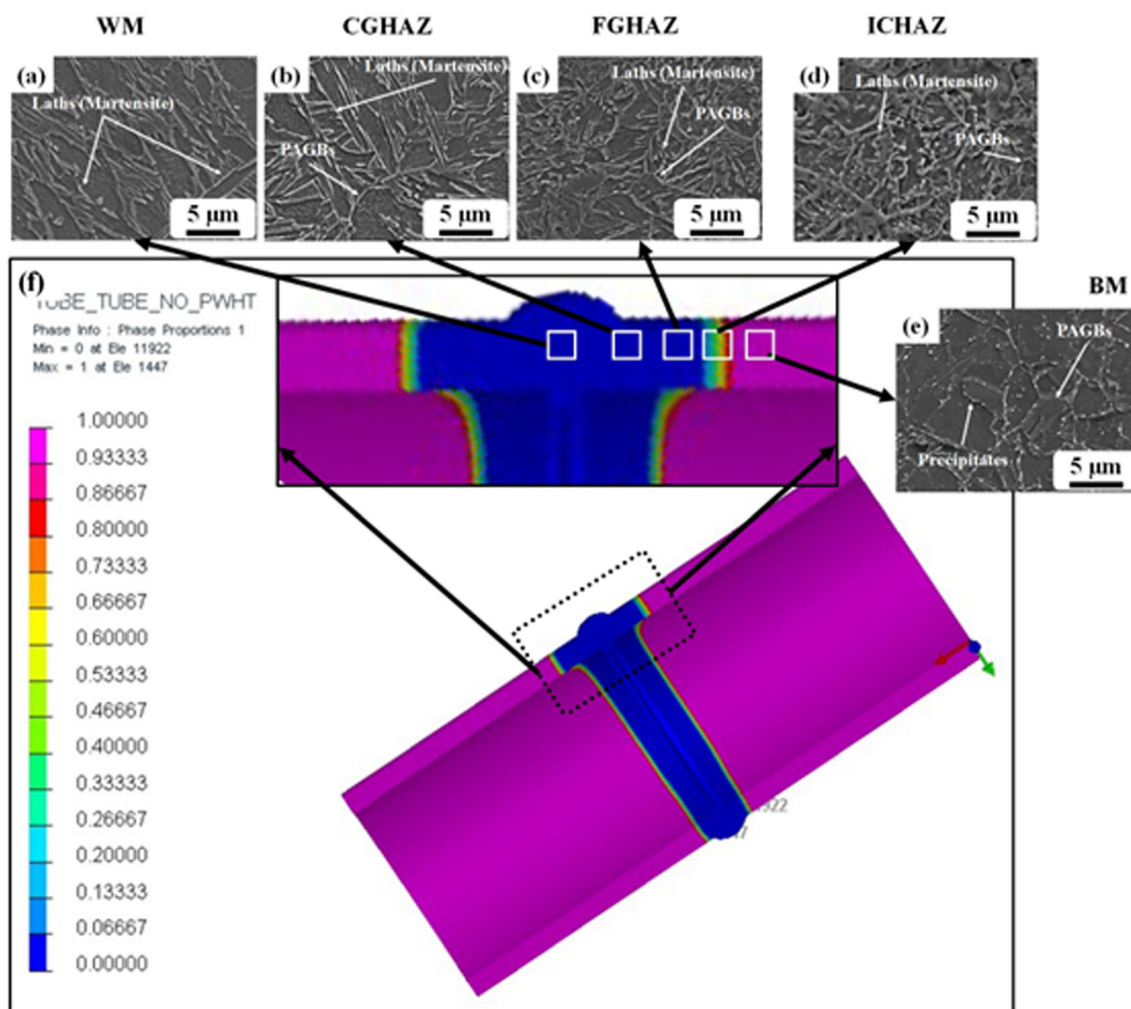
#### 3.1 Residual Stresses in As-Welded Similar Welds

Figure 2 shows the phase change across the weldments, where Fig. 2(a, b, c, d and e) show the SEM micrographs of weld metal (WM), coarse grain heat-affected zone (CGHAZ), fine grain heat-affected zone (FGHAZ), inter critical heat-affected zone (IC-HAZ) and base metal (BM). Figure 2 (f) shows the simulated phase proportion across the weldment. Note that the magenta color represents the ferrite phase and the blue color represents the martensitic phase in Fig. 2(f). It is known that WM and HAZs have experienced temperatures which are above an upper critical temperature, thus consistent with martensitic phase formation upon cooling (Ref 28). However, BM exhibits the initial microstructure (ferrite with carbides on grain boundaries), as shown in Fig. 2(e). The

simulated phase across the weldment demonstrates quite similar nature of phase transformation in the WM and HAZs, as shown in Fig. 2(f). There is a formation of an inter critical zone as well, see Fig. 2(d), where color changes from red to green to yellow (i.e., martensitic fraction changes), see Fig. 2(f). Moreover, the grain size was  $18 \pm 8$ ,  $5 \pm 2$ ,  $8 \pm 3$ , and  $11 \pm 3 \mu\text{m}$  respectively, for CGHAZ, FGHAZ, ICHAZ and BM regimes.

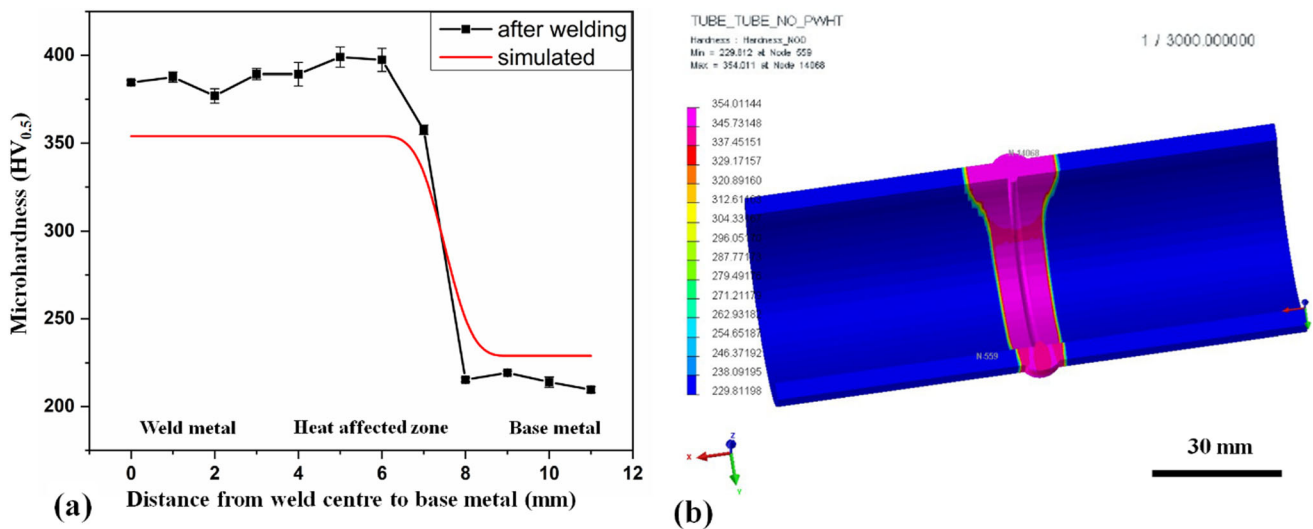
Figure 3 shows the comparison between experimental and simulated micro-hardness across the weldments. A similar variation in hardness for both can be seen. For instance, the experimentally measured hardness in WM and HAZs are 370 and  $\sim 400$  HV, respectively and then a sharp drop to  $\sim 213$  HV for BM) was observed. Similarly, the simulated curve exhibits a peak hardness of  $\sim 354$  HV in WM and HAZ and then a sharp drop to  $\sim 229$  HV in the BM region, as shown in Fig. 3(a, b). The peak hardness in WM and HAZs is mainly due to the martensitic transformation after welding. For more details on hardness variations in different HAZs, the reported investigation (Ref 22) can be referred.

Figure 4 depicts the measured residual stress across the weldments using two methods ( $\sin^2\psi$  and  $\cos \alpha$ ) and the corresponding validation through SYSWELD simulated resid-

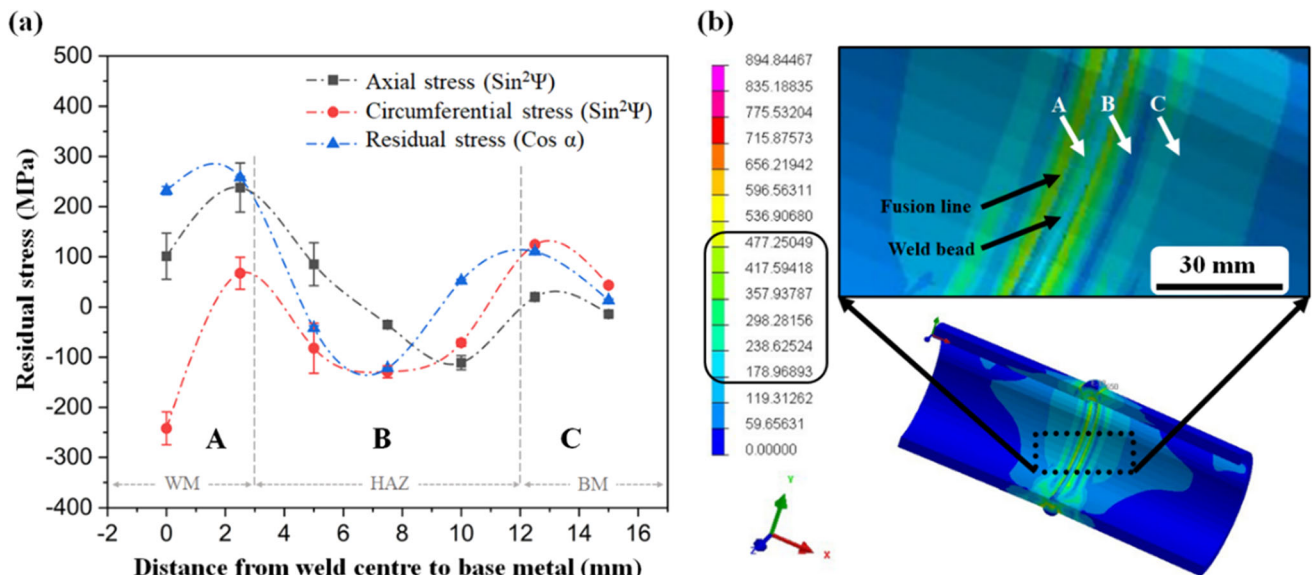


**Fig. 2** SEM micrograph of (a) weld metal (WM), (b) coarse grain heat affected zone (CGHAZ), (c) fine-grain heat affected zone (FGHAZ), (d) inter-critical heat affected zone (ICHAZ), (e) base metal (BM) and corresponding (f) simulated phase variation (magenta- ferritic, blue- martensitic), for as welded similar welds (T91-T91)





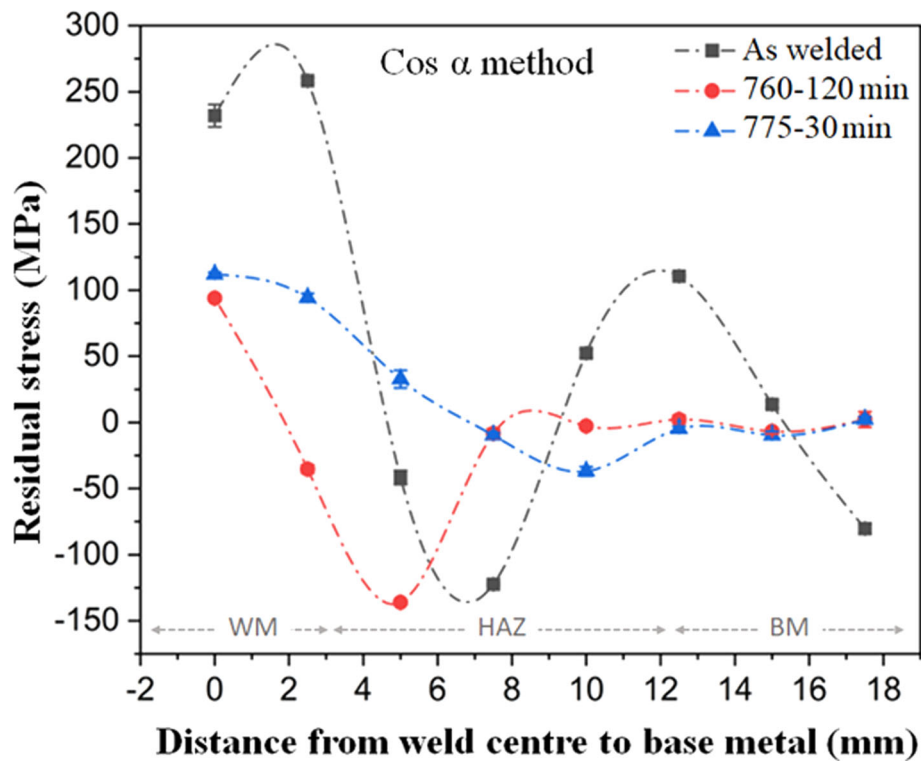
**Fig. 3** (a) Comparison of experimental and simulated microhardness variation and (b) simulated hardness, across the weldments after similar welds (T91-T91)



**Fig. 4** (a) Experimentally measured residual stresses using Sin<sup>2</sup>ψ and Cos α methods and (b) simulated residual stresses, for as welded condition of similar welds (T91-T91)

ual stresses for the as-welded condition. Accurately measuring residual stresses through experiments always poses a challenge, hence, two different experimental methods were used. Weldments can be divided into three regions WM (A), HAZ (B) and BM (C), depending on the nature of stresses, which shows a change in each of these regimes. Tensile axial residual stresses of  $101 \pm 37$  MPa were observed in WM using the sin<sup>2</sup>ψ-method, while circumferential residual stresses of  $242 \pm 33$  MPa were compressive, as shown in Fig. 4(a). In contrast, the cos α method measured a single magnitude of residual stresses in weldment and demonstrated the tensile residual stresses of  $231 \pm 8$  MPa. In WM (near the HAZ region), axial residual stress of  $238 \pm 41$  MPa and circumferential residual stress of  $\sim 69 \pm 32$  MPa were measured, both of which were tensile and demonstrated the peak value of residual stresses. At the same position, the cos α method

revealed the tensile peak residual stress of  $258 \pm 9$  MPa, as shown in Fig. 4(a). The simulated residual stresses suggested a  $\sim 190$  MPa in WM (as marked by a black arrow) and  $\sim 430$  MPa at the fusion line (as marked by a black arrow), as shown in Fig. 4(b). Kim et al. (Ref 34) also observed lower residual stresses in the weld centerline than on the fusion boundary in T91 steel using a numerical model and the neutron diffraction method. This is attributed to metallurgical processes during welding, such as shrinkage + quenching or shrinkage + phase transformation (Ref 35, 36). Therefore, 'M' shaped characteristics of residual stress distribution are expected in the case of ferrite-martensitic steel (Ref 37). They reported (Ref 34) a peak residual stress of  $\sim 700$  MPa, which is relatively higher than the yield stress ( $\sim 415$  MPa) and the ultimate tensile strength ( $\sim 585$  MPa) of T91 steel. Residual stresses in the HAZ changed from tensile to compressive, where axial and circum-



**Fig. 5** Influence of PWDT on residual stress distribution in similar welds (T91-T91) using the Cos  $\alpha$  method

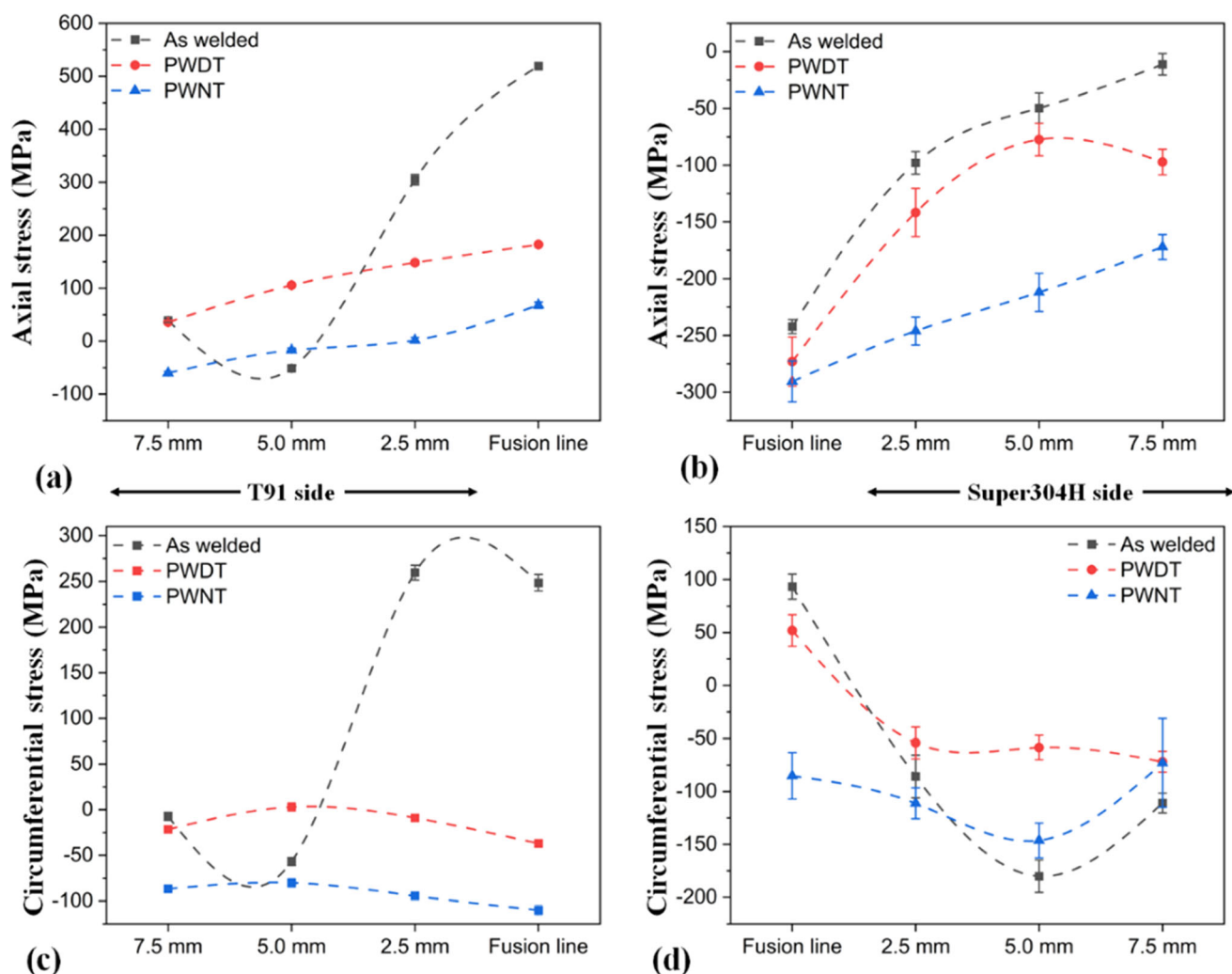
ferential stresses reached  $36 \pm 7$  and  $131 \pm 12$  MPa (in mid of HAZ), respectively, as shown in Fig. 4(a). The cos  $\alpha$  method revealed a similar nature and magnitude ( $122 \pm 3$  MPa) of residual stresses. Further, away from the fusion line, residual stresses decreased, as can be observed in experimental results, as shown in Fig. 4(a, b) by region B. The tensile residual stresses in the WM region and compressive residual stresses in HAZ are due to the solidification pattern. After welding, when solidification starts, T91 steel expands due to phase transformation from austenite to martensite. The martensitic crystal structure (body-centered tetragonal, BCT) increased its volume by 4.3% from the austenitic structure (face-centered cubic, FCC) due to its less dense structure. Due to this volume expansion, resulting from the predominant presence of martensitic structure in WM, a tensile residual stress is invariably induced in WM, which compresses the neighbouring HAZ, where compressive residual stresses get induced. Further, the experimental methods ( $\sin^2\psi$  and cos  $\alpha$ ) and as well the simulation results demonstrate that the residual stresses approach zero in the BM region (much away from the weld center), which signifies that BM is unaffected by the welding thermal cycle. Overall, both experimental and simulation methods have captured the nature of residual stresses well; however, some deviation in magnitude was present due to their variation in resolution ability. Nonetheless, the cos  $\alpha$  method exhibits good agreement with the  $\sin^2\psi$  method.

Both methods can be significant in evaluating residual stresses, with each having advantages over the other. For instance, the cos  $\alpha$  method is quick and more portable than the  $\sin^2\psi$  method. Hence, cos  $\alpha$  method is suitable for on-site applications as well. Further,  $\sin^2\psi$  method provides axial stress as well as circumferential stress, while cos  $\alpha$  provides in-plane normal and shear stress at particular exposed areas. On the

other hand, cos  $\alpha$  method is more viable for isotropic material and nearly flat surfaces, while  $\sin^2\psi$  method is useful for complex geometries such as narrow grooves/gaps (Ref 38). Both methods are capable of measuring residual stress for various materials (the ability to measure residual stresses for different materials depends on the target material, as reported elsewhere (Ref 39)). Moreover, the goniometer that provides different tilt angles in the  $\sin^2\psi$  method (more number of grains participate in measurement) provides a slightly different magnitude than the cos  $\alpha$  method for a particular region. The present investigation found that the repeatability (i.e., less error) of the cos  $\alpha$  method is better than  $\sin^2\psi$  for a specific area.

### 3.2 Residual Stresses in As-Welded Similar Welds: Influence of PWHT

The influence of PWHT on residual stress distribution across the weldment of similar T91 welds has been investigated using the cos  $\alpha$  method for as-welded and PWHT at 760 °C—120 min and 775 °C—30 min conditions, as shown in Fig. 5. It is well-known that the cos  $\alpha$  method is portable and quick, making it more useful for practical applications such as on-site inspections; hence it has been considered. The peak residual stress was  $258 \pm 6$  MPa near the fusion line after welding, with a significant deviation in its magnitude across weldment, as shown in Fig. 5. However, the residual stresses decreased after PWDT at 760 °C or 775 °C. It can be noticed that the variation in residual stress across weldment is relatively less for 775 °C—30 minutes due to higher temperature. This indicates that temperature has a more significant influence on residual stresses during PWHT. This is consistent with previous work (Ref 22), where better mechanical properties for PWDT at 775 °C—30 minutes have been observed. This has been attributed to uniform hardness and fine precipitate size.



**Fig. 6** (a, b) Axial and (c, d) circumferential residual stresses for the T91 side and Super304H side, respectively, for dissimilar welds of T91 and Super304H

Moreover, prolonged PWHT (say 2 h) coarsens the  $M_{23}C_6$  precipitates, which initiates decohesion from the precipitates-matrix interface (Ref 22).

### 3.3 Residual Stresses in As-Welded Dissimilar Welds with and without PWHT

DMW is a type of trimetallic joint in which T91 and Super304H were welded together with ERNiCr-3 as a filler. As a result, complications exist in measuring residual stresses using two different techniques for two different materials. Hence, using the  $\sin^2\psi$  approach, the residual stresses have been measured for all three conditions (as welded, PWDT and PWNT) from the T91 side (fusion line to BM) and the Super304H side (fusion line to BM), as shown in Fig. 6. Like similar welds, the T91 fusion line has a peak axial tensile residual stress of  $518 \pm 11$  MPa due to phase change and higher yield strength, see Fig. 6(a). However, the magnitude was much higher than similar welds due to heterogeneous metal joints with varying expansion and contraction thermal coefficients. After welding, the circumferential residual stress is less than the axial residual stress on the T91 side, as shown in Fig. 6(c), similar to the similar welds. The Super304H side, on

the other hand, shows compressive axial residual stresses due to the absence of phase change, while circumferential residual stresses are tensile with a lower magnitude of  $100 \pm 17$  MPa, see Fig. 6(b, d). After PWDT and PWNT heat treatments, there is a significant reduction in residual stress magnitude, as seen in similar welds. It's also worth noting that the deviation in residual stress distribution has dropped. The circumferential residual stress from the T91 side was less than axial residual stresses. Moreover, it can be noticed that the PWNT is more effective in reducing overall residual stress across weldment due to complete austenitizing and tempering, which is a unique heat treatment that is not reported yet for DMW (T91-Super304H). Hence the careful selection of PWHTs makes the dissimilar welded component out of danger of failure from the developed residual stresses.

## 4. Conclusions

An experimental and numerical investigation has been carried out to analyze the residual stresses in welded tubes of T91 steel. Additionally, the effectiveness of PWHTs (temper-



ature and time) on residual stresses has been studied for similar as well as dissimilar welds.

- Experimental ( $\cos \alpha$  and  $\sin^2 \psi$ ) and SYSWELDS simulation showed the peak tensile residual stresses in the WM region (close to the fusion line) are  $\sim 238$  MPa ( $\sin^2 \psi$ ), 258 MPa ( $\cos \alpha$ ) and  $\sim 430$  MPa (SYSWELDS), for T91 similar welds. SYSWELDS exhibited a similar nature of residual stresses across the weldment; however, the magnitudes were different.
- While in the case of dissimilar welds, peak tensile residual stresses of  $\sim 518$  MPa and peak compressive residual stresses of  $\sim 290$  MPa were observed at the fusion line of the T91 side and Super304H side, respectively. DMW exhibits relatively high residual stresses with large deviations across weldment due to varying thermal coefficients of expansion/contraction resulting from dissimilar metal joints.
- For similar welds of T91, the peak tensile residual stresses decreased from  $\sim 258$  to  $\sim 120$  MPa after PWDT ( $775^\circ\text{C}$ —30 min). Whereas, for dissimilar welds of T91 with Super304H, the peak tensile residual stresses were decreased from  $\sim 518$  to  $\sim 70$  MPa after PWNT, which is quite effective due to complete austenitizing.

## Acknowledgments

This project has been sponsored by NTPC NETRA. The authors also acknowledge the research infrastructure support of the department of materials science and engineering and central research facilities of IIT Delhi. The valuable discussion and support of Prof. S. Aravindan (Mechanical department, IIT Delhi) is duly acknowledged.

## References

1. R.L. Klueh, Elevated Temperature Ferritic and Martensitic Steels and Their Application to Future Nuclear Reactors, *Int. Mater. Rev.*, 2005, **50**, p 287–310. <https://doi.org/10.1179/174328005X41140>
2. A. Sauraw, A.K. Sharma, D. Fydrich, S. Sirohi, A. Gupta, A. Świerczyńska, C. Pandey, and G. Rogalski, Study on Microstructural Characterization, Mechanical Properties and Residual Stress of Gtaw Dissimilar Joints of p91 and p22 Steels, *Materials (Basel)*, 2021 <https://doi.org/10.3390/ma14216591>
3. J. Shen, P. Agrawal, T.A. Rodrigues, J.G. Lopes, N. Schell, Z. Zeng, R.S. Mishra, and J.P. Oliveira, Gas Tungsten Arc Welding of As-cast AlCoCrFeNi<sub>2.1</sub> Eutectic High Entropy Alloy, *Mater. Des.*, 2022, **223**, p 111176. <https://doi.org/10.1016/j.matdes.2022.111176>
4. J. Shen, P. Agrawal, T.A. Rodrigues, J.G. Lopes, N. Schell, J. He, Z. Zeng, R.S. Mishra, and J.P. Oliveira, Microstructure Evolution and Mechanical Properties in a Gas Tungsten Arc Welded Fe<sub>42</sub>Mn<sub>28</sub>-Co<sub>10</sub>Cr<sub>15</sub>Si<sub>5</sub> Metastable High Entropy Alloy, *Mater. Sci. Eng. A.*, 2023, **867**, p 144722. <https://doi.org/10.1016/j.msea.2023.144722>
5. J.G. Lopes, P. Rocha, D.A. Santana, J. Shen, E. Maawad, N. Schell, J.P. Oliveira, Impact of Arc-Based Welding on the Microstructure Evolution and Mechanical Properties in Newly Developed Cr<sub>29</sub>.7Co<sub>29</sub>.7Ni<sub>35</sub>.4Al<sub>4</sub>Ti<sub>1</sub>.2 Multi-principal Element alloy. *Adv. Eng. Mater.* (2023)
6. H.T. Kang, Y.L. Lee, and X.J. Sun, Effects of Residual Stress and Heat Treatment on Fatigue Strength of Weldments, *Mater. Sci. Eng. A.*, 2008, **497**, p 37–43. <https://doi.org/10.1016/j.msea.2008.06.011>
7. Y.C. Lin and S.C. Chen, Effect of Residual Stress on Thermal Fatigue in a Type 420 Martensitic Stainless Steel Weldment, *J. Mater. Process. Technol.*, 2003, **138**, p 22–27. [https://doi.org/10.1016/S0924-0136\(03\)00043-8](https://doi.org/10.1016/S0924-0136(03)00043-8)
8. D. Deng, Y. Zhou, T. Bi, and X. Liu, Experimental and Numerical Investigations of Welding Distortion Induced by CO<sub>2</sub> Gas Arc Welding in Thin-plate Bead-on Joints, *Mater. Des.*, 2013, **52**, p 720–729. <https://doi.org/10.1016/j.matdes.2013.06.013>
9. J.B. Ju, J.S. Lee, J.I. Jang, W.S. Kim, and D. Kwon, Determination of Welding Residual Stress Distribution in API X65 Pipeline Using a Modified Magnetic Barkhausen Noise Method, *Int. J. Press. Vessel. Pip.*, 2003, **80**, p 641–646. [https://doi.org/10.1016/S0308-0161\(03\)00131-5](https://doi.org/10.1016/S0308-0161(03)00131-5)
10. S. Paddea, J.A. Francis, A.M. Paradowska, P.J. Bouchard, and I.A. Shibli, Residual Stress Distribution in a P91 Steel-pipe Girth Weld Before and After Post Weld Heat Treatment, *Mater. Sci. Eng. A.*, 2012, **534**, p 663–672. <https://doi.org/10.1016/j.msea.2011.12.024>
11. V.I. Monin, T. Gurova, X. Castello, and S.F. Estefen, Analysis of Residual Stress State in Welded Steel Plates by X-ray Diffraction Method, *Rev. Adv. Mater. Sci.*, 2009, **20**, p 172–175
12. A. Yaghi, T.H. Hyde, A.A. Becker, W. Sun, and J.A. Williams, Residual Stress Simulation in Thin and Thick-walled Stainless Steel Pipe Welds Including Pipe Diameter Effects, *Int. J. Press. Vessel. Pip.*, 2006, **83**, p 864–874. <https://doi.org/10.1016/j.ijpvp.2006.08.014>
13. J.A. Francis, M. Turski, and P.J. Withers, Measured Residual Stress Distributions for Low and High Heat Input Single Weld Beads Deposited on to SA508 Steel, *Mater. Sci. Technol.*, 2009, **25**, p 325–334. <https://doi.org/10.1179/174328408X372074>
14. J. Shen, R. Gonçalves, Y.T. Choi, J.G. Lopes, J. Yang, N. Schell, H.S. Kim, and J.P. Oliveira, Microstructure and Mechanical Properties of Gas Metal Arc Welded CoCrFeMnNi Joints Using a 308 Stainless Steel Filler Metal, *Scr. Mater.*, 2023, **222**, p 115053. <https://doi.org/10.1016/j.scriptamat.2022.115053>
15. J. Shen, R. Gonçalves, Y.T. Choi, J.G. Lopes, J. Yang, N. Schell, H.S. Kim, and J.P. Oliveira, Microstructure and Mechanical Properties of Gas Metal Arc Welded CoCrFeMnNi Joints Using a 410 Stainless Steel Filler Metal, *Mater. Sci. Eng. A.*, 2022, **857**, p 144025. <https://doi.org/10.1016/j.msea.2022.144025>
16. D. Akbari and I. Sattari-Far, Effect of the Welding Heat Input on Residual Stresses in Butt-welds of Dissimilar Pipe Joints, *Int. J. Press. Vessel. Pip.*, 2009, **86**, p 769–776. <https://doi.org/10.1016/j.ijpvp.2009.07.005>
17. K. Tanaka, K. Suzuki, Y. Akiniwa, Evaluation of Residual Stresses by X-ray Diffraction, Yokendo (in Japanese) (2006)
18. Y. Maruyama, T. Miyazaki, and T. Sasaki, Development and Validation of an X-ray Stress Measurement Device Using an Imaging Plate Suitable for the  $\cos \alpha$  Method, *J. Soc. Mater. Sci. Japan.*, 2015, **64**, p 560–566. <https://doi.org/10.1179/1362171813Y.0000000132>
19. S. Kumar, R. Awasthi, C.S. Viswanadham, K. Bhanumurthy, and G.K. Dey, Thermo-metallurgical and Thermo-mechanical Computations for Laser Welded Joint in 9Cr-1Mo(V, Nb) Ferritic/Martensitic Steel, *Mater. Des.*, 2014, **59**, p 211–220. <https://doi.org/10.1016/j.matdes.2014.02.046>
20. R.P. Mahto, R. Kumar, and S.K. Pal, Characterizations of Weld Defects, Intermetallic Compounds and Mechanical Properties of Friction Stir Lap Welded Dissimilar Alloys, *Mater. Charact.*, 2020, **160**, p 110115. <https://doi.org/10.1016/j.matchar.2019.110115>
21. C. Pandey, M. Mohan Mahapatra, P. Kumar, J.G. Thakre, and N. Saini, Role of Evolving Microstructure on the Mechanical Behaviour of P92 Steel Welded Joint in As-welded and Post Weld Heat Treated State, *J. Mater. Process. Technol.*, 2019, **263**, p 241–255. <https://doi.org/10.1016/j.jmatprotec.2018.08.032>
22. R. Kumar, A. Varma, Y.R. Kumar, H. Vashishtha, J. Jain, and S. Neelakantan, Optimization of Post-weld Heat Treatment Condition of Arc-welded T91 Steel Tubes, *Int. J. Press. Vessel. Pip.*, 2020, **188**, p 104213. <https://doi.org/10.1016/j.ijpvp.2020.104213>
23. Z. Liang, Y. Gui, and Q. Zhao, Investigation of Microstructures and Mechanical Properties of T92 Martensitic Steel/Super304 Austenitic Steel Weld Joints Made with Three Welding Consumables, *Arch. Metall. Mater.*, 2018, **63**, p 1249–1256. <https://doi.org/10.24425/123798>
24. R. Kumar, A. Varma, Y.R. Kumar, S. Neelakantan, and J. Jain, Enhancement of Mechanical Properties Through Modified Post-weld Heat Treatment Processes of T91 and Super304H Dissimilar Welded Joint, *J. Manuf. Process.*, 2022, **78**, p 59–70. <https://doi.org/10.1016/j.jmapro.2022.04.008>
25. ESI Group, SYSWELD 2010 reference manual. Digital Version, **4** (2010)

26. C. Pandey, M.M. Mahapatra, P. Kumar, and N. Saini, Effect of Normalization and Tempering on Microstructure and Mechanical Properties of V-groove and Narrow-groove P91 Pipe Weldments, *Mater. Sci. Eng. A.*, 2017, **685**, p 39–49. <https://doi.org/10.1016/j.msea.2016.12.079>
27. M.Y. Kim, S.C. Kwak, I.S. Choi, Y.K. Lee, J.Y. Suh, E. Fleury, W.S. Jung, and T.H. Son, High-temperature Tensile and Creep Deformation of Cross-Weld Specimens of Weld Joint between T92 Martensitic and Super304H Austenitic Steels, *Mater. Charact.*, 2014, **97**, p 161–168. <https://doi.org/10.1016/j.matchar.2014.09.011>
28. R. Kumar, A. Varma, Y.R. Kumar, J. Jain, and S. Neelakantan, Microstructure Anomaly Upon High Temperature Exposure and Its Influence on the Mechanical Properties of a Modified 9Cr-1Mo Steel Weld, *Mater. Charact.*, 2022 <https://doi.org/10.1016/j.matchar.2022.111937>
29. R. Kumar, A. Gokhale, A. Varma, Y.R. Kumar, S. Neelakantan, and J. Jain, Role of Nb (C, N) and Cr Carbides on the Fracture Behaviour of Super304H Steel Using In-situ Tensile Studies, *Mater. Lett.*, 2023, **351**, p 135107. <https://doi.org/10.1016/j.matlet.2023.135107>
30. L. Lars-Erik, *Computational Welding Mechanics: Thermomechanical and Microstructural Simulation*, Woodhead Publishing Limited, Cambridge, 2007
31. J. Goldak, A. Chakravati, and M. Bibby, A New Finite Element Model for Welding Heat Sources, *Metall. Trans. B.*, 1984, **15**, p 299–305. <https://doi.org/10.1021/ac60352a844>
32. J. Moravec, Influence of Welding Parameters on Weld Pool's Geometry in Shielding Gas Welding. Polypress, Lib. (2011)
33. D. Kollár, B. Kövesdi, and J. Nézö, Numerical Simulation of Welding Process: Application in Buckling Analysis, *Period. Polytech. Civ. Eng.*, 2017, **61**, p 98–109. <https://doi.org/10.3311/PPci.9257>
34. S.H. Kim, J.B. Kim, and W.J. Lee, Numerical Prediction and Neutron Diffraction Measurement of the Residual Stresses for a Modified 9Cr-1Mo Steel Weld, *J. Mater. Process. Technol.*, 2009, **209**, p 3905–3913. <https://doi.org/10.1016/j.jmatprotec.2008.09.012>
35. H.H. Lai and W. Wu, Practical Examination of the Welding Residual Stress in View of Low-carbon Steel Welds, *J. Mater. Res. Technol.*, 2020, **9**, p 2717–2726. <https://doi.org/10.1016/j.jmrt.2020.01.004>
36. N.S. Rossini, M. Dassisti, K.Y. Benyounis, and A.G. Olabi, Methods of Measuring Residual Stresses in Components, *Mater. Des.*, 2012, **35**, p 572–588. <https://doi.org/10.1016/j.matdes.2011.08.022>
37. K.A. Venkata, S. Kumar, H.C. Dey, D.J. Smith, P.J. Bouchard, and C.E. Truman, Study on the Effect of Post Weld Heat Treatment Parameters on the Relaxation of Welding Residual Stresses in Electron Beam Welded P91 Steel Plates, *Procedia Eng.*, 2014, **86**, p 223–233. <https://doi.org/10.1016/j.proeng.2014.11.032>
38. M. Belassel, J. Pineault, N. Caratanasov, M. Brauss, Comparison of Residual Stress Measurement Techniques and Implementation Using x-ray diffraction., *Eur. Conf. Residual Stress.* **43** (2016)
39. M.E. Fitzpatrick, A.T. Fry, P. Holdway, F.A. Kandil, J. Shaackleton, L. Suominen, Determination of Residual Stresses by X-ray Diffraction - Issue 2, (2005) 31

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



# Federated learning inspired privacy sensitive emotion recognition based on multi-modal physiological sensors

Neha Gahlan<sup>1</sup> · Divyashikha Sethia<sup>1</sup>

Received: 5 June 2023 / Revised: 21 August 2023 / Accepted: 27 August 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

## Abstract

Traditional machine learning classifiers can automatically evaluate human behaviour and emotion recognition tasks. However, prior research work does not secure users' privacy and personal information because they need complete access to sensitive physiological data. The recently introduced Federated Learning (FL) paradigm can address this problem. FL allows the local model updates to be sent to a central server, combining them to create a global model. It does not allow the global model to access the raw data used to train it. Motivated by the core concept of FL, this paper proposes a novel FL-based Multi-modal Emotion Recognition System (F-MERS) framework combining EEG, GSR, ECG, and RESP physiological sensors data. It uses Multi-layer Perceptron (MLP) as a base model for classifying complex emotions in three dimensions: Valence, Arousal, and Dominance (VAD). The work validates the F-MERS framework with three emotion benchmark datasets, DEAP, AMIGOS, and DREAMER, achieving accuracies of 87.90%, 89.02%, and 79.02%, respectively. It is the first FL-enabled framework for recognizing complex emotions in three dimensions (VAD) with multi-modal physiological sensors. The proposed study assesses the F-MERS framework in two scenarios: (1). Subject dependent and (2). Subject independent, making the framework more generalized and robust. The experimental outcomes indicate that the F-MERS framework is scalable, efficient in communication, and offers privacy preservation over the baseline Non-FL MLP model.

**Keywords** Emotion recognition system · Federated learning · Physiological sensors · Multi-modal · Privacy · MLP

## 1 Introduction

Emotions influence a person's physical health and decision-making abilities [1]. For instance, people are more prone to suffering from poor mental health during emotionally stressful times. Also, when a person is not feeling well, their emotional state is unbalanced. Hence, determining the emotional states is vital to ensure improved emotional wellness. There are two categories of indicators for recognizing emotions, as described below:

- **Physical indicators:** One is human bodily indicators, such as facial expression [2, 3], speech [4], gesture [5], Eye-tracking [6, 7], posture, and others, which have the advantage of being easy to collect. Nevertheless, it is quite easy for people to alter their body signs, such as their voice or facial expression, to hide their genuine emotions while interacting with others. People might, for instance, smile during a formal social gathering even if they are experiencing bad emotions. Hence, there is no way to guarantee the correctness of these indicators.
- **Physiological indicators:** These indicators capture the electrical activities (physiological responses) of the human body using physiological sensors like - Electroencephalogram (EEG) [8–10], Electrocardiogram (ECG) [11], Electrodermal Activity (EDA) [12], Heart Rate (HR) [13], Blood Volume Pulse (BVP) [14] and Respiration Rate (RESP) [15]. These indicators can

---

✉ Neha Gahlan  
nehagahlan\_2k21phdcs02@dtu.ac.in

Divyashikha Sethia  
divyashikha@dtu.ac.in

<sup>1</sup> Department of Software Engineering, Delhi Technological University, Rohini, New Delhi 110042, India

map a person's complex emotions with supporting approaches and cautions, including self-reported feelings. Indeed, mapping emotions based on physiological signals is a complex and challenging task, but it is an essential and valuable area of research. Unfortunately, a single physiological sensor is unable to capture emotional changes accurately. Therefore, combining multiple physiological sensors and self-reported assessments can help recognize complex emotions. Hence, emotion recognition using multiple physiological sensors is significant in research and real applications [16].

Multi-modal emotion recognition based on physiological sensors has garnered much attention in previous research due to the limitations of unimodal information (use of a single sensor) for emotion recognition. However, using a single physiological signal is unreliable as artefacts and noises in the signals distort the signals' features, which are input for training the ML models to classify emotions resulting in unjustified and inaccurate recognition. Recent research on multi-modal emotion recognition has extensively used several sensors' constructive fusions. Information fusion for multi-modal emotion recognition has been the main focus of most earlier efforts. The two primary components are feature-level and decision-level fusions. Busso et al. [17] provided an easy feature-level fusion method for concatenating all feature vectors into one large vector to train a classifier. Physiological measures from wearable devices are fed into automated artificial intelligence systems to classify emotional states. Machine Learning (ML) and Deep Learning (DL) methods enable the automated evaluation of large amounts of data and the establishment of correlations between measurements made under various conditions. The traditional emotion recognition systems include algorithms such as Support Vector Machine (SVM) [18], Decision Trees (DT) [19], K-Nearest Neighbor (KNN) [20, 21], Recurrent Neural Network (RNN) [22], Convolutional Neural Network (CNN) [23, 24], Artificial Neural Network (ANN) and LSTM (Long Short-Term Memory) [25]. These algorithms were successful in attaining higher accuracies with physiological sensors. However, when used for emotion recognition based on physiological sensors, they have a considerable detriment in terms of user data privacy. These techniques employ the user's complete physiological data to train a classification model. In such environments where a user's physiological data is being collected and stored, more security measures are needed to protect that data from being accessed by data attackers. It could be a serious issue, as physiological data is extremely sensitive and could be used for nefarious purposes if it falls into the wrong hands. To address this issue of data privacy,

McMahan et al. [26] introduced in 2016 a new paradigm called Federated Learning (FL).

## 1.1 Research questions

This paper aims to address the following research questions. These are two forms: Main Question (MQ) and a Specific Question (SQ).

**MQ1:** Why is privacy essential for physiological data?

**SQ1:** How can complex emotions be mapped into different dimensions?

**SQ2:** How physiological signals contribute to emotion recognition?

**SQ3:** Why is multi-modality required in emotion recognition?

**SQ4:** How can machine learning be used for automated emotion recognition systems?

**SQ5:** How federated learning paradigm is preserving data privacy in emotion recognition?

**Motivation for data privacy to sensitive physiological sensors:** Physiological data is considered sensitive because it reveals a lot about a person's health status and can potentially provide insights into their emotional state, behaviour, and habits. One of the significant issues in keeping physiological data while identifying emotions is privacy concerns. When exposed to various situations, physiological sensors such as EEG, GSR, HR, and others record a human's brain responses, skin conductance, and heart rates. These sensors evaluate and help map to track a person's thoughts, moods, emotions, and, most importantly, health state. Conventional Emotion Recognition Systems (ERS) use sensitive, physiological data for training and classification procedures without any privacy standards, allowing data attackers easy access and resulting in private data leaks. Recent incidents of data leaking from emotion recognition systems have been frequent. The electroencephalogram (EEG)-based biometric authentication systems comprises the individual difference in EEG signals as the only corresponding identification and performs authentication [27]. The breach of such sensitive EEG data will destroy authentication mechanisms. Millions of consumers' biometric data records comprising fingerprint scans and face recognition records were leaked by some third-party attackers and made public 2019 [28]. Therefore, it is essential to ensure the use of physiological sensor data for emotion recognition in a way that protects people's privacy and autonomy. Technological advances make collecting and analyzing this kind of sensitive data easier.



## 1.2 Federated learning background

The core concept of FL is training ML and DL models on distinct data distributed across different devices or clients, preserving the privacy of local data [29–31]. FL is an ideal solution to prevent the loss of privacy of users' data as it enables sending model gradients (weights) of the local data to the server in place of the complete raw data [32]. The server combines them to create a better global model, uses them for training the ML model, and sends the updated gradients (weights) back to the users.

The proposed approach aims to develop a global model via the Client and Server architecture of Federated Learning. Clients, Servers, and Aggregation frameworks are the three essential components of the FL paradigm that generates a high-quality global model [33, 34]. The following are the components of FL in detail:

- **Clients:** These are the initial locations for storing local and entire raw data. They develop the local model independently and communicate the weights to the global model (at the server).
- **Servers:** A powerful computational single node is frequently used as the server. It manages client communication. It allows the global model to perform averaging on the local model weights received from the clients and then use them for training before sending the updated weights back to the clients.
- **Aggregation Framework (FedAvg [35]):** Google offers FedAvg, an aggregation technique for forming a global model by aggregating client-side updates (as shown in the Fig. 1). FedAvg meets the essential privacy protection and data security standards by aggregating weights received from local models. It

enables multiple devices to train a machine-learning model while storing the user's raw data locally. This method relies on an optimization technique - Stochastic Gradient Descent (SGD) and produces outcomes without sacrificing user privacy or data sharing.

## 1.3 Main contribution of the paper

The following are the paper's main contributions:

- Proposal of Federated Learning-based Multi-modal Emotion Recognition System (F-MERS) framework, combining the physiological sensors: EEG, GSR, ECG, and RESP from the three benchmark datasets: AMI-GOS [36], DEAP [37] and DREAMER [38].
- The proposed F-MERS framework utilizes an MLP classifier as a base model for classifying emotional states. It uses the three-dimensional model of emotions (VAD), i.e., Valence, Arousal, and Dominance. The experimental results prove the Federated-inspired F-MERS framework to be as accurate as the base Non-FL MLP model, indicated by the comparable accuracies achieved by both the base MLP model and F-MERS framework.
- The proposed study validates the F-MERS framework in two scenarios: Subject-dependent and Subject independent, making the framework more generalized and robust.
- The proposed study calculates the training time and aggregation time of the proposed framework F-MERS, which proves it to be scalable and efficient in communication. The experimental results indicate that the proposed framework successfully recognizes emotions

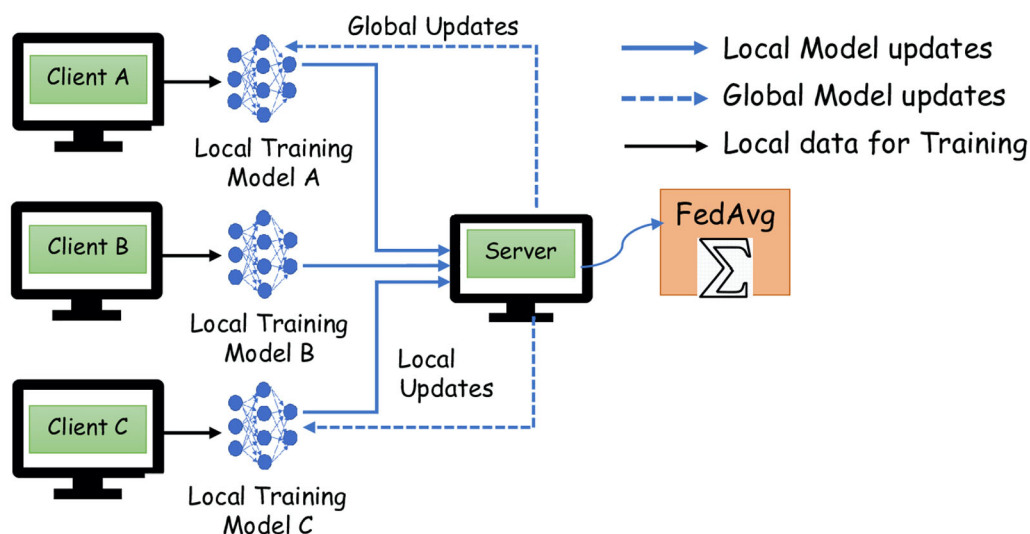


Fig. 1 Client-server architecture

without compromising users' privacy with sensitive information measuring.

## 1.4 Organization of the paper

The rest of the paper comprises Sect. 2, which illustrates the related work using physiological signals for emotion recognition with FL. Section 3 explains the proposed methodology. Section 4 gives the experimental outcomes for various model configurations. Section 5 discusses the results outcomes, and Sect. 6 explains the lessons learnt. The paper concludes with Sect. 7.

## 2 Related work

Only a few research studies have worked on physiological sensors for emotion recognition systems in an FL environment, as indicated in Table 1. The current study proposes a federated learning-enabled emotion recognition framework based on multi-modal physiological sensors for the first time. Previously, single physiological sensors such as EEG, Electrodermal Activity (EDA), Respiration Belt (RB), and Galvanic Skin Response (GSR) were employed in the federated learning environment to classify the emotional states ref.

Nandi and Xhafa [39] developed Fed-ReMECS, a real-time emotion classification framework using FL that leverages physiological data from EDA and RESP wearable sensors. The authors classify emotional states from the DEAP [37] dataset into two-dimensional emotional states: arousal and valence. The results conclude an average accuracy of 81.92% along with the objective of privacy preservation of the participating subjects. The proposed framework is limited to utilizing only peripheral physiological sensor data from EDA and RESP, which are less effective than EEG for recognizing human emotions. The research study failed to map complex emotions due to the absence of a dominance dimension.

Tara [40] employed the physiological sensor data GSR from the CASE [41] dataset to train a federated CNN-based model for emotion recognition. The author found that the proposed federated CNN architecture achieved the same accuracy as the non-federated centralized CNN, i.e. 79%. However, the research study used only single physiological sensor data from GSR. It recognized the emotions in arousal and valence dimensions. However, it did not map the complex emotions in the dominance dimension.

Gao et al. [42] proposed a heterogeneous federated learning approach for training the fully connected neural network over EEG data while maintaining each user's data privacy. The authors conduct an extensive experiment

using the dataset MindBigData [42], which collects data from wearable devices. The research study uses a very less known dataset, which needs a suitable emotion-elicitation environment and stimuli. The study is limited to only EEG physiological sensor data. The research did not fulfil the objective of emotion state classification as no emotion mapping is present in any dimension.

Ayaan et al. [27] proposed a FedEmo framework for emotion recognition using EEG signals based on the FL environment. This framework uses ANN as a classifier and FedAvg [35] to preserve the privacy of EEG data. The proposed framework is validated on the DREAMER [38] dataset for classifying three emotional states (valence, arousal, and dominance) and achieves an average accuracy of 57.4%. The research study is limited to only EEG physiological sensor data.

### 2.1 Research gaps

The related work on FL for emotion recognition has restrictions and limitations in their proposed study, such as: 1). None of the prior work on FL has focused on complex emotions, including the dominance dimension, to gain essential insight into the depth and intensity of emotions beyond simple positive and negative emotion classification. 2). The related work experiments on either one physiological sensor or fusing only peripheral sensors, not including EEG while fusing. The absence of EEG data in the fusion limits the depth and accuracy of the results obtained from those experiments. 3). None of the prior work on FL is validated for both scenarios, where the emotions are dependent on the individual subjects and where they are independent of the individual subjects. Only one is validated for subject-independent scenarios (Nandi et.al. [39]). 4). none of the related works have discussed the communication efficiency of their FL framework for emotion recognition. However, in the current study, we address both of these aspects.

The proposed framework can fill these research gaps and limitations in the literature described in the related work. The proposed study validates the framework with multi-modalities, fusing EEG, ECG, GSR, and RESP, for recognizing complex emotions in three dimensions (VAD), which is not performed in existing FL works. The proposed framework is more generalized and robust compared to the existing FL works for emotions as it validates the framework for three benchmark datasets of emotions and measures the communication efficiency of the proposed framework by analyzing the training time, aggregation time, and scalability other than evaluation metrics.



**Table 1** Recent physiological signal-based federated learning research work for emotion recognition

Refs.	Dataset	PS <sup>1</sup>	CM <sup>2</sup>	Tool	Algo	Avg. Accuracy (%)	Modality
Nandi et al. [39]	DEAP [37]	EDA, RB	FFNN	TFF	FedAvg	81.92	Bi-Modal
Tara Hassani [40]	CASE [41]	GSR	CNN	TFF	FedAvg	79	Single
Gao et al. [42]	MindBigData [42]	EEG	CNN-FC	PySfyt	FedAvg	86	Single
Ayaan et al. [27]	DREAMER [38]	EEG	ANN	TFF	FedAvg	57.4	Single

<sup>1</sup> Physiological Signals (PS)<sup>2</sup> Classification Model (CM)

### 3 Methodology

The methodology for the proposed framework is illustrated in this section as follows.

#### 3.1 Emotional model

Russell [45] established two categories as the basis for defining emotions: Arousal and Valence (as shown in Fig. 2a). The model lacks the mapping and depiction of more subtle emotions. Afterwards, Mehrabian and Russell [46, 47] together extended the two-dimensional model into a three-dimensional model known as the VAD (Valence-Arousal-Dominance) model (as shown in Fig. 2b). This approach introduces a third dimension of dominance, ranging from submissive to dominant emotions. Fear and wrath, for example, are easily distinguished via the dominant axis, which maps fear on the submissive axis and wrath on the dominant axis. As a result, this paper employs Mehrabian and Russell's VAD model for the proposed framework.

#### 3.2 Datasets description

The proposed framework validates the three emotion benchmark datasets: AMIGOS<sup>1</sup> [36], DEAP<sup>2</sup> [37], and DREAMER<sup>3</sup> [38]. Brief descriptions of all the datasets are given in Table 2, giving an overview of the no. of subjects in all three datasets, the different physiological signals from which the data is collected, the video stimuli and their durations shown to the subjects and their label matrices. For AMIGOS and DREAMER, the Emotiv EPOC Neuro-headset with 14 channels for collecting EEG data, and SHIMMER is used to collect the ECG and GSR data. For DEAP, the BioSemi ActiveTwo system with 32 electrodes for collecting EEG data and other peripheral physiological signals. To evaluate the proposed F-MERS framework, the experimental study (in Sect. 3.4) uses the multi-modal

signals as EEG, ECG and GSR from AMIGOS dataset; EEG, GSR and RESP from the DEAP dataset; EEG and ECG from DREAMER dataset.

These datasets have certain limitations concerning the subjects' demographic information (personal data), such as gender and age. The number of male and female subjects is disclosed, but their identities are intentionally anonymized to ensure and respect the privacy concerns of the subjects. For more details about the dataset, readers can find the dataset links provided as footnotes and their base paper references.

##### 3.2.1 Data pre-processing

The proposed study uses the preprocessed data given by the data owners as it has achieved state-of-the-art results as referenced [38, 39, 48]. The steps are given in below:

1. In the AMIGOS dataset, NaN values are detected in the ECG signals for the 28th subject for the 9th video, which is removed from the experimental setup.
2. The EEG, GSR, RESP and ECG data are downsampled to 128Hz. (For all three datasets)
3. EEG data is filtered from 4.0 to 45.0 Hz with a bandpass filter. (For all three datasets)
4. The ECG and GSR data are filtered with a cutoff frequency of 60Hz using a low-pass filter. (For AMIGOS)

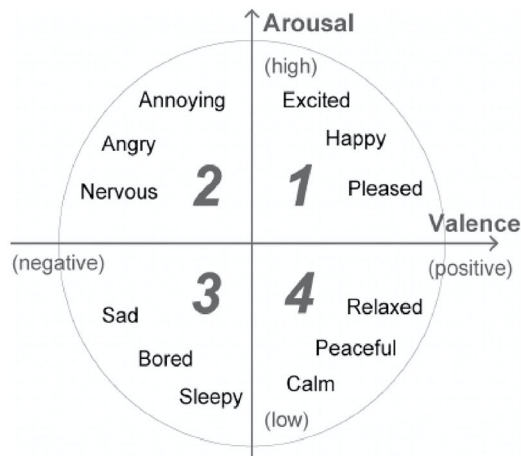
##### 3.2.2 Data clipping

- **AMIGOS:** This study focuses on the experiment with the 16 short videos since longer videos are less likely to elicit stable emotions [49]. Each video displayed to the subject in the experiment is of varying duration (as shown in Table 3), from which the study uses the last 50 s of data of every video and clips the starting portion.

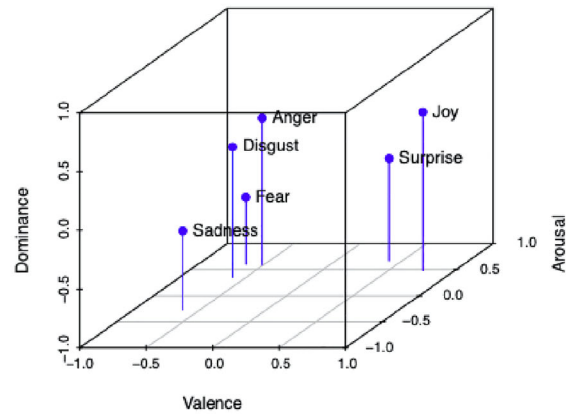
Data Matrix = [total subjects  $\times$  videos (per subject)  $\times$  samples (per video)] = [40  $\times$  16  $\times$  6400 (50  $\times$  128)]

- **DEAP:** The proposed study uses the data from the last 60 s of every video stimuli shown to the subjects after removing the 3 s baseline.

<sup>1</sup> <http://www.eecs.qmul.ac.uk/mmv/datasets/amigos/index.html>.<sup>2</sup> <https://www.eecs.qmul.ac.uk/mmv/datasets/deap/download.html>.<sup>3</sup> <https://zenodo.org/record/546113#%23.ZEn3hi8RpQI>.



(a) Russel's 2D model of Emotions



(b) 3D model of Emotions

**Fig. 2** Emotional models [43, 44]**Table 2** Brief description of all the datasets

Description/dataset	AMIGOS	DEAP	DREAMER
No. of subjects	40	32 (16 male, 16 female)	23 (14 male, 9 female)
Physiological signals	EEG, ECG, GSR	EEG, GSR, RESP	EEG, ECG
Video content	16 short videos, 4 long videos	40 videos	18 videos
Video duration	57–155 s, 14–15 min	63 s	65–393 s
Label matrix	16 × 3	40 × 3	18 × 3
Emotion states	Arousal, Valence, Dominance		
Emotion assessment	SAM (Self Assessment Manikins)		
EEG electrodes	14	32	14
ECG electrodes	2	–	2
GSR electrodes	1	1	–
RESP electrodes	–	1	–

Data Matrix =[total subjects × videos (per subject) × samples (per video)]=[32 × 40 × 7680 (60 × 128)]

- **DREAMER:** The proposed study uses the data from the last 60 s of every video stimuli shown to the subjects.

Data Matrix =[total subjects × videos (per subject) × samples (per video)]=[23 × 18 × 7680 (60 × 128)]

### 3.2.3 Data Labelling

Subjects from all three datasets use SAM [50] to rate their valence, arousal, and dominance emotional states. On a scale of 1–9 (in AMIGOS, DEAP) and 1–5 (in DREAMER). The emotional states of low and high valence, arousal, and dominance are defined by a threshold float value of 4.5 (in AMIGOS, DEAP) and 3 (in DREAMER) as shown in Table 4, and illustration of mapping of

emotions on the three-dimensional VAD emotional model as shown in Table 5.

### 3.3 Feature extraction & feature fusion

The proposed framework uses the following extracted features (as given in Table 6) from the physiological data: EEG, ECG (Right and Left channels), GSR and RESP. The framework applies a sliding window of 4 s with 50% overlap for the feature extraction methods for all three sensors.

**Feature fusion:** There are two methods for fusing physiological data from various sensors: Decision-level and Feature-level. Recent studies [38, 51–53] evaluated coherence between multi-modal signals to allow feature-level fusion and discovered that it enhanced overall accuracy over decision-level fusion. Hence, this work employs Feature-Level Fusion (FLF) as shown in Fig. 3 to use rich

**Table 3** Details of the 16 short videos shown to each subject in AMIGOS [36]

Video ID	No. of samples	Duration (in sec)	Videos
10	12225	96	August Rush
13	7229	57	Love Actually
138	15160	122	The Thin Red Line
18	10575	83	House of Flying Daggers
19	16106	126	Exorcist
20	8335	65	My Girl
23	14265	112	My Bodyguard
30	9717	76	Silent Hill
31	19886	155	Prestige
34	8417	66	Pink Flamingos
36	8698	68	Black Swan
4	11621	91	Airplane
5	14347	112	When Harry Met Sally
58	8181	64	Mr Beans Holiday
80	13047	102	Love Actually
9	9630	75	Hot Shots

**Table 4** Emotion state ratings

Ratings	AMIGOS	DEAP	DREAMER
Low (Arousal/Valence/Dominance)	1–4.5	1–4.5	1–3
High (Arousal/Valence/Dominance)	4.5–9	4.5–9	3–5

**Table 5** Mapping of emotions on 3D VAD

Emotions	Arousal	Valence	Dominance
Disgust	High	Low	High
Happiness	Low	High	Low
Surprise	High	High	Low
Anger	High	Low	High
Fear	High	Low	Low
Calm	Low	High	Low
Sorrow	Low	Low	Low
Excitement	High	High	High

data from distinct modalities. The FLF process involves performing feature extraction independently for each sensor data. After feature extraction, the resulting feature vectors from each sensor concatenate into a single fused feature vector. This fused feature vector represents the information extracted from all the sensor modalities (EEG, ECG, GSR, and RESP) used in the learning process. The final feature vectors obtained are:

For AMIGOS:

$$A_{EEG+ECG+GSR} = [f_{EEG}(14 * 17 = 238) + f_{ECG}(32 * 2 = 64) + f_{GSR}(20)] \quad (1)$$

For DEAP:

$$D_{EEG+GSR+R} = [f_{EEG}(32 * 17 = 544) + f_{GSR}(20) + f_R(6)] \quad (2)$$

For DREAMER:

$$Dr_{EEG+ECG} = [f_{EEG}(14 * 17 = 238) + f_{ECG}(32 * 2 = 64)] \quad (3)$$

### 3.4 Architecture of F-MERS framework

The current study proposes an FL framework using multi-modal physiological data for emotion recognition (F-MERS). The architecture for the proposed F-MERS is given below stepwise in detail. Figure 4 illustrates the architecture with marked respective steps.

- **Step 1: Data collection and preprocessing**

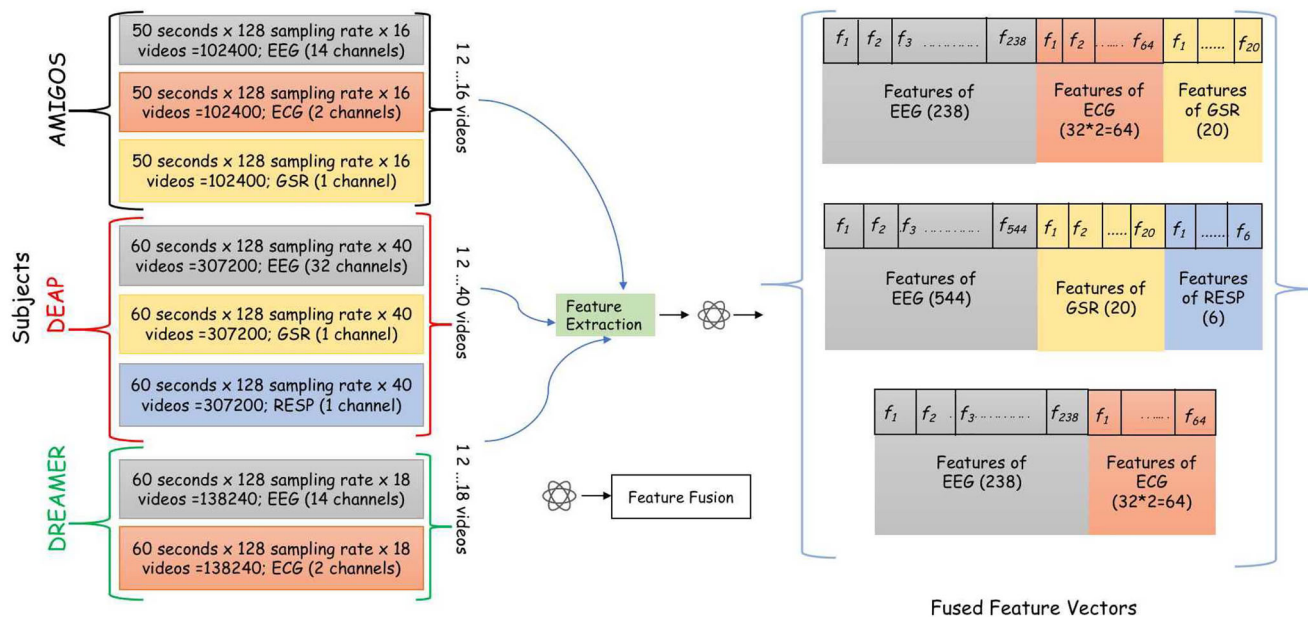
Firstly, the proposed framework collects multi-modal physiological data from all the subjects. It includes physiological signals EEG, ECG, GSR and RESP data. This data then goes for preprocessing, serving as input to the next step. section 3.2 explains the details of the dataset.

- **Step 2: Feature extraction and fusion**

The framework performs feature extraction on the collected physiological data for the Statistical, Time-

**Table 6** Features extracted from the physiological signals

PS <sup>1</sup>	Domain	Features with description
EEG (17)	Time, statistical, frequency	<b>Hjorth Features: Complexity, Mobility and Activity, Fractal Dimension: Higuchi, Petrosian, Spectral Entropy:</b> (measure of the distribution or randomness of the power spectrum of a signal), <b>Sample Entropy:</b> (measure of the irregularity or complexity of a time series), <b>SVD Entropy:</b> (measure of complexity of a signal using singular value decomposition), <b>Bandpower (alpha, beta, theta, delta):</b> (power or energy contained within specific frequency bands of a signal), <b>Mean, Median, Standard Deviation (STD), 1st and 2nd difference:</b> (average value, median, standard deviation, change between consecutive values a signal)
ECG (32)	Time, statistical, frequency	<b>mean_nni, median_nni:</b> (average and middle of the NN intervals between successive heartbeats), <b>nni_50, pnni_50, nni_20, pnni_20:</b> (number and percentage of NN intervals differing by more than 50/20 milliseconds), <b>hrv_mean, hrv_sdmn:</b> (mean value, standard deviation of the differences between adjacent NN intervals), <b>hrv_rmssd:</b> (root mean square of successive differences between NN intervals), <b>range_nni:</b> (difference between the maximum and minimum NN intervals), <b>cvsd:</b> (coefficient of variation of successive differences between NN intervals), <b>cvnni:</b> (coefficient of variation of NN intervals), <b>mean_hr, max_hr, min_hr, std_hr:</b> (average heart rate, maximum, minimum, standard deviation calculated from NN intervals) <b>triangular_index:</b> (measure of the overall shape of the RR interval histogram), <b>tinn:</b> triangular interpolation of NN intervals, <b>total_power</b> (total power of the frequency spectrum of the ECG signal), <b>vlf</b> (power in the very low-frequency range of the ECG signal), <b>lf</b> (power in the low-frequency range of the ECG signal), <b>hf</b> (power in the high-frequency range of the ECG signal), <b>lf_hf_ratio</b> (ratio of LF power to HF power), <b>lfnu</b> (normalized low-frequency power), <b>hfnu</b> (normalized high-frequency power), <b>csi</b> (complexity index of the heart rate variability), <b>cvi</b> (cardiopulmonary coupling index), <b>sd1</b> (standard deviation of points perpendicular to the line of identity), <b>sd2</b> (standard deviation along the line of identity), <b>ratio_sd2_sd1</b> (ratio of sd2 to sd1), <b>sampen</b> (sample entropy, a measure of the complexity of the signal)
GSR (20)	Statistical, time	<b>mean_gsr, var_gsr, skew_gsr, kurtosis_gsr, std_gsr:</b> (average value, variance, skewness, kurtosis, standard deviation of the GSR signal), <b>SCL (tonic) slope:</b> slope of the tonic (slow-changing) component of the skin conductance level (SCL)), <b>SCR (phasic) peaks:</b> (The number of peaks representing the phasic (rapid-changing) component of the skin conductance response (SCR)), <b>Statistical features applied to SCR, SCL:</b> mean_scl, var_scl, skew_scl, kurtosis_scl, std_scl, slope_scl, mean_scr, var_scr, skew_scr, kurtosis_scr, std_scr, max_scr, scr_peaks
R <sup>2</sup> (6)	Statistical	<b>mean_rb, median_rb, var_rb, skew_rb, kurtosis_rb, std_rb:</b> (average value, median value, variance, skewness, kurtosis, standard deviation of the Respiration signal)

<sup>1</sup> Physiological Signal(PS)<sup>2</sup> Respiration (R)**Fig. 3** Feature Fusion

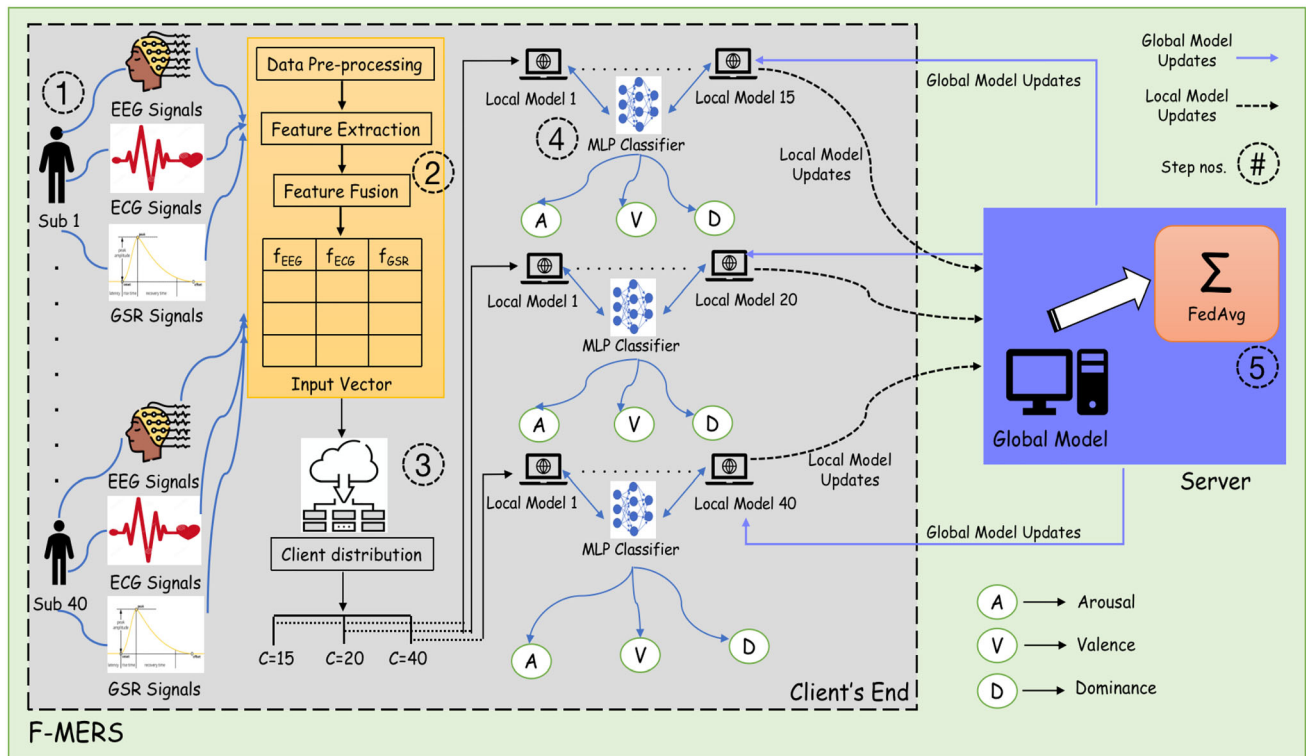


Fig. 4 Architecture of Proposed F-MERS Framework (Illustration with AMIGOS dataset)

domain, and Frequency-domain features. It fuses the extracted features into a single concatenated feature vector. The section 3.3 gives a detailed explanation for it.

### • Step 3: Data division for FL environment

In this step, we define the participating clients in the FL framework. For creating a federated framework, the framework divides the datasets into three different client divisions, with each client consisting of one subject's multi-stream physiological data, followed similarly by two other datasets. An illustration is shown in Fig. 5 for the AMIGOS dataset showing that the first experiment is with clients = 15, the second with clients = 20, and the third is clients = 40. It works similarly for the other two datasets with different no. of client as given in Table 7.

### • Step 4: Model selection and training

The proposed study uses Multi-Layer Perceptron (MLP) neural network in a Federated learning environment as a baseline classifier. It consists of the following parameters:

- The MLP model takes the concatenated feature vector of the physiological sensors data in the input layer.
- At hidden layers, the model has three Dense Layers.
- The Rectified Linear Activation Unit (ReLU) is the activation function for each Dense layer as it has

faster convergence and is computationally inexpensive.

- Dropout layers after each hidden layer regularise the neural network at a rate of 0.7. It is only applied during the training phase and turned off during evaluation, as the goal is to use the entire network's representation power during evaluation.
- The output layer gives a single output and employs the Sigmoid activation function.
- The federated and non-federated frameworks employ the SGD optimizer with a 0.05 learning rate.

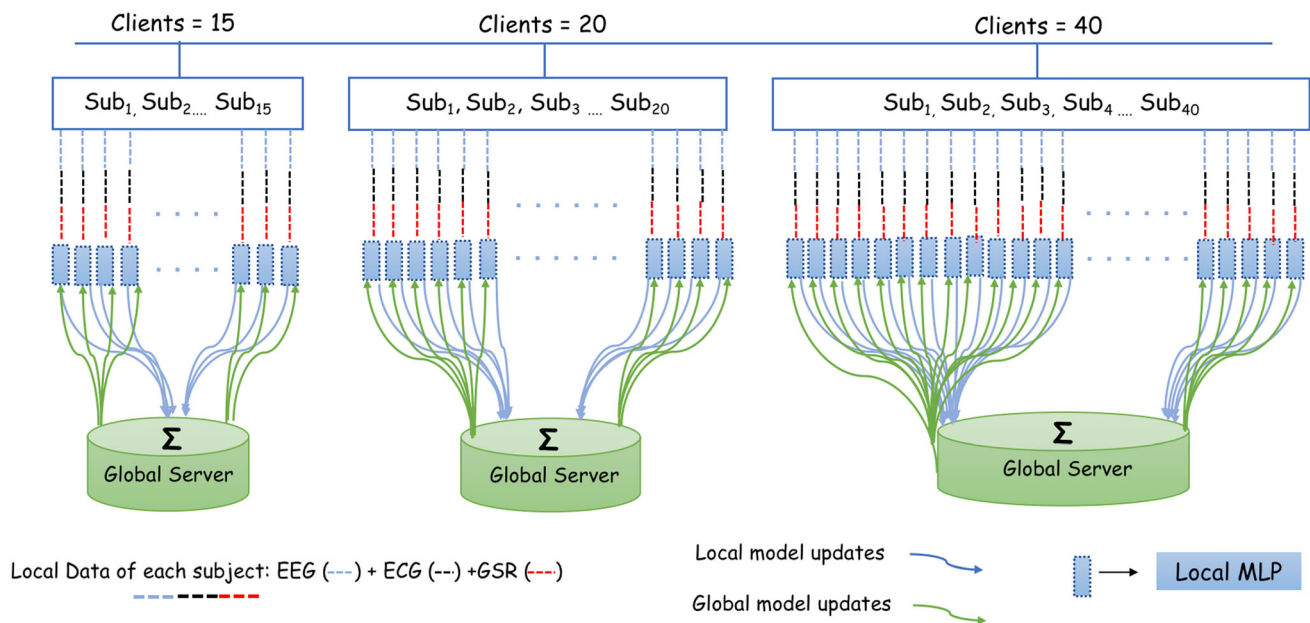
### • Step 5: Creating FL environment

TensorFlow Federated (TFF) [54] is used as an FL tool in the proposed framework. After splitting the dataset, the framework distributes it to multiple virtual clients. It creates a federated learning environment using TFF to produce FedAvg, a federated averaging algorithm [35]. TFF employs a distributed aggregation protocol [55] to collect and aggregate model updates from the clients. Equation 4 gives the computation for FedAvg.

$$w_t^g = \frac{1}{N_{total}} \sum_{i=1}^{N_{total}} w_{t,i}^l \quad (4)$$

Here  $w_t^g$  is the aggregated weight at the global server in time  $t$ ,  $w_{t,i}^l$  are the weights received from all local





**Fig. 5** Data divisions of clients (Illustration with AMIGOS dataset)

**Table 7** Data partitioning for subjects into clients for each dataset. (1 Client= 1 subject)

AMIGOS (40)		
Client = 15	Client = 20	Client = 40
DEAP (32)		
Client=10	Client=16	Client = 32
DREAMER (23)		
Client=7	Client=11	Client = 23

models in time  $t$ , and  $N_{total}$  is the total number of the local model participating for aggregation. It employs the horizontal federated learning approach [27]. The Algorithm 1 describes the detailed steps for the algorithm for FedAvg.

- **Creation of local model (client end):** The local model is created using the multi-layer perceptron, initially taking the feature vector as input from each client's multi-stream physiological data. The models from each client are then sent to the server to generate a global model.
- **Creation of global model (server end):** After receiving the local model from the clients, the server performs the federated averaging (FedAvg). It then sends the global model back to the clients.
- **Local model updates (client end):** Following aggregation, the participating client receives an updated global model from the server. The process repeats until the model converges or completes the number of iterations. The experiments are performed with three rounds of iterations (Rounds = 100, 200, 500).



**Algorithm 1 Federated Averaging (FedAvg [35])****Server Execution:**

```

initialize  $w_0$ 
for each round  $r = 1, 2, \dots$  do
   $m \leftarrow \max(\text{int}(C * m), 1)$ ;
   $S_t \leftarrow$  (random set of  $m$  clients);
  for each client  $k \in S_t$  in parallel do
     $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$  end
   $w_{t+1} \leftarrow \sum_{k=1}^K \frac{n_k}{n} w_{t+1}^k$ 

```

**ClientUpdate(k,w):**//Run on client  $k$ 

```

 $A \leftarrow$  (split  $P_k$  into batches of size  $B$ )
for each local epoch  $i$  from 1 to  $E$  do
  for batch  $a \in A$  do
     $w \leftarrow w - \eta \Delta l(w; a)$ 

```

```

return  $w$  to server

```

**3.5 Validation scenarios of the framework**

The study validates the proposed FL approach in two scenarios: One is Subject-dependent, and the other is Subject-independent.

- **The subject-dependent scenario** uses data from each client for training and testing, i.e., 80% of each of the client's data for training and 20% of each of the client's data for testing.
- **The subject-independent scenario** uses different sets of training and testing clients, which are in the ratio of 80% (training) and 20% (testing). For example, the model uses 12 clients for training and the rest 3 for testing.

Table 8 presents the no. of subjects utilized for training and testing the experiment in the subject-independent scenario separately for each dataset, along with the details of the three experiments performed.

**3.6 Evaluation metrics**

The following metrics evaluate the proposed framework:

- **Loss function:** Binary Crossentropy (as indicated in Eq. 5). Here  $y$  is the binary indicator (0 or 1) for the class label's correct classification, and  $p$  is the predicted probability for the observations for class labels.

$$-(y \log(p) + (1 - y) \log(1 - p)) \quad (5)$$

**Table 8** Data partitioning of subjects into clients for the proposed F-MERS in Subject-independent scenario for all three datasets

Dataset = AMIGOS (40)			
1 Client = 1 subject	Client = 15	Client = 20	Client = 40
Training/Testing clients	12/3	16/4	32/8
Dataset = DEAP (32)			
1 Client = 1 subject	Client = 10	Client = 16	Client = 32
Training/Testing clients	8/2	13/3	26/8
Dataset = DREAMER (23)			
1 Client = 1 subject	Client = 7	Client = 11	Client = 23
Training/testing clients	5/2	9/2	18/5

**Table 9** Evaluation Metrics

Metric	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
F1-Score	$\frac{TP}{TP+\frac{1}{2}(FP+FN)}$

- **Metrics:** Confusion Matrix, Binary Accuracy and F1-Score (as indicated in Table 9). Here, TP is the True Positives, TN is the True Negative, FP is the False Positives, and FN is the False Negative. The sci-kit

learn library of Python calculates the Accuracy and F1-Scores.

### 3.6.1 Scalability metrics

To measure the scalability of the proposed FL-based F-MERS framework, this work uses the following metrics:

- **Training Time:** The time it takes to train the model on a dataset helps indicate its scalability. A scalable federated learning model should be able to handle bigger datasets and models with appropriate training times, even when data is distributed across different devices.
- **Model Accuracy:** The model should remain accurate as the number of clients and data increases.

### 3.6.2 Communication efficiency metrics

Federated Learning is a distributed machine learning approach that uses data stored on multiple local models for aggregation at a single global server. Its communication efficiency can be measured using aggregation time.

- **Aggregation Time:** The time required to aggregate the model updates received from the local client models impacts the communication efficiency of the federated learning model. If the model aggregation time is long, it can result in increased communication latency and increased network congestion.

## 3.7 Experimental setup

The proposed study experiments with two perspectives: The non-federated learning environment (Non-FL) and the Federated learning environment (FL). The study validates both of these approaches in Subject-dependent and Subject-independent scenarios. These approaches are employed to assess the effectiveness of the proposed framework. For both perspectives, this work utilizes Google Colab's Pro plus GPU and Python 3.8 to run the tests and trials on a MacBook Air with a 1.6 GHz dual-core Intel core i5. Each round of aggregation in the federated environment requires clients to train one epoch locally, with batch size 256 [24, 56].

## 4 Experimental results

This section presents the experimental results of the proposed F-MERS framework for recognizing emotional states. As the global server has no access to the local data,

which is carefully confined to each client, the emotion classification is done by each client's local model. Hence, the global model's accuracy is derived by averaging the accuracy value of each local model after receiving the global model updates from the global server. The following subsections present the experimental results for both Subject-dependent and Subject-independent scenarios.

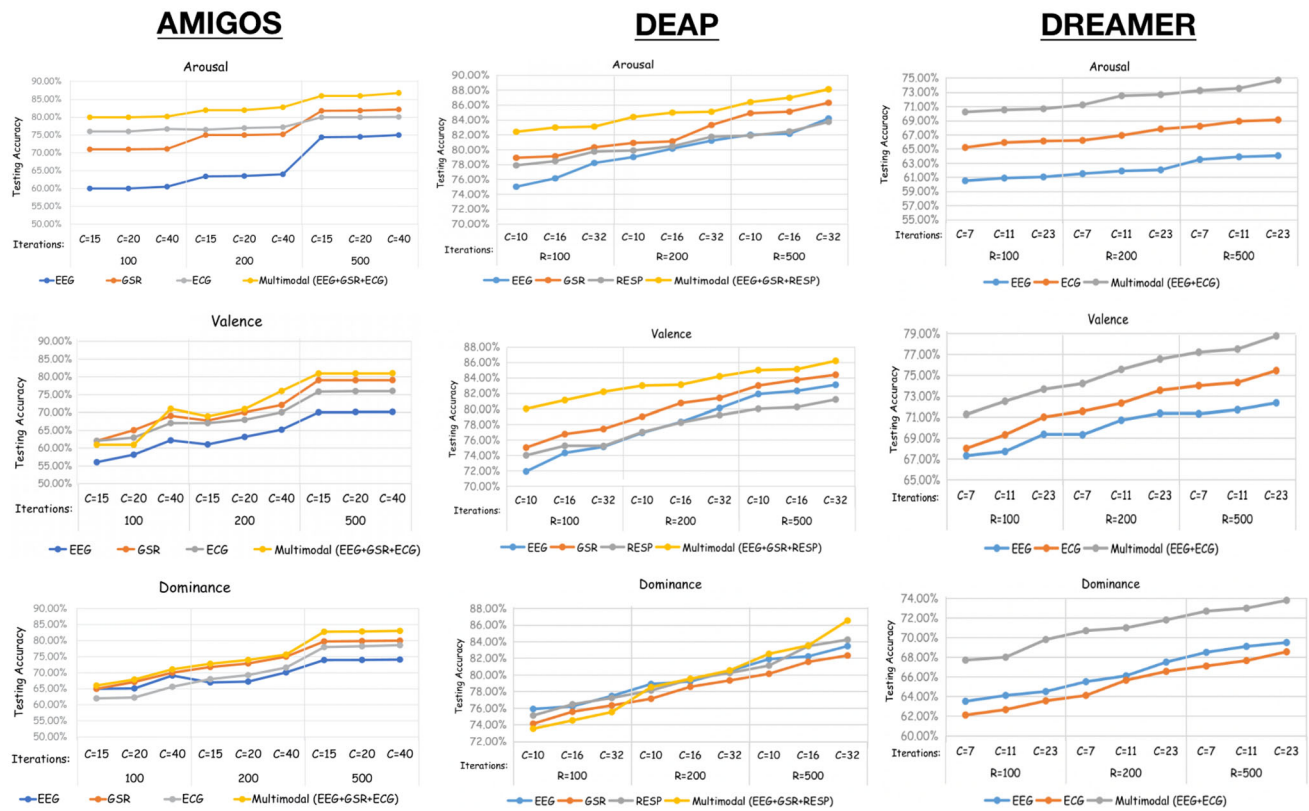
### 4.1 Subject-dependent results

This section presents the experimental results of the proposed F-MERS for the subject-dependent scenario with all three datasets AMIGOS, DEAP and DREAMER.

Figure 6 shows the graphical representation of accuracy scores of the proposed F-MERS framework with all rounds of aggregation for all the modalities. It shows that the proposed multi-modal framework performs best for all three emotional states when rounds=500 among the rest of the rounds.

Table 10 compares the binary classification performance of the proposed F-MERS framework with different client distributions for the global model. The results consist of the best outcomes obtained, i.e., with 500 rounds for all three emotional states. It clearly illustrates that the multi-modal framework performs better than single modalities, validated by different client distributions. The different client distributions represent the ability of the proposed framework to handle large amounts of data without compromising its performance, making it scalable. The proposed F-MERS achieves an accuracy of 88.10% (arousal), 86.20% (valence), and 86.52% (dominance) with the DEAP dataset for client=32, 86.80% (arousal), 80.98% (valence), and 83.06% (dominance) with the AMIGOS dataset for client=40, 74.66% (arousal), 78.12% (valence), and 73.80% (dominance) with the DREAMER dataset for client=23, for the multi-modal physiological sensors.

The proposed work also compares the efficacy of a federated learning environment (F-MERS) to non-federated learning (conventional deep learning MLP) for emotion recognition based on physiological sensors. The objective is to achieve comparability between the federated paradigm and the non-federated deep learning MLP model (as shown in Table 11) along with the addition of data privacy considerations. The proposed multi-modal FL framework achieves an average accuracy of 86.94% (vad), which is comparable with multi-modal Non-FL achieving 87.10% (vad) with the DEAP dataset for client = 32. For the AMIGOS dataset, the proposed multi-modal FL framework achieves an average accuracy of 83.61% (vad), which is comparable with multi-modal Non-FL achieving 83.64% (vad) with client = 40. For the DREAMER dataset, the proposed multi-modal FL framework achieves an average accuracy of 75.39% (vad), which is comparable



**Fig. 6** Testing accuracies of the global model for all three labels with different client distribution & different no. of rounds for the proposed F-MERS framework for all the three datasets in **Subject-dependent** scenario

with multi-modal Non-FL achieving 75.61% (vad) with client = 23.

Relative to these results, the confusion matrix is presented in Table 12 for a better presentation of the classification task performed.

## 4.2 Subject-independent results

This section presents the experimental results of the proposed F-MERS for the subject-independent scenario with all three datasets AMIGOS, DEAP and DREAMER.

Figure 7 shows the graphical representation of accuracy scores of the proposed F-MERS framework with all rounds of aggregation for all the modalities. It shows that the proposed multi-modal framework performs best for all three emotional states when rounds=500 among the rest of the rounds. It clearly illustrates that the multi-modal framework performs better than single modalities, validated by different client distributions. Hence, we present the results of the best outcomes obtained, i.e., with the multi-modal physiological data for 500 rounds for all three emotional states in Table 13.

Table 13 compares the binary classification performance of the proposed F-MERS framework with different client distributions in the subject-independent scenario for

multi-modal physiological data. The different client distributions represent the ability of the proposed framework to handle large amounts of data without compromising its performance, making it scalable. The proposed F-MERS achieves an accuracy of 86.50% (arousal), 89.02% (valence), and 84.02% (dominance) with the DEAP dataset for client=32. 87.90% (arousal), 82.10% (valence), and 82.06% (dominance) with the AMIGOS dataset for client=40. 74.33% (arousal), 79.02% (valence), and 72.24% (dominance) with the DREAMER dataset for client=23, for the multi-modal physiological sensors.

The proposed work compares the efficacy of a federated learning environment (F-MERS) to non-federated learning (Non-FL MLP) for emotion recognition based on physiological sensors in subject-independent scenarios. The objective is to compare the federated paradigm and the non-federated deep learning MLP model (as shown in Table 14) and add data privacy considerations. The proposed multi-modal FL framework achieves an average accuracy of 86.51% (vad), which is comparable with multi-modal Non-FL achieving 86.72% (vad) with the DEAP dataset for client=32. For the AMIGOS dataset, the proposed multi-modal FL framework achieves an average accuracy of 84.02% (vad), which is comparable with multi-modal Non-FL achieving 83.45% (vad) with client=40.

**Table 10** Accuracy and F1-Score Results (while Testing) for the proposed F-MERS framework in different client distributions for 500 Rounds of iterations with all three datasets in *Subject-dependent* scenario

Metrics	Accuracy			F1-Score		
(AMIGOS)						
PS <sup>1</sup> /Clients(C)	C = 15 (%)	C = 20 (%)	C = 40 (%)	C = 15 (%)	C = 20	C = 40
(Arousal)						
EEG	74.40	74.50	75.00	0.741	0.742	0.758
GSR	81.80	81.89	82.20	0.811	0.821	0.834
ECG	80.00	80.00	80.10	0.801	0.803	0.811
Multimodal (EEG+GSR+ECG)	<b>86.00</b>	<b>86.00</b>	<b>86.80</b>	<b>0.851</b>	<b>0.851</b>	<b>0.854</b>
(Valence)						
EEG	70.00	70.11	70.15	0.691	0.698	0.704
GSR	79.00	79.00	79.03	0.791	0.796	0.801
ECG	75.80	75.92	76	0.759	0.76	0.762
Multimodal (EEG+GSR+ECG)	<b>80.90</b>	<b>80.90</b>	<b>80.98</b>	<b>0.81</b>	<b>0.81</b>	<b>0.811</b>
(Dominance)						
EEG	73.97	74.00	74.10	0.741	0.743	0.755
GSR	79.78	79.89	80.00	0.798	0.798	0.801
ECG	78.00	78.26	78.60	0.781	0.78	0.786
Multimodal (EEG+GSR+ECG)	<b>82.80</b>	<b>82.88</b>	<b>83.06</b>	<b>0.825</b>	<b>0.829</b>	<b>0.833</b>

Metrics	Accuracy			F1-Score		
(DEAP)						
PS <sup>1</sup> /Clients(C)	C = 10 (%)	C = 16 (%)	C = 32 (%)	C = 10	C = 16	C = 32
(Arousal)						
EEG	82.00	82.13	84.20	0.822	0.831	0.837
GSR	84.90	85.11	86.31	0.851	0.841	0.859
RESP	81.88	82.45	83.73	0.811	0.825	0.833
Multimodal (EEG+GSR+RESP)	<b>86.40</b>	<b>86.98</b>	<b>88.10</b>	<b>0.853</b>	<b>0.868</b>	<b>0.876</b>
(Valence)						
EEG	81.91	82.31	83.10	0.818	0.831	0.802
GSR	83.00	83.74	84.40	0.829	0.842	0.842
RESP	80.00	80.23	81.20	0.818	0.822	0.821
Multimodal (EEG+GSR+RESP)	<b>85.00</b>	<b>85.12</b>	<b>86.20</b>	<b>0.854</b>	<b>0.865</b>	<b>0.864</b>
(Dominance)						
EEG	81.89	82.21	83.45	0.825	0.823	0.843
GSR	80.12	81.56	82.32	0.814	0.821	0.822
RESP	81.11	83.45	84.22	0.804	0.833	0.841
Multimodal (EEG+GSR+RESP)	<b>82.52</b>	<b>83.52</b>	<b>86.52</b>	<b>0.819</b>	<b>0.834</b>	<b>0.861</b>

Metrics	Accuracy			F1-Score		
(DREAMER)						
PS <sup>1</sup> /Clients(C)	C = 7 (%)	C = 11 (%)	C = 23 (%)	C = 7	C = 11	C = 23
(Arousal)						
EEG	63.50	63.88	64.04	0.631	0.647	0.651

**Table 10** (continued)

Metrics	Accuracy			F1-Score		
(DREAMER)						
PS <sup>1</sup> /Clients(C)	C = 7 (%)	C = 11 (%)	C = 23 (%)	C = 7	C = 11	C = 23
ECG	68.20	68.90	69.10	0.685	0.693	0.701
Multimodal (EEG+ECG)	<b>73.21</b>	<b>73.50</b>	<b>74.66</b>	<b>0.736</b>	<b>0.748</b>	<b>0.761</b>
(Valence)						
EEG	71.30	71.70	72.35	0.723	0.722	0.732
ECG	74.07	74.30	75.43	0.731	0.744	0.767
Multimodal (EEG+ECG)	<b>77.20</b>	<b>77.50</b>	<b>78.12</b>	<b>0.763</b>	<b>0.777</b>	<b>0.782</b>
(Dominance)						
EEG	68.50	69.10	69.50	0.661	0.698	0.701
ECG	67.10	67.65	68.55	0.681	0.671	0.691
Multimodal (EEG+ECG)	<b>72.70</b>	<b>73.00</b>	<b>73.80</b>	<b>0.711</b>	<b>0.732</b>	<b>0.735</b>

<sup>1</sup> Physiological Signal(PS)**Table 11** Comparison of average Testing accuracy values for all labels for the proposed federated F-MERS (With FL) and non-federated framework (Non-FL) with all three datasets in *Subject-dependent* scenario with maximum no. of clients

Environment Epochs(E)/Rounds(R) PS <sup>1</sup>	Non FL E = 500 (Arousal) (%)	With FL R = 500 (%)	Non FL E = 500 (Valence) (%)	With FL R = 500 (%)	Non FL E = 500 (Dominance) (%)	With FL R = 500 (%)
AMIGOS (with clients = 40)						
EEG	75.10	75.00	70.35	70.15	74.15	74.10
GSR	82.30	82.20	79.11	79.03	80.11	80.00
ECG	80.11	80.0	76.0	76.0	78.55	78.60
Multimodal (EEG+GSR+ECG)	<b>86.81</b>	<b>86.80</b>	<b>81</b>	<b>80.98</b>	<b>83.11</b>	<b>83.06</b>
DEAP (with clients=32)						
EEG	84.51	84.20	83.18	83.10	83.80	83.45
GSR	86.52	86.31	84.53	84.40	82.45	82.32
RESP	83.92	83.73	81.28	81.20	84.30	84.22
Multimodal (EEG+GSR+RESP)	<b>88.20</b>	<b>88.10</b>	<b>86.34</b>	<b>86.20</b>	<b>86.78</b>	<b>86.52</b>
DREAMER (with clients=23)						
EEG	65.00	64.04	72.35	72.15	69.50	69.12
ECG	70.11	69.10	75.10	75.20	68.55	68.23
Multimodal (EEG+ECG)	<b>74.81</b>	<b>74.66</b>	<b>78.23</b>	<b>78.10</b>	<b>73.80</b>	<b>73.43</b>

<sup>1</sup> Physiological Signal(PS)

Moreover, for the DREAMER dataset, the proposed multimodal FL framework achieves an average accuracy of 75.19% (vad), which is comparable with multi-modal Non-FL achieving 75.64% (vad) with client=23.

The confusion matrix is presented in Table 15 for a better presentation of the classification task performed.

Table 16 assesses the scalability and communication efficiency of the proposed framework. For measuring scalability, we compute the training time for classification, which implies that the proposed approach can handle

different data distributions well with acceptable training times, making the proposed framework scalable. Furthermore, we compute the time taken for aggregation in the FL environment to measure the communication efficiency of the model reflected by different rounds of aggregations, which converges at R=500 with optimal performance. The time differences in training and aggregation time are very minute for subject-dependent and subject-independent scenarios. Table 16 presents the average training time of both scenarios.

## 5 Discussion

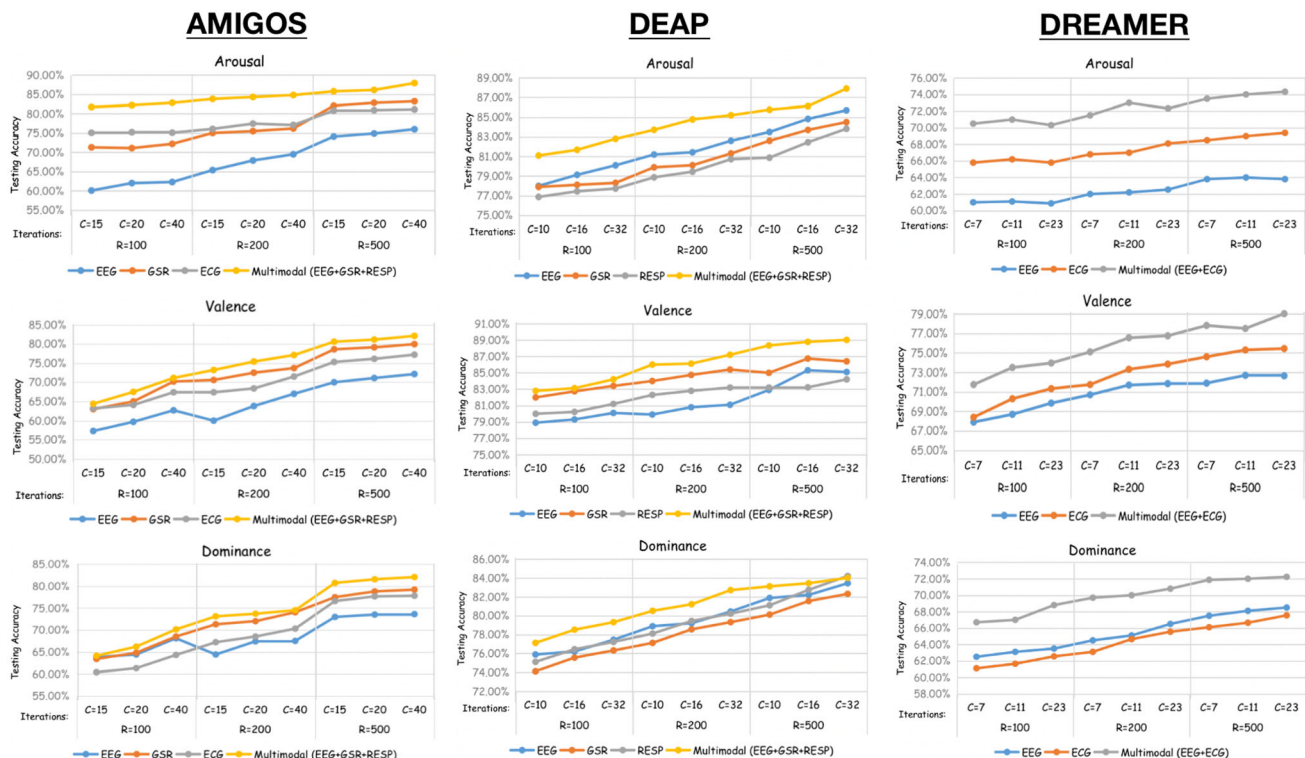
- Proposed FL framework in both subject-dependent and subject-independent scenario:** The proposed study aims to validate the FL framework F-MERS through experiments conducted in two different

**Table 12** Confusion Matrix for testing the global model for all three labels with all three datasets for the multi-modal proposed F-MERS framework in **subject-dependent** scenario

Valence			Arousal			Dominance		
Class	Low	High	Class	Low	High	Class	Low	High
<b>AMIGOS</b>								
Low	0.154	0.085	Low	0.191	0.058	Low	0.154	0.068
High	0.108	0.653	High	0.076	0.675	High	0.104	0.674
<b>DEAP</b>								
Low	0.242	0.059	Low	0.254	0.049	Low	0.242	0.059
High	0.078	0.621	High	0.069	0.628	High	0.077	0.622
<b>DREAMER</b>								
Low	0.194	0.109	Low	0.186	0.123	Low	0.175	0.121
High	0.108	0.589	High	0.129	0.562	High	0.142	0.562

scenarios: Subject-Dependent (SD) and Subject-Independent (SID). The results obtained in these scenarios show that the performance of the F-MERS framework is slightly better (1-2%) in the Subject-independent scenario compared to the Subject-dependent scenario. However, it is noteworthy that the performance difference between the two scenarios is very minute, which is validated by all three datasets. It indicates (Tables 10, 11 and 13, Table 14) that the proposed F-MERS framework performs well in both scenarios and proves to be robust and generalized.

- Proposed FL framework comparison with existing FL work for emotion:** To compare the proposed framework with the existing works in FL for emotion recognition, the same evaluation grounds are included, i.e. same datasets and validation approach. Table 17 compares the proposed framework F-MERS with previous work based on FL for emotion recognition using physiological sensors. The results in the table conclude that the proposed work with multi-modal physiological sensors' data is outperforming the previous works in both the validation scenarios of subject-dependent (SD) and subject-independent (SID) in terms of accuracy, communication efficiency, and scalability.



**Fig. 7** Testing accuracies of the global model for all three labels with different client distribution & different no. of rounds for the proposed F-MERS framework for all the three datasets in **Subject-independent** scenario



**Table 13** Accuracy and F1-Score Results (while Testing) for the proposed F-MERS framework in different client distributions for 500 Rounds of iterations with all three datasets in **Subject-independent** scenario

Metrics	Accuracy			F1-Score		
(AMIGOS)						
PS <sup>1</sup> /Clients (C)	C = 15 (%)	C = 20 (%)	C = 40 (%)	C = 15	C = 20	C = 40
(Arousal)						
Multimodal (EEG+GSR+ECG)	85.76	86.22	87.90	0.805	0.882	0.891
(Valence)						
Multimodal (EEG+GSR+ECG)	80.58	81.12	82.10	0.795	0.791	0.813
(Dominance)						
Multimodal (EEG+GSR+ECG)	80.73	81.56	82.06	0.822	0.812	0.824
Metrics	Accuracy			F1-Score		
(DEAP)						
PS <sup>1</sup> /Clients	C = 10 (%)	C = 16 (%)	C = 32 (%)	C = 10	C = 16	C = 32
(Arousal)						
Multimodal (EEG+GSR+RESP)	85.22	85.72	86.50	0.864	0.878	0.888
(Valence)						
Multimodal (EEG+GSR+RESP)	88.34	88.78	89.02	0.891	0.883	0.893
(Dominance)						
Multimodal (EEG+GSR+RESP)	83.12	83.45	84.02	0.841	0.833	0.854
Metrics	Accuracy			F1-Score		
(DREAMER)						
PS <sup>1</sup> /Clients	C = 7 (%)	C = 11 (%)	C = 23 (%)	C = 7	C = 11	C = 23
(Arousal)						
Multimodal (EEG+ECG)	73.50	74.01	74.33	0.702	0.749	0.768
(Valence)						
Multimodal (EEG+ECG)	77.81	78.52	79.02	0.789	0.801	0.804
(Dominance)						
Multimodal (EEG+ECG)	71.88	72.01	72.24	0.722	0.744	0.761

<sup>1</sup> Physiological Signal(PS)

## 6 Lessons learnt

Based on the experiments performed and findings, the learning for the research questions (as mentioned in Sect. 1.1) are as follows:

- **MQ1: Why is privacy essential for physiological data?**

**Soln:** Physiological data is inherently personal and highly sensitive, encompassing signals such as EEG, ECG, RESP, GSR, Heart Rate, and others. Data from these signals are unique to each individual, revealing a lot about their health status, potentially providing

insights into their emotional state, behaviour, and habits and hence is considered sensitive. Compromising the privacy of physiological data can have profound repercussions on an individual's life. It opens the door for data attackers and exposes them to the risk of data breaches, which, in turn, can lead to various threats [28]. These threats include the potential exposure of an individual's health status, emotional stability, and even their biometric identity based on physiological signals. Therefore, safeguarding privacy is crucial when it comes to handling physiological data. For this, the proposed study ensures that physiological sensor data

**Table 14** Comparison of average Testing accuracy values for all labels for the proposed federated F-MERS (With FL) and non-federated framework (Non-FL) with all three datasets in **Subject-independent** scenario with maximum no. of clients

Environment Epochs(E)/Rounds(R) AMIGOS (with clients = 40)	Non FL E = 500	With FL R = 500	Non FL E = 500	With FL R = 500	Non FL E = 500	With FL R = 500
PS <sup>1</sup>	(Arousal) (%)		(Valence) (%)		(Dominance) (%)	
EEG	76.21	75.09	71.15	70.95	73.76	73.15
GSR	82.45	82.20	79.91	79.46	80.11	80.00
ECG	81.56	81.14	76.35	76.20	77.86	77.50
Multimodal (EEG+GSR+ECG)	<b>88.12</b>	<b>87.90</b>	<b>81.81</b>	<b>82.10</b>	<b>82.23</b>	<b>82.06</b>

Environment Epochs(E)/Rounds(R) DEAP (with clients=32)	Non FL E = 500	With FL R = 500	Non FL E = 500	With FL R = 500	Non FL E = 500	With FL R = 500
PS <sup>1</sup>	(Arousal) (%)		(Valence) (%)		(Dominance) (%)	
EEG	84.89	84.45	85.67	85.53	84.67	84.55
GSR	85.25	85.16	85.53	85.89	81.45	81.29
RESP	82.12	82.24	82.12	82.06	80.56	80.34
Multimodal (EEG+GSR+RESP)	<b>86.88</b>	<b>86.50</b>	<b>89.12</b>	<b>89.02</b>	<b>84.16</b>	<b>84.02</b>

Environment Epochs(E)/Rounds(R) DREAMER (with clients = 23)	Non FL E = 500	With FL R = 500	Non FL E = 500	With FL R = 500	Non FL E = 500	With FL R = 500
PS <sup>1</sup>	(Arousal) (%)		(Valence) (%)		(Dominance) (%)	
EEG	66.10	64.78	74.53	74.21	68.87	68.43
ECG	71.05	70.10	77.86	77.68	67.88	67.10
Multimodal (EEG+ECG)	<b>75.02</b>	<b>74.33</b>	<b>79.23</b>	<b>79.02</b>	<b>72.68</b>	<b>72.24</b>

<sup>1</sup> Physiological Signal(PS)

for emotion recognition is used to protect people's privacy and autonomy.

- **SQ1:How can complex emotions be mapped into different dimensions?**

**Soln:** Complex emotions like fear, wrath, guilt, resentment, anxiety, and others have different levels of arousal, valence and dominance, are difficult to distinguish and cannot be mapped on the 2-dimensional emotion model by Russel [45]. Hence, the proposed study adopted Mehrabian's 3-dimensional model of emotions to map these complex emotions [46, 47]. The 3-dimensional model of emotion maps complex emotions into three dimensions: Arousal, Valence and Dominance. It understands better how emotions are experienced and influence one's behaviour. The 3-dimensional model of emotions maps emotions as arousal refers to the physiological activation or intensity level of emotions, ranging from low arousal (calm, relaxed) to high arousal (excited, anxious). Valence represents the pleasantness or unpleasantness of emotions, ranging from positive (happiness, joy) to negative (anger,

sadness). Dominance reflects an emotion's sense of control or power, ranging from feeling dominant (empowered, in control) to feeling submissive (helpless, powerless). This information develops better ways to cope with complex emotions and build stronger relationships.

- **SQ2: How physiological signals contribute to emotion recognition?**

**Soln:** Physiological signals provide valuable information about the body's physiological responses and indicate different emotional states.

- EEG signals are electric impulses recorded to analyze brain functions [57, 58]. The brain controls all of the emotional behaviours of people, including physical movement, sensory processing, language & communication, memory, and emotions.
- GSR measures the skin's electrical conductivity and is also known as Electrodermal Activity (EDA) [59–61]. Skin conductivity varies with skin moisture level (sweating), showing variations in the

- Autonomous Nerve System associated with arousal, reflecting emotions such as stress, anxiety, and surprise. It is, in particular, a measure of arousal.
- ECG signals are electric signals acquired to trace the action of the human heart and the potential fluctuations transmitted to the skin surface due to the heart's electrical activity (the contraction and

relaxation of heart muscles) [62, 63]. Electrodes linked to the skin surface detect it.

- The respiratory rate (RESP) physiological signal represents the frequency of breaths a person takes per minute, mirroring the inhaling and exhaling air rate. These breathing patterns closely connect to the emotional states [37]. When experiencing emotions like stress, fear, anxiety, or excitement, breathing quickens and becomes shallower. In contrast, during moments of relaxation and calmness, our breathing becomes slower and deeper. This interplay between respiratory rates highlights the importance of conscious breathing techniques in managing emotions [64]

• **SQ3: Why is multi-modality required in emotion recognition?**

**Soln:** Here, multi-modality refers to the fusion or combination of different physiological signals required for emotion recognition. These different physiological signals provide complementary information about emotions. Multiple physiological signals capture a more comprehensive picture of emotional states and increase emotion recognition accuracy [17]. Emotion recognition systems that rely on a single physiological signal

**Table 15** Confusion Matrix for testing the global model for all three labels with all three datasets for the multi-modal proposed F-MERS framework in **subject-independent** scenario

Valence			Arousal			Dominance		
Class	Low	High	Class	Low	High	Class	Low	High
AMIGOS								
Low	0.154	0.085	Low	0.191	0.045	Low	0.154	0.068
High	0.094	0.667	High	0.076	0.688	High	0.112	0.666
DEAP								
Low	0.242	0.051	Low	0.254	0.049	Low	0.212	0.069
High	0.059	0.648	High	0.089	0.608	High	0.091	0.628
DREAMER								
Low	0.194	0.129	Low	0.206	0.101	Low	0.17	0.15
High	0.128	0.549	High	0.109	0.584	High	0.13	0.55

**Table 16** Average values (for both the subject-dependent and subject-independent scenario) of Aggregation Time and Training time for all the data distribution and iterations of the proposed F-MERS

(AMIGOS)									
Measures (in seconds)	C = 15			C = 20			C = 40		
	R = 100	R = 200	R = 500	R = 100	R = 200	R = 500	R = 100	R = 200	R = 500
Training Time	99.157	194.701	508.756	99.427	198.958	510.943	103.376	215.104	520.434
Aggregation Time	76.458	152.774	398.618	84.924	155.722	410.698	90.293	162.842	419.394
(DEAP)									
Measures (in seconds)	C = 10			C = 16			C = 32		
	R = 100	R = 200	R = 500	R = 100	R = 200	R = 500	R = 100	R = 200	R = 500
Training Time	130.412	220.011	550.123	145.411	240.812	568.132	158.631	260.421	596.287
Aggregation Time	90.812	182.434	410.509	100.555	198.231	450.918	115.342	190.412	470.314
(DREAMER)									
Measures (in seconds)	C = 7			C = 11			C = 23		
	R = 100	R = 200	R = 500	R = 100	R = 200	R = 500	R = 100	R = 200	R = 500
Training Time	110.609	200.354	518.145	120.512	220.989	540.456	125.39	235.799	550.422
Aggregation Time	77.512	155.445	400.867	86.432	160.254	415.821	92.367	170.256	421.443
Clients (C)									
Rounds (R)									

**Table 17** Comparison of proposed F-MERS framework with previous work based on FL for emotion recognition

Ref	Dataset	PS <sup>1</sup>	CM <sup>2</sup>	Tool	Algorithm	Avg. Accuracy	Modality	Validation	CE <sup>3</sup>	S <sup>4</sup>
Nandi et al.[39]	DEAP[37]	EDA+RB	FFNN	TFF	FedAvg	81.92% (VA)	Bi-Modal	SID	×	✓
Tara Hassani [40]	CASE[41]	GSR	CNN	TFF	FedAvg	79% (V), 68% (A)	Single	SD	×	×
Gao et al.[42]	MindBigData[42]	EEG	CNN - FC	PySfyt	FedAvg	86 %	Single	SD	×	×
Ayaan et al.[27]	DREAMER[38]	EEG	ANN	TFF	FedAvg	63.33% (V), 56.7% (A), 52.2% (D)	Single	SD	×	×
Proposed Work (F-MERS)	DEAP[36]	EEG+GSR+ RESP	MLP	TFF	FedAvg	89.02% (V), 86.50% (A), 84.02% (D)	Multi-Modal	SID	✓	✓
	DREAMER	EEG+ECG				79.02% (V), 74.33% (A), 72.24% (D)		SID		
	AMIGOS	EEG+ECG+ GSR				80.10% (V), 87.90% (A), 81.06% (D)		SID		
Proposed Work (F-MERS)	DEAP[36]	EEG+GSR+ RESP	MLP	TFF	FedAvg	86.20% (V), 88.10% (A), 86.52% (D)	Multi-Modal	SD	✓	✓
	DREAMER	EEG+ECG				78.10% (V), 74.66% (A), 73.43% (D)		SD		
	AMIGOS	EEG+ECG+GSR				80.98% (V), 86.80% (A), 83.06% (D)		SD		

<sup>1</sup> Physiological Signals (PS)<sup>2</sup> Classification Model (CM)<sup>3</sup> Communication Efficiency (CE)<sup>4</sup> Scalability (S)<sup>5</sup> Subject Independent (SID)<sup>6</sup> Subject Dependent (SD)

are vulnerable to noise, artefacts, or biases inherent in that particular modality (physiological signal). Combining multiple modalities (physiological signals) creates more robust and adaptable systems capable of recognizing different emotions more accurately and precisely. In response, the proposed study combines the participating subjects' EEG, ECG, GSR and RESP signals.

- **SQ4: How can machine learning be used for automated emotion recognition systems?**

**Soln:** Physiological signals data (EEG, ECG, GSR, RESP) obtained from different wearable sensors for recognizing emotions is preprocessed, from which the relevant features are extracted. The relevant features extracted are the inputs to train the Machine Learning and Deep Learning algorithms like SVM [18], DT [19], and KNN [20, 21], RNN [22], CNN [23, 24], and LSTM [25] for classifying the different emotion states. These algorithms are successful in attaining higher accuracies with physiological sensor data. It is worth noting that the success of these automated machine learning-based emotion recognition systems depends on the quality and diversity of the training data, the choice of relevant features, and the selection and optimization of the machine learning algorithm. The proposed study accommodates this by implementing an MLP classifier for emotion state classification.

- **SQ5: How federated learning paradigm is preserving data privacy in emotion recognition?**

**Soln:** The traditional ML and DL architectures require complete access to the physiological data for training the model in an automated emotion recognition system. It compromises the privacy of the data as it requires complete access to physiological data for training purposes, giving easy access to data attackers. A new paradigm called Federated Learning (FL) is introduced by McMahan et al. [26] to resolve the issue of data privacy. FL is a promising approach which creates a decentralized environment with a local and global model at the client and server end, respectively [32]. It allows the local model updates to be sent to a central server, combining them to create a global model [29–31]. This approach does not allow the global model to access the raw data used for training and hence preserves the privacy of the sensitive physiological data. The proposed study accommodates this approach by proposing the F-MERS architecture for emotion recognition, preserving the privacy of the sensitive physiological data while achieving good accuracy results.

## 7 Conclusion

Federated learning is essential for emotion recognition with physiological data because it enables data privacy, communication efficiency, and scalability, all while improving the accuracy and generalizability of emotion recognition models. This study proposes a novel FL-based Multi-modal Emotion Recognition System (F-MERS) framework to successfully and accurately classify human emotional states while protecting sensitive physiological information. The proposed framework improves prior work in emotion recognition by generating a federated environment using federated averaging (FedAvg) at the server. The training and classification are performed at the client's end to protect data privacy from data breaches and sensitive information scenarios by not sharing the complete raw data (available at the clients' end) with other entities and the global server. The proposed framework's contributions are prominently evident from the results stating that the multi-modal framework outperforms single modalities, including EEG, ECG, GSR, and RESP. The proposed experiments unveil that dominance, arousal, and valence play a pivotal role in recognizing complex human emotions. This contribution of the proposed study provides valuable insights into a deeper understanding of human emotions. The three datasets (AMIGOS [36], DEAP [37] and DREAMER [38]) validate the results for different iterations and varying rounds concluding the model to be efficient communication-wise, scalable, and performs accurately. It disagrees with the prior emotion recognition works in that they have not considered the privacy concerns for the user's physiological data. The proposed study concludes that emotion recognition with a single modality is less accurate than multi-modal sensor data. Hence, the proposed FL-enabled multi-modal emotion recognition system can assist in better personalized emotional care with the security of personal data privacy while dealing with emotional distress.

The proposed study has some restrictions, opening the area for future investigations and experimentation. These are: (1). To experiment with the decision-level fusion of the different modalities (the proposed study adopts feature-level fusion), (2). Combining the physiological indicators with other physical indicators of emotions like eye-tracking [6, 7], speech [4] and gesture [5] can be a more generalized approach for emotion detection. The proposed FL framework can be extended in the future for the fusion of physical with physiological indicators (the proposed study adopts only physiological indicators).

**Funding** The authors have no competing interests to declare that are relevant to the content of this article.

**Data availability** Data sharing is not applicable to this article as no datasets were generated during the current study. The datasets mentioned in the current study are cited with their base papers and their links are also given in the paper.

## Declarations

**Conflict of interest** The authors have no relevant financial or non-financial interests to disclose.

## References

- Salovey, P., Rothman, A.J., Detweiler, J.B., Steward, W.T.: Emotional states and physical health. *Am. Psychol.* **55**(1), 110 (2000). <https://doi.org/10.1037/0003-066X.55.1.110>
- Zhang, Y.-D., Yang, Z.-J., Lu, H.-M., Zhou, X.-X., Phillips, P., Liu, Q.-M., Wang, S.-H.: Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation. *IEEE Access* **4**, 8375–8385 (2016). <https://doi.org/10.1109/ACCESS.2016.2628407>
- Alhussein, M.: Automatic facial emotion recognition using weber local descriptor for e-Healthcare system. *Clust. Comput.* **19**, 99–108 (2016). <https://doi.org/10.1007/s10586-016-0535-3>
- Mao, Q., Dong, M., Huang, Z., Zhan, Y.: Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Trans. Multimed.* **16**(8), 2203–2213 (2014). <https://doi.org/10.1109/TMM.2014.2360798>
- Gunes, H., Piccardi, M.: Bi-modal emotion recognition from expressive face and body gestures. *J. Netw. Comput. Appl.* **30**(4), 1334–1345 (2007). <https://doi.org/10.1016/j.jnca.2006.09.007>
- Yang, M., Cai, C., Hu, B.: Clustering based on eye tracking data for depression recognition. *IEEE Trans. Cognit. Dev. Syst.* (2022). <https://doi.org/10.1109/TCDS.2022.3223128>
- Yang, M., Feng, X., Ma, R., Li, X., Mao, C.: Orthogonal-moment-based attraction measurement with ocular hints in video-watching task. *IEEE Trans. Comput. Soc. Syst.* (2023). <https://doi.org/10.1109/TCSS.2023.3268505>
- Li, Z., Tian, X., Shu, L., Xu, X., Hu, B.: Emotion recognition from EEG using RASM and LSTM. In: *Internet Multimedia Computing and Service: 9th International Conference, ICIMCS 2017, Qingdao, China, August 2017. Revised Selected Papers 9*, pp. 310–318. Springer (2018) [https://doi.org/10.1007/978-981-10-8530-7\\_30](https://doi.org/10.1007/978-981-10-8530-7_30)
- Malviya, L., Mal, S.: Cis feature selection based dynamic ensemble selection model for human stress detection from eeg signals. *Clust. Comput.* (2023). <https://doi.org/10.1007/s10586-023-04008-8>
- Gahlan, N., Sethia, D.: Three dimensional emotion state classification based on EEG via empirical mode decomposition. In: *2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1)*, pp. 1–6 (2023). <https://doi.org/10.1109/ICAIA57370.2023.10169633>
- Valenza, G., Citi, L., Lanatà, A., Scilingo, E.P., Barbieri, R.: Revealing real-time emotional responses: a personalized assessment based on heartbeat dynamics. *Sci. Rep.* **4**(1), 1–13 (2014). <https://doi.org/10.1038/srep04998>
- Benedek, M., Kaernbach, C.: A continuous measure of phasic electrodermal activity. *J. Neurosci. Methods* **190**(1), 80–91 (2010). <https://doi.org/10.1016/j.jneumeth.2010.04.028>
- Peter, C., Ebert, E., Beikirch, H.: A wearable multi-sensor system for mobile acquisition of emotion-related physiological data. In: *International Conference on Affective Computing and Intelligent*



- Interaction, pp. 691–698. Springer (2005). [https://doi.org/10.1007/11573548\\_89](https://doi.org/10.1007/11573548_89)
14. Schmidt, P., Reiss, A., Duerichen, R., Van Laerhoven, K.: Wearable affect and stress recognition: a review. *Hum.-Comput. Interact.* (2018). <https://doi.org/10.48550/arXiv.1811.08854>
  15. Krumova, E.K., Frettlöh, J., Klauenberg, S., Richter, H., Wasner, G., Maier, C.: Long-term skin temperature measurements—a practical diagnostic tool in complex regional pain syndrome. *Pain* **140**(1), 8–22 (2008). <https://doi.org/10.1016/j.pain.2008.07.003>
  16. Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., Xu, X., Yang, X.: A review of emotion recognition using physiological signals. *Sensors* **18**(7), 2074 (2018). <https://doi.org/10.3390/s18072074>
  17. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S.: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: *Proceedings of the 6th International Conference on Multimodal Interfaces*, pp. 205–211 (2004). <https://doi.org/10.1145/1027933.1027968>
  18. Tuncer, et al.: LEDPatNet19: automated emotion recognition model based on nonlinear LED pattern feature extraction function using EEG signals. *Cognitive Neurodynamics* (2021). <https://doi.org/10.1007/s11571-021-09748-0>
  19. Cheng, J., Chen, M., Li, C., Liu, Y., Song, R., Liu, A., Chen, X.: Emotion recognition from multi-channel EEG via deep forest. *IEEE J. Biomed. Health Inf.* **25**(2), 453–464 (2020). <https://doi.org/10.1109/JBHI.2020.2995767>
  20. Nasoz, F., Alvarez, K., Lisetti, C.L., Finkelstein, N.: Emotion recognition from physiological signals using wireless sensors for presence technologies. *Cognition Technol. Work* **6**(1), 4–14 (2004). <https://doi.org/10.1007/s10111-003-0143-x>
  21. Zhou, Z., Asghar, M.A., Nazir, D., Siddique, K., Shorfuazzaman, M., Mehmood, R.M.: An AI-empowered affect recognition model for healthcare and emotional well-being using physiological signals. *Clust. Comput.* **26**(2), 1253–1266 (2023). <https://doi.org/10.1007/s10586-022-03705-0>
  22. Duan, R.-N., Zhu, J.-Y., Lu, B.-L.: Differential entropy feature for EEG-based emotion classification. In: *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 81–84 (2013). <https://doi.org/10.1109/NER.2013.6695876>. IEEE
  23. Meng, M., Zhang, Y., Ma, Y., Gao, Y., Kong, W.: EEG-based emotion recognition with cascaded convolutional recurrent neural networks. *Pattern Anal. Appl.* (2023). <https://doi.org/10.1007/s10044-023-01136-0>
  24. Dar, M.N., Akram, M.U., Khawaja, S.G., Pujari, A.N.: CNN and LSTM-based emotion charting using physiological signals. *Sensors* **20**(16), 4551 (2020). <https://doi.org/10.3390/s20164551>
  25. Tang, H., Liu, W., Zheng, W.-L., Lu, B.-L.: Multimodal emotion recognition using deep neural networks. In: *International Conference on Neural Information Processing*, pp. 811–819 (2017). Springer [https://doi.org/10.1007/978-3-319-70093-9\\_86](https://doi.org/10.1007/978-3-319-70093-9_86)
  26. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: *Artificial Intelligence and Statistics*. PMLR (2017). <https://doi.org/10.48550/arXiv.1602.05629>
  27. Anwar, M.A., Agrawal, M., Gahlan, N., Sethia, D., Singh, G.K., Chaurasia, R.: FedEmo: A privacy-preserving framework for emotion recognition using EEG physiological data. In: *2023 15th International Conference on COMMunication Systems & NETWORKS (COMSNETS)*, pp. 119–124 (2023). <https://doi.org/10.1109/COMSNETS56262.2023.10041308>. IEEE
  28. Data Breach. <https://tinyurl.com/2p8b57ax> (Accessed 3 Feb 2023)
  29. Zhu, H., Xu, J., Liu, S., Jin, Y.: Federated learning on non-IID data: a survey. *Neurocomputing* **465**, 371–390 (2021). <https://doi.org/10.1016/j.neucom.2021.07.098>
  30. Rahman, A., Hossain, M.S., Muhammad, G., Kundu, D., Debnath, T., Rahman, M., Khan, M.S.I., Tiwari, P., Band, S.S.: Federated learning-based AI approaches in smart healthcare: concepts, taxonomies, challenges and open issues. *Clust. Comput.* (2022). <https://doi.org/10.1007/s10586-022-03658-4>
  31. Xu, J., Lin, J., Liang, W., Li, K.-C.: Privacy preserving personalized blockchain reliability prediction via federated learning in IoT environments. *Clust. Comput.* **25**(4), 2515–2526 (2022). <https://doi.org/10.1007/s10586-021-03399-w>
  32. Li, D., Luo, Z., Cao, B.: Blockchain-based federated learning methodologies in smart environments. *Clust. Comput.* **25**(4), 2585–2599 (2022). <https://doi.org/10.1007/s10586-021-03424-y>
  33. Sánchez Sánchez, P.M., Huertas Celdrán, A., Buendía Rubio, J.R., Bovet, G., Martínez Pérez, G.: Robust federated learning for execution time-based device model identification under label-flipping attack. *Clust. Comput.* (2023). <https://doi.org/10.1007/s10586-022-03949-w>
  34. Zeng, R., Zeng, C., Wang, X., Li, B., Chu, X.: A comprehensive survey of incentive mechanism for federated learning. *Mach. Learn.* (2021). <https://doi.org/10.48550/arXiv.2106.15406>
  35. Rahman, K.J., Ahmed, F., Akhter, N., Hasan, M., Amin, R., Aziz, K.E., Islam, A.M., Mukta, M.S.H., Islam, A.N.: Challenges, applications and design aspects of federated learning: a survey. *IEEE Access* **9**, 124682–124700 (2021). <https://doi.org/10.1109/ACCESS.2021.3111118>
  36. Miranda-Correa, J.A., Abadi, M.K., Sebe, N., Patras, I.: Amigos: a dataset for affect, personality and mood research on individuals and groups. *IEEE Trans. Affective Comput.* **12**(2), 479–493 (2018). <https://doi.org/10.1109/TAFFC.2018.2884461>
  37. Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: Deap: a database for emotion analysis; using physiological signals. *IEEE Trans. Affective Comp.* **3**(1), 18–31 (2011). <https://doi.org/10.1109/TAFFC.2011.15>
  38. Katsigiannis, et al.: DREAMER: a database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE J. Biomed. Health Inf.* **22**(1), 98–107 (2017). <https://doi.org/10.1109/JBHI.2017.2688239>
  39. Nandi, A., Khafa, F.: A federated learning method for real-time emotion state classification from multi-modal streaming. *Methods* (2022). <https://doi.org/10.1016/j.ymeth.2022.03.005>
  40. Hassani, T.: Federated emotion recognition with physiological signals-GSR (2021)
  41. Sharma, K., Castellini, C., van den Broek, E.L., Albu-Schaeffer, A., Schwenker, F.: A dataset of continuous affect annotations and physiological signals for emotion analysis. *Sci. Data* **6**(1), 1–13 (2019). <https://doi.org/10.1038/s41597-019-0209-0>
  42. Gao, D., Ju, C., Wei, X., Liu, Y., Chen, T., Yang, Q.: Hhhfl: hierarchical heterogeneous horizontal federated learning for electroencephalography. *Signal Proc.* (2019). <https://doi.org/10.48550/arXiv.1909.05784>
  43. Yang, Y.-H., Chen, H.H.: Machine recognition of music emotion: a review. *ACM Trans. Intell. Syst. Technol. (TIST)* **3**(3), 1–30 (2012). <https://doi.org/10.1145/2168752.2168754>
  44. Balan, O., Moise, G., Petrescu, L., Moldoveanu, A., Leordeanu, M., Moldoveanu, F.: Emotion classification based on biophysical signals and machine learning techniques. *Symmetry* **12**(1), 21 (2019)
  45. Lang, P.J.: The emotion probe: studies of motivation and attention. *Am. Psychol.* **50**(5), 372 (1995). <https://doi.org/10.1037/0003-066X.50.5.372>
  46. Mehrabian, A., Russell, J.A.: An approach to environmental psychology. The MIT Press, Cambridge (1974)
  47. Russell, J.A., Mehrabian, A.: Evidence for a three-factor theory of emotions. *J. Res. Pers.* **11**(3), 273–294 (1977). [https://doi.org/10.1016/0092-6566\(77\)90037-X](https://doi.org/10.1016/0092-6566(77)90037-X)



48. Nandi, A., Xhafa, F., Subirats, L., Fort, S.: Reward-penalty weighted ensemble for emotion state classification from multimodal data streams. *Int. J. Neural Syst.* **32**(12), 2250049 (2022). <https://doi.org/10.1142/S0129065722500496>
49. Ortony, A., Clore, G.L., Collins, A.: *The cognitive structure of emotions* Cambridge, vol. 9. Cambridge University Press, Cambridge (1988)
50. Morris, J.D.: Observations: SAM: the Self-Assessment Manikin; an efficient cross-cultural measurement of emotional response. *J. Advert. Res.* **35**(6), 63–68 (1995)
51. Khaleghi, B., Khamis, A., Karray, F.O., Razavi, S.N.: Multisensor data fusion: a review of the state-of-the-art. *Inf. Fusion* **14**(1), 28–44 (2013). <https://doi.org/10.1016/j.inffus.2011.08.001>
52. Durrant-Whyte, H., Henderson, T.C.: *Multisensor data fusion*. Springer handbook of robotics, pp. 867–896. Springer, Berlin (2016). [https://doi.org/10.1007/978-3-319-32552-1\\_35](https://doi.org/10.1007/978-3-319-32552-1_35)
53. Chen, J., Hu, B., Xu, L., Moore, P., Su, Y.: Feature-level fusion of multimodal physiological signals for emotion recognition. In: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 395–399 (2015). <https://doi.org/10.1109/BIBM.2015.7359713>. IEEE
54. Google: Tensorflow federated. <https://www.tensorflow.org/federated> Accessed 1 Feb 2023)
55. El Mokadem, R., Ben Maissa, Y., El Akkaoui, Z.: Federated learning for energy constrained devices: a systematic mapping study. *Clust. Comput.* **26**(2), 1685–1708 (2023). <https://doi.org/10.1007/s10586-022-03763-4>
56. Proadhan, R.A., Akter, S., Mujib, M.B., Adnan, M.A., Pias, T.S.: Emotion recognition from brain wave using multitask machine learning leveraging residual connections. In: *International Conference on Machine Intelligence and Emerging Technologies*, pp. 121–136 (2022). [https://doi.org/10.1007/978-3-031-34622-4\\_10](https://doi.org/10.1007/978-3-031-34622-4_10). Springer
57. Lin, Y.-P., Wang, C.-H., Wu, T.-L., Jeng, S.-K., Chen, J.-H.: Multilayer perceptron for eeg signal classification during listening to emotional music. In: *TENCON 2007-2007 IEEE Region 10 Conference*, pp. 1–3 (2007). IEEE <https://doi.org/10.1109/TENCON.2007.4428831>
58. Ravi Kumar, M., Srinivasa Rao, Y.: Epileptic seizures classification in eeg signal based on semantic features and variational mode decomposition. *Clust. Comput.* **22**, 13521–13531 (2019). <https://doi.org/10.1007/s10586-018-1995-4>
59. Nourbakhsh, N., Wang, Y., Chen, F., Calvo, R.A.: Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In: *Proceedings of the 24th Australian Computer-human Interaction Conference*, pp. 420–423 (2012). <https://doi.org/10.1145/2414536.2414602>
60. Liu, M., Fan, D., Zhang, X., Gong, X.: Human emotion recognition based on galvanic skin response signal feature selection and svm. In: 2016 International Conference on Smart City and Systems Engineering (ICSCSE), pp. 157–160 (2016). <https://doi.org/10.1109/ICSCSE.2016.0051>. IEEE
61. Lang, P.J., Bradley, M.M., Cuthbert, B.N., et al.: *International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual*. NIMH, Center for the Study of Emotion & Attention Gainesville, FL, (2005)
62. Delaney, J., Brodie, D.: Effects of short-term psychological stress on the time and frequency domains of heart-rate variability. *Perceptual and motor skills* **91**(2), 515–524 (2000). <https://doi.org/10.2466/pms.2000.91.2.515>
63. Bong, S.Z., Murugappan, M., Yaacob, S.: Analysis of Electrocardiogram (ECG) signals for human emotional stress classification. In: *International Conference on Intelligent Robotics, Automation, and Manufacturing*, pp. 198–205 (2012). Springer [https://doi.org/10.1007/978-3-642-35197-6\\_22](https://doi.org/10.1007/978-3-642-35197-6_22)
64. Homma, I., Masaoka, Y.: Breathing rhythms and emotions. *Exp. Physiol.* **93**(9), 1011–1021 (2008). <https://doi.org/10.1113/expphysiol.2008.042424>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



**Neha Gahlan** is a Ph.D. student in the Software Engineering Department at the Delhi Technological University. She received her bachelor's and master's degrees in Computer Science Engineering from Guru Gobind Singh Indraprastha University, India, in 2019 and 2021, respectively. Her research interests include human cognition, human-machine interaction, edge computing, and signal processing.



**Divyashikha Sethia** is an Assistant professor at the Delhi Technological University, serving from 2010 till present. She has worked in the software industry in Telecom and Networking in software and test automation development in leading companies like Cisco System, 2Wire, Force10 Networks, and Future Software (acquired by Aricent) for more than 5 years in the US and India.

# Fin field-effect-transistor engineered sensor for detection of MDA-MB-231 breast cancer cells: A switching-ratio-based sensitivity analysis

Bhavya Kumar<sup>\*</sup> and Rishu Chaujar<sup>†</sup>

*Department of Applied Physics, Delhi Technological University, Delhi 110042, India*



(Received 7 February 2023; revised 23 May 2023; accepted 1 September 2023; published 19 September 2023)

The present study describes the utilization of a gallium-arsenide gate-stack gate-all-around (GaAs-GS-GAA) fin field-effect transistor (FinFET) to accomplish the electrical identification of the breast cancer cell MDA-MB-231 by monitoring the device switching ratio. The proposed sensor uses four nanocavities carved beneath the gate electrodes for enhanced detection sensitivity. MDA-MB-231 (cancerous) and MCF-10A (healthy) breast cells have a distinct dielectric constant, and it changes when exposed to microwave frequencies spanning across 200 MHz and 13.6 GHz, which modifies the electrical characteristics, allowing for early diagnosis. First, a percentage shift in the primary DC characteristics is presented to demonstrate the advantage of GS-GAA FinFET over conventional FinFET. The sensor measures the switching-ratio-based sensitivity, which comes out to be 99.72% for MDA-MB-231 and 47.78% for MCF-10A. The sensor was tested for stability and reproducibility and found to be repeatable and sufficiently stable with settling times of 55.51, 60.80, and 71.58 ps for MDA-MB-231 cells, MCF-10A cells, and air, respectively. It can distinguish between viable and nonviable cells based on electrical response alterations. The possibility of early detection of cancerous breast cells using Bruggeman's model is also discussed. Further, the impact of biomolecule occupancy and frequency variations on the device sensitivity is carried out. This study also explains how to maximize the sensing performance by adjusting the fin height, fin width, work function, channel doping, temperature, and drain voltage. Lastly, this article compared the proposed breast cancer cell detectors to existing literature to evaluate their performance and found considerable improvement. The findings of this research have the potential to establish GaAs-GS-GAA FinFET as a promising contender for MDA-MB-231 breast cancer cell detection.

DOI: [10.1103/PhysRevE.108.034408](https://doi.org/10.1103/PhysRevE.108.034408)

## I. INTRODUCTION

Cancer is not an infectious illness; instead, it is caused by a malfunction in the DNA of a cell or tissue [1]. These cells do not perform their usual functions but rather proliferate and replicate in an uncontrolled manner, resulting in the formation of a tumor. In 2020, according to WHO fact sheets, cancer was the leading cause of mortality worldwide, accounting for around 10 million deaths, or almost one in every six, with breast cancer (2.26 million cases) as the most prevalent cancer, followed by lung cancer (2.21 million cases), colon and rectum cancer (1.93 million cases), and so on [2]. The formation of malignant tumors in women's breasts is the primary cause of breast cancer, and the lifetime chance of developing it is 12%. The most common cancerous breast cells are MDA-MB-231, MCF-7, T47D, and Hs578t, while MCF-10A is a healthy nontumorigenic breast cell [1]. Compared to MCF-7 and T47D, Hs578t and MDA-MB-231 cells are regarded as the most invasive. Since invasive breast cancer cells are so dangerous and may spread rapidly, diagnosis at an early stage is very important. Early diagnosis may aid in more effective disease management, and more than 70% of cases are expected to be cured with early detection [3,4].

Cell isolation separates one or more particular cell populations from a heterogeneous mixture of cells. Targeted cells are identified, isolated, and then segregated by kind. Many cell isolation techniques are available depending on the kind of cells being separated, with a few significant ones covered in this paper, each having pros and cons. The computer-controlled micropipette (CCMP) method uses a small glass or quartz micropipette with a fine tip that a computer can control to precisely separate cells. CCMP involves manipulating a micropipette towards a suspended cell and applying a tiny suction pressure to partly aspirate the cell within the micropipette. As suction pressure rises, the cell deforms and flows into the micropipette. Researchers have widely employed this approach to explore the adhesion force measurements [5] and mechanical characteristics of diverse cells [6,7]. Fluorescence-activated cell sorting (FACS) is a flow cytometry method that uses fluorescence characteristics to separate cells. FACS begins with labeling cells with fluorescent dyes that attach to specific cell surface markers. In front of a laser, the suspended cells are passed in a stream of droplets, each containing a single cell. This stream is then directed via a series of lasers, which activate the cell-bound fluorophores, resulting in light scattering and fluorescent emissions. The fluorescence detecting system recognizes cells of interest based on the wavelengths generated by the laser excitation. Due to its wide application, FACS research includes bacteria [8], protoplasts [9], bone marrow

<sup>\*</sup>bhavyakumar\_phd2k18@dtu.ac.in

<sup>†</sup>chaujar.rishu@dtu.ac.in

cells [10], etc. Microfluidics is a cell separation technique that uses fluid manipulation on a microscopic scale. Cell isolation approaches based on microfluidics vary depending on their size, density, compressibility, electrical and magnetic characteristics, etc. The membrane filtering technique uses thin membrane layers with micropores to detect and isolate cells based on their size [11]. Cells of different densities and compressibilities may be separated using acoustic waves in a process called acoustophoresis [12]. In dielectrophoresis, nonuniform electric fields separate and isolate cells based on their dielectric properties [13]. Cells that contain magnetic nanoparticles may be identified and separated via magnetic cell sorting [14]. Laser microdissection is a high-resolution technique for isolating cells from their surrounding tissues that employs a laser beam and direct microscopic visualization. The sample is mounted on a microscope slide, and an infrared or ultraviolet laser selectively cuts off the cells of interest. After that, the sliced cells are collected for further examination. This technique has been extensively used in the research of liver illnesses [15], mass spectrometry [16], etc.

X-ray mammography, sonography, and magnetic resonance imaging (MRI) scans are some screening methods for breast cancer identification. Currently, x-ray mammography is the predominant method for breast cancer detection. Although this technology has made significant strides in this sector, there have been reports of many drawbacks [17,18]. In addition, x-ray mammography's specificity and sensitivity drastically decrease to 89% and 67% for dense breasts, and the method also includes radiation exposure dangers [19]. While sonography may be a cost-effective tool in the fight against breast cancer, the accuracy of a diagnosis relies on the experience of the person doing the procedure, and hence it may provide erroneous findings at times [20]. MRIs with improved contrast have a higher sensitivity (93–100%) [21], but they are difficult to use, costly, and limited to hospital use.

The microwave imaging technique was created to overcome the drawbacks of conventional imaging. This method uses the large dielectric difference between healthy and cancerous tissue to perform microwave imaging and heating [22–24]. Scientists have been interested in how different types of malignant cells behave and may be detected when exposed to microwave frequencies [25,26]. Kim *et al.* calculated the dielectric characteristics of fatty glandular, fibro, and malignant breast tissues from 50 MHz to 5 GHz frequency [27]. It was noted that the dispersion of malignant tissues differs from that of healthy breast tissue. In the frequency range of 50–900 MHz, Joines *et al.* investigated the dielectric characteristics of human tissues and found that cancerous tissues have greater conductivity and permittivity than normal tissues [28]. Many investigations have been conducted to know the dielectric characteristics of various human body components, such as the liver [29]. Dielectric characteristics of numerous *in vitro* breast malignant cell lines have been examined between 200 MHz and 13.6 GHz by Hussein *et al.* [1]. It was found that breast cancer cells, because of their high water content, exhibit varying dielectric characteristics, leading to enhanced scattering at microwave frequencies.

Compared to imaging methods, molecular biotechnology tests may detect breast cancer sooner. However, they cannot substitute imaging techniques but complement imaging

methods for diagnosing breast cancer. Molecular biotechnology examines biomarkers like nucleic acid, proteins, cells, and tissues of patients. Effective molecular biotechnology examination tools utilized for identifying breast cancer cells include quantitative polymerase chain reaction (qPCR), mass spectrometry (MS), single-cell resonant waveguide gratings (SCRWGs), digital holographic microscopy (DHM), etc. The qPCR, or real-time PCR, quantifies DNA and gene expression levels in samples. qPCR has been used to assess circulating tumor cells in many solid tumors, such as breast cancer [30]. The qPCR technology may guide breast cancer therapy by monitoring mRNA expression [31]. Matrix-assisted laser desorption/ionization (MALDI) mass spectrometry imaging (MSI) [32], surface-enhanced laser desorption/ionization (SELDI) MS [33], and liquid chromatography-tandem mass spectrometry (LC-MS/MS) [34] are some of the MS-based techniques utilized for breast cancer diagnosis. The concentration of adhesion proteins inside the cell-substrate contact zone causes a change in the refractive index, which may be monitored in real-time using SCRWGs. The SCRWGs technology was used to monitor the adhesion of HeLa cancer cells [35]. DHM technology provides high-resolution three-dimensional (3D) imaging of transparent biological specimens such as live cells and tissues. DHM may be utilized to capture digital holograms of breast tissues and analyze their malignancy using a deep learning approach [36]. Each technique has pros and cons, depending on the nature of the investigated molecules.

Microelectromechanical systems (MEMS) based sensors [37,38], fiber Bragg grating (FBG) [39,40], and optical sensors [41] have also played significant roles in identifying breast cancer cells. However, field-effect-transistor (FET) based devices have recently attracted attention in biosensing applications for their many benefits, including small size, low cost, high sensitivity, suitability for CMOS (complementary metal-oxide-semiconductor) technology, controllable electrical response, and reproducibility [42,43]. Tunnel field-effect transistors (TFETs) [44], high-electron-mobility transistors (HEMTs) [45], and fin field-effect transistors (FinFETs) [46] have been used in the past in breast cancer diagnosis. FinFET demonstrates superior performance compared to planar devices and optimizes the short-channel effects (SCEs) by allowing for high scalability, decreased power utilization, and longer battery life [47,48]. The gate-all-around (GAA) design provides enhanced electrostatic control over the channel, higher packing density, a steep subthreshold slope, and high current driving capability [49,50]. During MOS technology downscaling, the gate-stack (GS) design inhibits the increase in off-state leakage current ( $I_{\text{off}}$ ) [51]. Furthermore, the GS design eliminates the mobility degradation and threshold voltage instability that occur with direct deposition of high- $k$  dielectrics on silicon substrates [52,53]. Figure 1 shows a quick comparison of the conventional FinFET and GS-GAA FinFET devices concerning the percentage change in the fundamental DC performance characteristics. It is worth noting that, at a supply voltage of 0.5 V, the GS-GAA FinFET device provides an 86.26% lower off-current ( $I_{\text{off}}$ ), which leads to an improvement of 693.48% in the switching ratio (SR).

Gallium arsenide (GaAs) is considered in the fin area because it has several superior electrical properties to silicon, including high electron mobility and a large energy band gap



Parameters	Conv. FinFET	GS-GAA FinFET	Conv. FinFET	GS-GAA FinFET
	$V_{ds} = 0.05V$	$V_{ds} = 0.05V$	$V_{ds} = 0.5V$	$V_{ds} = 0.5V$
$I_{on}$ ( $\mu A$ )	2.19	2.32	21.00	22.83
$I_{off}$ (pA)	10.37	1.94	22.78	3.13
SR ( $\times 10^6$ )	0.21	1.19	0.92	7.30
SS (mV/dec)	67.52	61.49	67.05	61.86
$V_{th}$ (V)	0.27	0.29	0.25	0.27

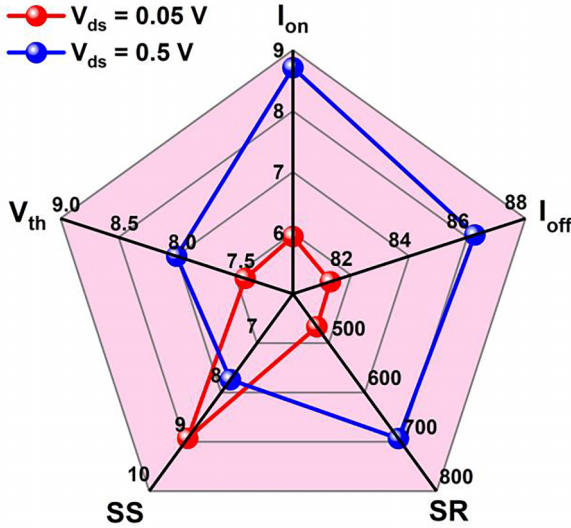


FIG. 1. Conventional FinFET vs GS-GAA FinFET concerning the percentage change in DC performance characteristics.

[54]. The absence of GaAs-based thermodynamically stable, high-quality insulators to augment device standards like SiO<sub>2</sub> on silicon is the primary difficulty with GaAs-based devices. Nonetheless, molecular beam epitaxy (MBE) and atomic layer deposition (ALD) have successfully built high-quality dielectrics atop III-V semiconductors [55–57]. Aluminum oxide (Al<sub>2</sub>O<sub>3</sub>) is favored as a gate dielectric because of its capacity to stay noncrystalline throughout manufacturing processing, excellent GaAs interface quality, huge 9 eV energy band gap, and high thermal stability [58]. Consequently, we proposed the GaAs-GS-GAA FinFET.

The proposed device employs a GAA design that encloses the gate on all four sides; consequently, four nanocavities are carved beneath the gate electrodes toward the source area for enhanced detection sensitivity. The presence or absence of breast cancer cells affects the dielectric constant of the cavity area. The change in the dielectric constant alters the device's electrical properties, which may then be utilized to identify the presence of sickness in the body. Thus, we used a GaAs-GS-GAA FinFET device in this study to identify breast cancer cells based on their dielectric constant value. MCF-10A and MDA-MB-231 breast cells were chosen for examination and may be produced by the procedure described by Hussein *et al.* [1]. Simulations were run to test the device sensitivity regarding switching ratio by analyzing the drain current characteristics for air (cell free), MCF-10A, and MDA-MB-231 cells. The efficiency of any given sensor is directly proportional to the degree to which it can recognize with a high level of accuracy or precision. Therefore, we utilized Eq. (1)

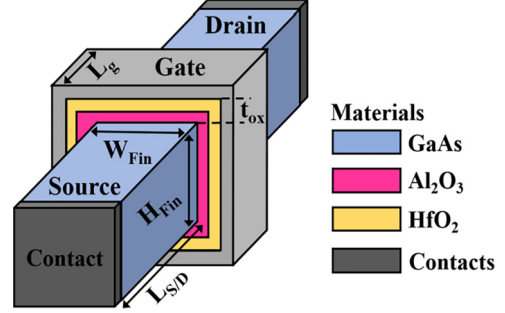


FIG. 2. Symmetric 3D view of the GaAs-GS-GAA FinFET.

to calculate the switching-ratio-based sensitivity ( $S_{SR}$ ):

$$S_{SR}(\%) = \left| \frac{SR_{(air)} - SR_{(healthy/cancerous\ cell)}}{SR_{(air)}} \right| \times 100. \quad (1)$$

Further, the healthy and malignant breast cells were taken together in various concentrations, and an investigation was carried out to identify an MDA-MB-231 infection, even in small amounts. When breast cancerous and healthy cells are combined in different concentrations, the effective dielectric constant ( $\epsilon_{eff}$ ) is determined using the formulas from Bruggeman's model [59,60]:

$$\epsilon_{eff} = \frac{H_b + \sqrt{H_b^2 + 8\epsilon_c\epsilon_h}}{4}, \quad (2)$$

with  $H_b = (2 - 3C_m)\epsilon_c - (1 - 3C_m)\epsilon_h$ .

$\epsilon_c$  and  $\epsilon_h$  are the dielectric constants of cancerous and healthy cells, and  $C_m$  represents the healthy cell fractional volume. Next, the effect of biomolecule occupancy on the device's sensitivity is explored. In biomolecule detection, the cavity area is assumed to have been completely filled. However, during biomolecule immobilization, the target biomolecule only fills a portion of the cavity areas, leaving some empty space, which can change the proposed device's electrical performance for different target biomolecules. Thus, there is a need to consider the biomolecule occupancy factor ( $\gamma_{Bio}$ ) as it can potentially affect the sensitivity of the sensor.  $\gamma_{Bio}$  is defined as follows

TABLE I. Values of different parameters used for simulation.

Parameters	Symbol	Value	Unit
Source/drain length	$L_{S/D}$	50	nm
Gate length	$L_g$	50	nm
Cavity length	$C_L$	25	nm
Cavity height	$C_H$	3	nm
Oxide thickness	$t_{ox}$	3	nm
Fin height	$H_{Fin}$	30	nm
Fin width	$W_{Fin}$	15	nm
Gate thickness	$G_t$	5	nm
Channel doping	$N_{Ch}$	$1 \times 10^{16}$	cm <sup>-3</sup>
Source/drain doping	$N_{S/D}$	$5 \times 10^{18}$	cm <sup>-3</sup>
Work function	$\phi_m$	4.65	eV
Temperature	$T$	300	K
Gate-source voltage	$V_{gs}$	1.5	V
Drain-source voltage	$V_{ds}$	0.5	V

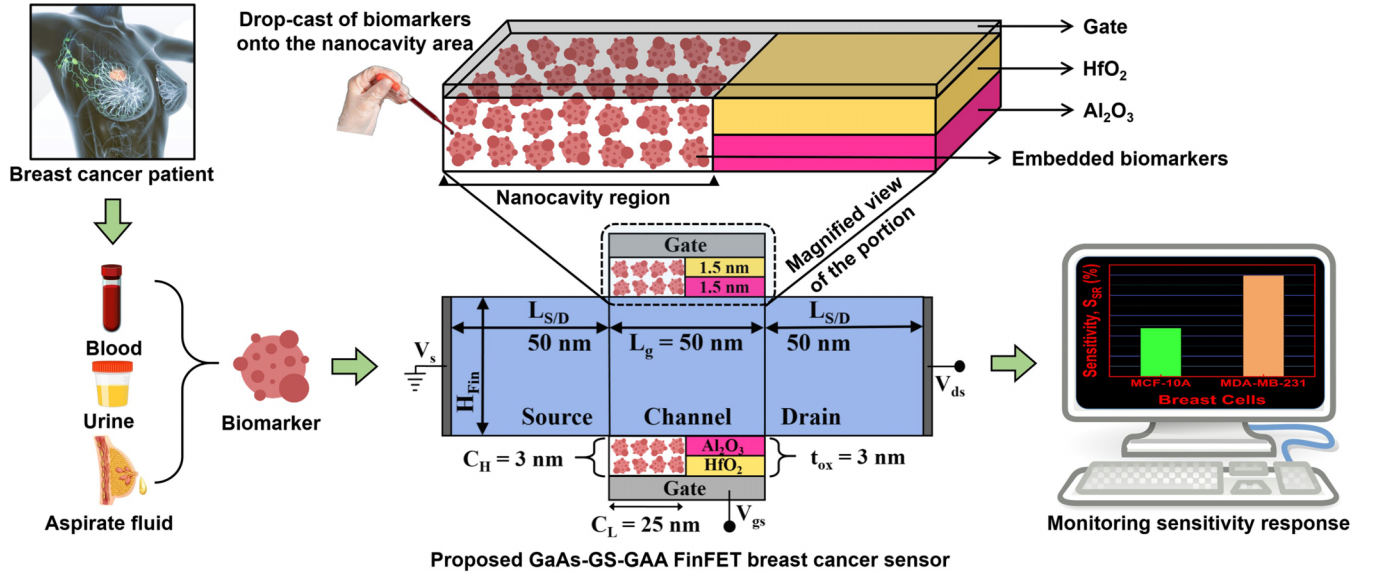


FIG. 3. Diagrammatic representation of the operation of a GaAs-GS-GAA FinFET sensor for breast cancer cell recognition.

[44]:

$$\gamma_{\text{Bio}}(\%) = \frac{\text{Thickness of cavity filled}}{\text{Total thickness of the cavity}} \times 100. \quad (3)$$

The proposed study also investigates how changes to the frequency and device's physical parameters, like fin height, fin width, work function, channel doping, temperature, and drain voltage, influence the device's sensitivity. Based on the findings, optimal device settings for maximizing sensitivity may be selected. Finally, the effectiveness of the proposed breast cancer cell detector is compared to that of already existing breast cancer detectors.

## II. DEVICE ARCHITECTURE AND SIMULATION APPROACH

Figure 2 illustrates the symmetric 3D view of the GaAs-GS-GAA FinFET. Table I contains detailed descriptions of the device's structural parameters. The fin area is made of GaAs material. The simulations adhere to the width quantization property by keeping fin width ( $W_{\text{Fin}}$ ) at a constant proportional multiple of fin height ( $H_{\text{Fin}}$ ) [61,62]. In FinFET devices, it is recommended that the  $W_{\text{Fin}}$  should be less than one-third of the gate length ( $L_g$ ) and  $H_{\text{Fin}}$  should be in the  $0.6L_g$  to  $0.8L_g$  range to minimize the SCEs [63,64]. We considered that recommendation during the device dimension consideration. The gate oxide has a combination of coatings of  $\text{Al}_2\text{O}_3$  and  $\text{HfO}_2$ . All sections are uniformly n-type doped with lower channel doping than the source/drain doping to lessen the parasitic capacitance. At 200 MHz frequency, the dielectric constants ( $k$ ) for MCF-10A and MDA-MB-231 are 4.33 and 24.50, while at 13.6 GHz, they drop to 2.76 and 16.65, respectively [1]. For air, which does not have any cells,  $k$  is 1. Figure 3 presents a diagrammatic representation of the operation of a GaAs-GS-GAA FinFET sensor for breast cancer cell recognition. The biomarker was drop-cast onto the nanocavity area carved beneath the gate electrodes to analyze the desired parameters using a technique based on dielectric modulation.

The GaAs-GS-GAA FinFET structure was simulated using the SILVACO-Atlas 3D simulator [65]. The Poisson and continuity equations are frequently used in the device simulation, but additional equations and models are also needed to improve device simulation results. As a result, the simulations include a wide variety of physical models. Quantum confinement effects are an essential design consideration for rapidly scalable devices. To consider the consequences of quantum confinement, the Bohm quantum potential (BQP) model uses a position-dependent quantum potential ( $Q$ ) with parameters  $\gamma = 1.4$  and  $\alpha = 0.3$  [66]. Fermi-Dirac statistics, Crowell-Size impact ionization, concentration-dependent mobility, Klaassen tunneling, Shockley-Read-Hall (SRH) recombination, and band gap narrowing are the other

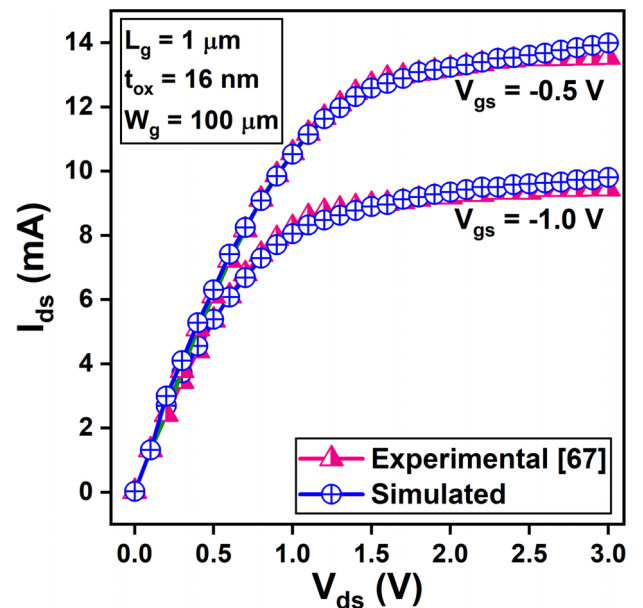


FIG. 4. Calibration curve of an  $\text{Al}_2\text{O}_3/\text{GaAs}$  MOSFET.

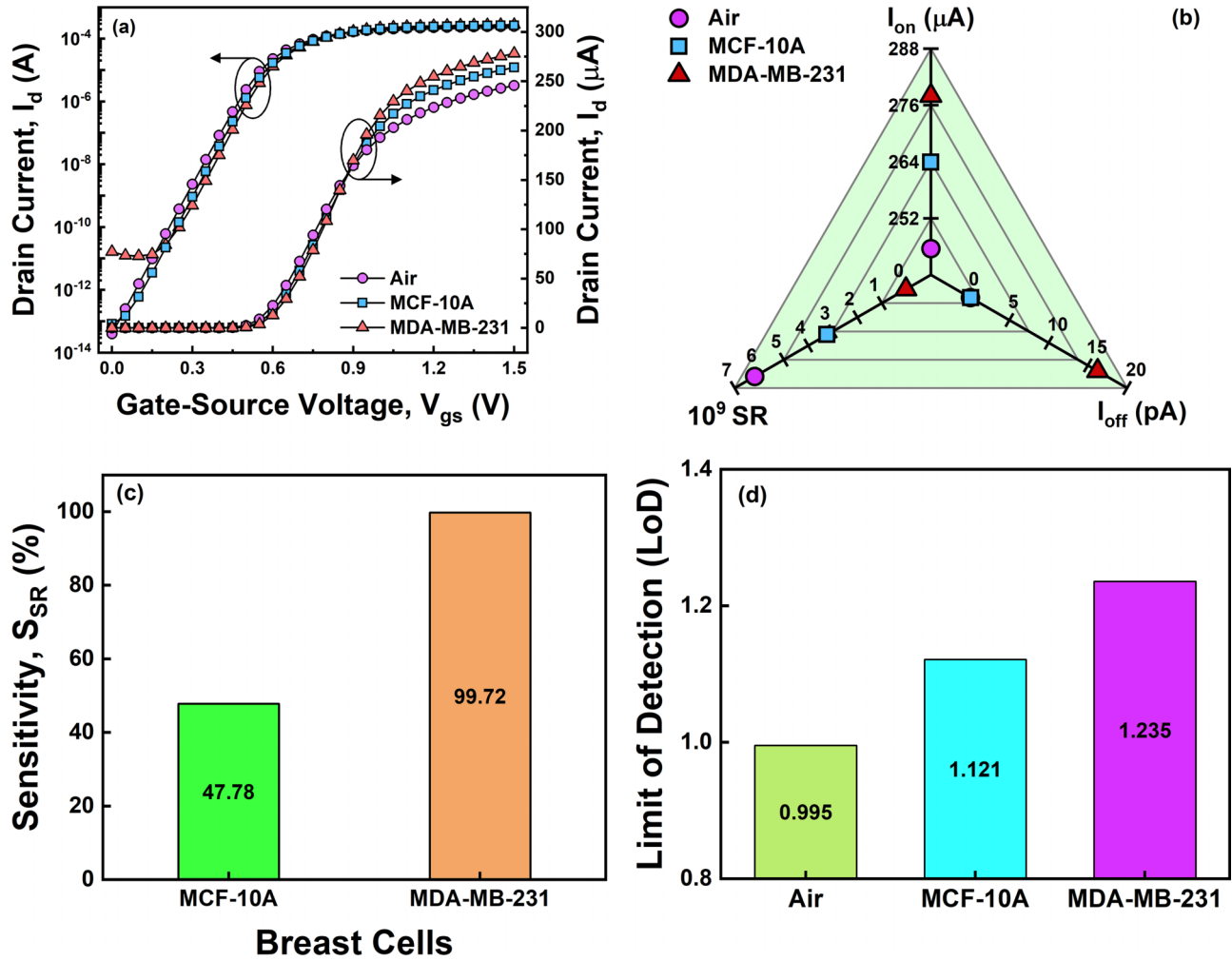


FIG. 5. (a) Transfer characteristics of the proposed sensor for air, MCF-10A, and MDA-MB-231 in linear and logarithmic form. (b) Fluctuation in  $I_{on}$ ,  $I_{off}$ , and SR for air, MCF-10A, and MDA-MB-231. (c)  $S_{SR}$  comparison of MDA-MB-231 and MCF-10A cells. (d) LoD plot for air, MCF-10A, and MDA-MB-231.

standard models that have been incorporated [65]. Additionally, drain current characteristics are simulated using Newton and Block iteration methods for breast cancer cell identification.

We used the findings published by Ye *et al.* [67] to verify the simulation models discussed before. The output characteristics of an  $\text{Al}_2\text{O}_3/\text{GaAs}$  MOSFET operated with  $V_{ds} = -0.5$  V and  $V_{gs} = -1.0$  V are revealed in Fig. 4, along with the experimental and simulated results. The fact that the data sets obtained via simulation and experiment are comparable adds credibility to the simulation models chosen. The steps in creating the proposed GaAs-GS-GAA FinFET device were thoroughly covered in our earlier article [68], including a flowchart illustrating the whole process. Moreover, the cavity region can be created by dry etching the gate dielectric toward the source area.

### III. RESULTS AND DISCUSSION

#### A. Switching-ratio-based sensitivity analysis

Figure 5(a) depicts the transfer characteristics ( $I_d - V_{gs}$ ) of the proposed sensor for the air, MCF-10A, and MDA-MB-231

in linear and logarithmic forms. The sensor is examined at  $V_{ds} = 0.5$  V field bias conditions. The drain current increases with the gate-source voltage ( $V_{gs}$ ) and attains maximum value for MDA-MB-231. In contrast, the opposite trend is observed in the leakage current and degrades significantly for the MDA-MB-231 cancer cell. Figure 5(b) provides a clearer picture of the fluctuation in on-current ( $I_{on}$ ), off-current ( $I_{off}$ ), and switching ratio ( $SR = I_{on}/I_{off}$ ) (which is subsequently employed as a sensitivity parameter). It can be seen that  $I_{on}$  is higher for the MDA-MB-231 cancer cell than for air and healthy cells. It is because introducing the breast cancer cells in the cavity region leads to enhanced effective gate oxide, which in turn increases the coupling between the channel region and gate metal, and thereby on-current. The difference in  $I_{off}$  between air and MCF-10A is not very large, but for MDA-MB-231 it is noticeably greater, which brings the SR down to a rather remarkable level. Figure 5(c) compares MDA-MB-231 and MCF-10A cells in terms of their  $S_{SR}$ . The graph reveals that the  $S_{SR}$  for MDA-MB-231 is 99.72%, and it is 47.78% for MCF-10A. The presence of MDA-MB-231 cells causes a more pronounced shift in  $I_{off}$ , which ultimately enhances the device's sensitivity. Figure 5(d) demonstrates the



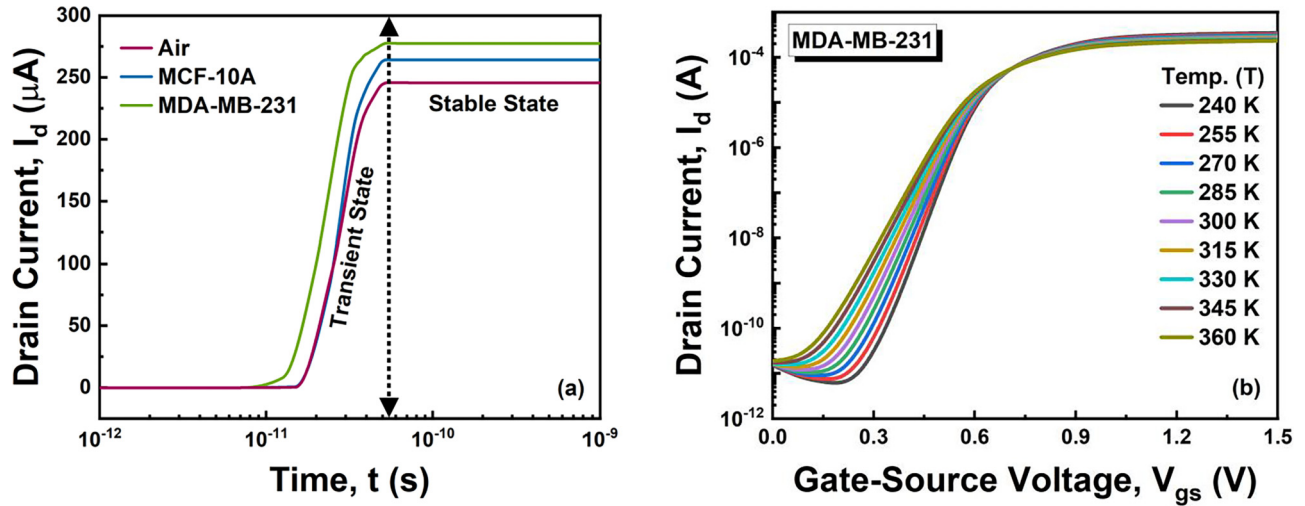


FIG. 6. (a) Transient response of the drain current for air and for MCF-10A and MDA-MB-231 cells. (b) Temperature dependence of the  $I_d$ – $V_{gs}$  characteristics of the MDA-MB-231 cancer cell.

limit of detection (LoD) plot for the three samples considered. The lowest concentration of an analyte that can be accurately identified from a sample with a high degree of certainty is known as the LoD [69]. LoD is calculated using the response of slope and standard deviation (SD) of the intercept. The slope and SD of the intercept for MDA-MB-231, MCF-10A, and air are analyzed using the transfer characteristics curve [Fig. 5(a)]. The slope is 0.0002 for all samples, and the SD values come out to be 0.0000749, 0.0000679, and 0.0000603 for MDA-MB-231, MCF-10A, and air. As a result, the LoD obtained for MDA-MB-231 is slightly higher than those for air and MCF-10A. The relevance of reducing execution variability between devices and enhancing sensitivity while designing and producing nano-FET biosensors is provided by these results.

### B. Stability and reproducibility analysis

The transient analysis of air and of MCF-10A and MDA-MB-231 cell lines was carried out to test the stability of the proposed sensor. The time it takes for the drain current to settle from its transient state to its steady state is called the settling time ( $t_{\text{sett}}$ ) [70,71]. The transient response is simulated by applying  $V_{gs}$  with an amplitude of 1.5 V, a ramp time of  $5 \times 10^{-11}$  s, a stop time of  $1 \times 10^{-9}$  s, and a step time of  $1 \times 10^{-12}$  s. Figure 6(a) shows the transient response of the drain current for air and for MCF-10A and MDA-MB-231 cells. It was observed that the drain current is higher for the MDA-MB-231 cell compared to MCF-10A and air, due to which  $t_{\text{sett}}$  for the MDA-MB-231 cell is somewhat lower (55.51 ps) than  $t_{\text{sett}}$  for the MCF-10A cell (60.80 ps) and air (71.58 ps). After  $t_{\text{sett}}$ , the current becomes steady at a magnitude equal to that at  $V_{gs} = 1.5$  V [Fig. 5(a)]. The effect of temperature on the transfer characteristics is investigated to further probe the stability of our proposed sensor. Figure 6(b) displays the temperature dependence of the  $I_d$ – $V_{gs}$  characteristics of the MDA-MB-231 cancer cell; it can be observed that the  $I_d$ – $V_{gs}$  curves do not vary significantly between 240 and 360 K.

Second, to examine reproducibility, it is necessary to test the sensor's repeatability under controlled use settings. As shown in Fig. 7, we conducted the transient simulation of the proposed sensor for MDA-MB-231 cancerous cells over four cycles with a gap of 30 minutes between each cycle. The drain current is measured for four cycles, and the findings reveal that the drain current can be reliably reproduced with unnoticeable deviation. The above data suggest that the GaAs-GS-GAA FinFET is sufficiently stable and reproducible.

### C. Cell viability

The term “oxidative stress” pertains to a state of imbalance between the generation of reactive oxygen species and the capacity of cells to counteract the consequent harm. Cells undergo either apoptotic or necrotic cell death when their antioxidant defense mechanisms are overwhelmed by high amounts of oxidative stress. Zou *et al.* employed a silicon-based attenuated total reflectance terahertz time-domain spectroscopy (ATR THz-TDS) system to monitor cell mortality caused by oxidative stress in MCF-10A breast

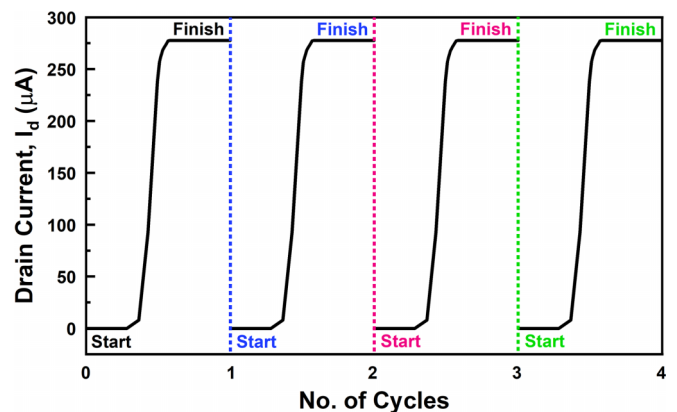


FIG. 7. Transient response of the proposed sensor for MDA-MB-231 cancerous cells over four cycles.

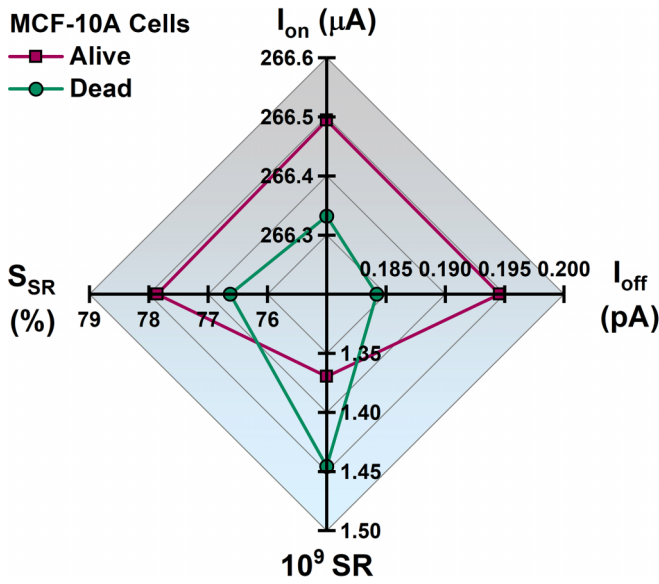


FIG. 8. Variation in  $I_{on}$ ,  $I_{off}$ , SR, and  $S_{SR}$  of living and dead MCF-10A breast cells.

cells [72]. This study shows the THz dielectric responses of living and dead MCF-10A breast cells, in which cell death is induced by oxidative stress using a high concentration of hydrogen peroxide (10 mM,  $H_2O_2$ ). Thus, with the assistance of this study, we have extracted the dielectric constants at 0.3 THz frequency for MCF-10A breast cells before and after the oxidative stress to monitor the electrical response of the proposed sensor on the live and dead cells. The sample under consideration (MCF-10A) may be produced by the procedure described by Zou *et al.* Unfortunately, to the author's best knowledge, no analysis has been performed to characterize the dielectric responses of dead MDA-MB-231 cancerous breast cells. As a result, the electrical response of the proposed sensor on the dead MDA-MB-231 cancerous breast cells could not be determined. The spider-chart depiction of the variation in  $I_{on}$ ,  $I_{off}$ , SR, and  $S_{SR}$  of living and dead MCF-10A breast cells is shown in Fig. 8. The induction of cell death through oxidative stress reduces  $I_{on}$  from 266.49 to 266.33  $\mu A$  and  $I_{off}$  from 0.195 to 0.184 pA. Since the reduction in  $I_{off}$  is bigger than the drop in  $I_{on}$ , SR is increased by 5.55%, and  $S_{SR}$  is lowered from 77.86% to 76.62%. Despite the relatively minor changes in the electrical response, the sensor under consideration may differentiate between viable and nonviable cells.

#### D. Early detection

Figures 9(a) and 9(b) demonstrate the transfer characteristics for five different combinations in linear and logarithmic form. In the graph, HC represents the healthy cell and CC is the cancerous cell. The presence of 90% HC and 10% CC indicates a very low quantity of cancerous breast cells, while the presence of 10% HC and 90% CC indicates a very high concentration of breast cancer cells. An enlarged view of the peaks generated by mixing various numbers of healthy and malignant cells is also shown in the insets of Figs. 9(a) and 9(b). It is visible that the drain current and the leakage

current increase with the rise in the concentration of MDA-MB-231 cancerous cells. Figure 9(c) exhibits the spider-chart representation of the variance in  $I_{on}$ ,  $I_{off}$ , and SR over the five different combinations.  $I_{on}$  increases from 265 to 277  $\mu A$ ,  $I_{off}$  increases by  $\sim 10^2$  orders, and SR decreases by 98.48% when the concentration of cancerous cells is raised from 10% to 90%. Thus, the developed sensor can detect the presence of breast cancer cells, even at low concentrations, allowing for early illness diagnosis.

#### E. Effect of biomolecule occupancy on sensitivity

In order to examine the biomolecule occupancy of the device, five different sites were analyzed: 20%, 40%, 60%, 80%, and 100%. A section of the cavity is covered with biomolecules, while the remaining space is filled with air or left vacant to study the effect of biomolecule occupancy on sensor sensitivity. Figure 10(a) shows the  $I_{on}$ ,  $I_{off}$ , and SR for a healthy MCF-10A cell, while Fig. 10(b) shows the same data for a malignant MDA-MB-231 cell for the five considered occupancy combinations of biomolecules. The increase in the  $\gamma_{Bio}$  leads to an increase in the effective dielectric and capacitance in the cavity area, ultimately increasing  $I_{on}$ . For MCF-10A,  $I_{on}$  is 256  $\mu A$  at 20% biomolecule occupancy, which rises to 264  $\mu A$  for 100% occupancy. Similarly,  $I_{on}$  is 267  $\mu A$  at 20%, which rises to 278  $\mu A$  at 100% occupancy for MDA-MB-231.  $I_{off}$  and SR show the same trend and improve with the increase in  $\gamma_{Bio}$  for MDA-MB-231 and MCF-10A cells. The sensitivity performance for healthy and malignant cells is plotted against different  $\gamma_{Bio}$  in Fig. 10(c). The sensitivity  $S_{SR}$  for MDA-MB-231 and MCF-10A decreases slightly with the increase in  $\gamma_{Bio}$ .

#### F. Effect of device parametric variation on sensitivity

Figures 11(a) to 11(f) collectively demonstrate the  $S_{SR}$  of the proposed sensor against the deviation of the mentioned parameters for the MDA-MB-231 cancerous cell. The fin height ( $H_{Fin}$ ) varies from 30 to 40 nm with a step size of 2.5 nm. The  $S_{SR}$  of the proposed device increases with the surge in  $H_{Fin}$ , as shown in Fig. 11(a). Fin width ( $W_{Fin}$ ) is altered from 5 to 15 nm with a step size of 2.5 nm. Figure 11(b) displays the sensitivity  $S_{SR}$  as a function of  $W_{Fin}$  and shows an improvement in  $S_{SR}$  with the rise in  $W_{Fin}$ , thus following the path of  $H_{Fin}$ . The work function ( $\phi_m$ ) is varied from 4.55 to 4.75 eV with an increase of 0.5 eV. The  $S_{SR}$  is 85.17% for 4.55 eV, which rises to 99.99% for 4.75 eV, as depicted in Fig. 11(c). Next is channel doping ( $N_{Ch}$ ), which is considered from  $1 \times 10^{16} \text{ cm}^{-3}$  to  $1 \times 10^{18} \text{ cm}^{-3}$ . Figure 11(d) exhibits the variation of  $S_{SR}$  with  $N_{Ch}$  and shows that the increase in the  $N_{Ch}$  results in sensitivity degradation. The temperature ( $T$ ) range is from 200 to 400 K, with a measurement taken for every 50 K, as portrayed in Fig. 11(e). The reduction in the  $S_{SR}$  is marginal from 200 to 300 K, but afterward  $S_{SR}$  decreases significantly with the reduction in  $T$ . Lastly, in Fig. 11(f), the sensitivity is plotted against the drain voltage ( $V_{ds}$ ), which varies from 0.1 to 0.5 V with a step size of 0.1 V. The  $S_{SR}$  is relatively lower at 0.1 V but increases after that with the increase in  $V_{ds}$ , with the highest value being recorded at  $V_{ds} = 0.5$  V. Thus, to summarize, the increased levels of

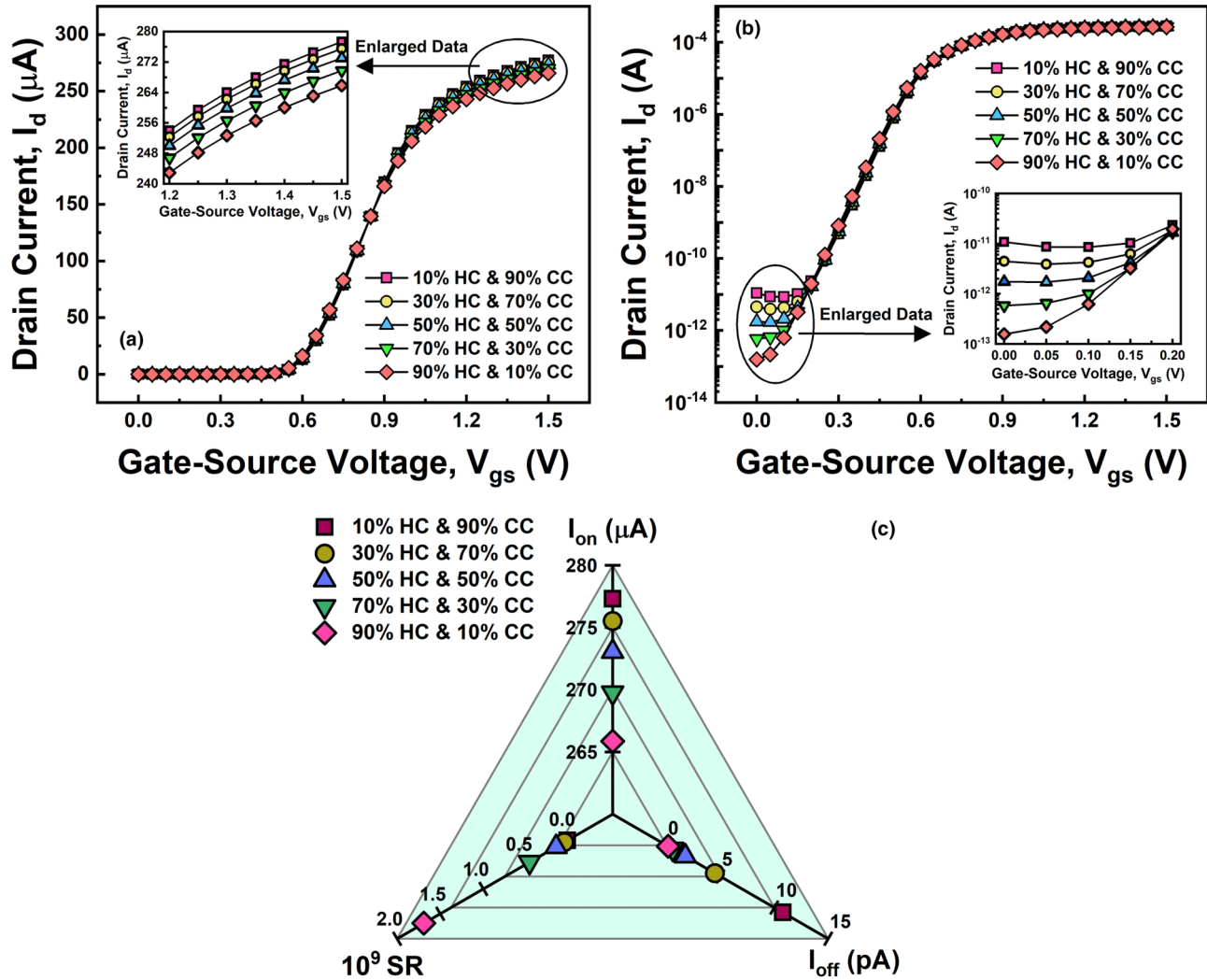


FIG. 9. Transfer characteristics for five different combinations in (a) linear and (b) logarithmic form. (c) Variation in  $I_{on}$ ,  $I_{off}$ , and SR for the combinations considered.

$H_{Fin}$ ,  $W_{Fin}$ ,  $\phi_m$ , and  $V_{ds}$ , and the decreased levels of  $N_{Ch}$  and  $T$ , make it simpler to identify the breast cancer cells.

### G. Effect of frequency on sensitivity

We evaluated four parameters, namely  $I_{on}$ ,  $I_{off}$ , SR, and  $S_{SR}$  of MDA-MB-231 and MCF-10A breast cells at 13.6 GHz, to study the effect of frequency on these parameters. The comparative statistics for 200 MHz and 13.6 GHz in tabular form are shown in Fig. 12, along with a plot of the percentage change in each performance parameter mentioned above for MDA-MB-231 and MCF-10A breast cells. We evaluated the percentage change considering 200 MHz as the initial value and 13.6 GHz as the final value and plotted the actual percentage change to understand better the impact of the frequency on the respective parameter. When the frequency is raised from 200 MHz to 13.6 GHz,  $I_{on}$  decreases by 1.93% and  $S_{SR}$  by 2.85%, and  $I_{off}$  and SR improve by 3.70% and 2.48%, respectively, for MCF-10A. Similarly, for MDA-MB-231,  $I_{on}$  decreases by 0.86% and  $S_{SR}$  by 0.69%, and  $I_{off}$  and SR improve by 71.57% and 247.06%, respectively, with the rise in

frequency. Thus, the proposed sensor detection sensitivity is significantly better at 200 MHz compared to 13.6 GHz for MDA-MB-231 and MCF-10A.

### H. Comparison with published breast cancer detectors

An evaluation of the proposed breast cancer sensor against existing breast cancer detectors is required to determine its efficacy. Table II gives an overview of the proposed FinFET breast cancer sensor with other already published breast cancer sensors regarding the change in drain current ( $\Delta I_{ds}$ ) and drain current sensitivity ( $S_{Id}$ ). The  $\Delta I_{ds}$  data were unavailable for the reduced graphene oxide (rGO) encapsulated nanoparticle (NP) based FET biosensor, so we mentioned the device's sensitivity, which is about 3.9%. The highest reported  $\Delta I_{ds}$  was 6 μA for the AlGaN/GaN HEMT structure, with  $\Delta I_{ds}$  for the remaining devices being relatively low. The proposed GaAs-GS-GAA FinFET sensor exhibits improved results compared to the breast cancer detectors mentioned in Table II, with  $\Delta I_{ds}$  of about 32.5 μA and  $S_{Id}$  of 13.21%.

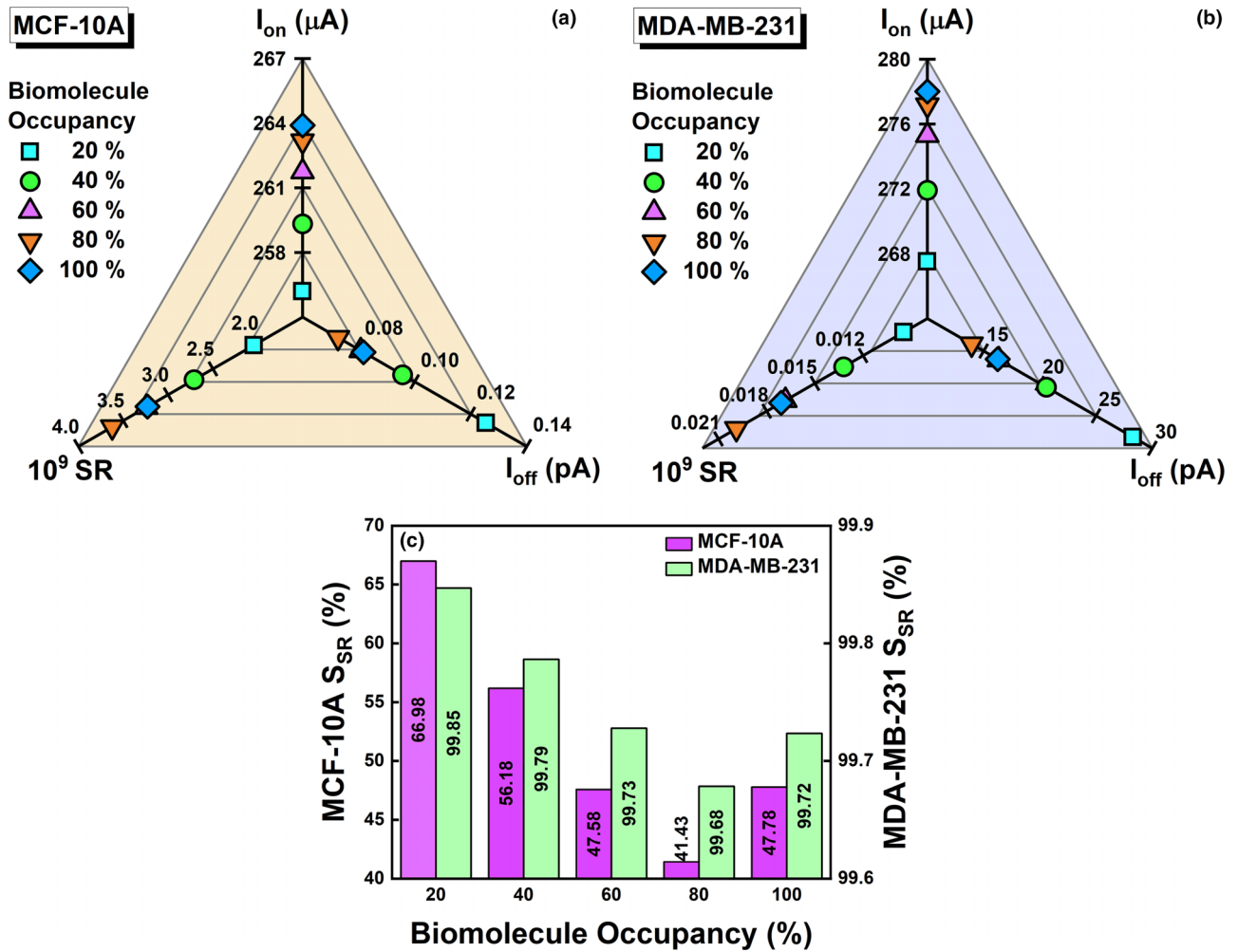


FIG. 10. Biomolecule occupancy impact on  $I_{on}$ ,  $I_{off}$ , and SR for (a) MCF-10A and (b) MDA-MB-231. (c) Sensitivity performance for healthy and malignant cells against different  $\gamma_{Bio}$ .

#### IV. CONCLUSION

The current work details the usage of a GaAs-GS-GAA FinFET to monitor the device switching ratio to achieve the electrical identification of the MDA-MB-231 breast cancer cell. In order to increase the detection sensitivity, the suggested sensor makes use of four nanocavities that are carved

underneath the gate electrodes. To emphasize the advantages of GS-GAA FinFET over traditional FinFET, a percentage change in the crucial electrical parameters is displayed for both types of FinFET. The switching-ratio-based sensitivity of the sensor is measured for healthy and malignant breast cells and turns out to be 47.78% and 99.72%, respectively. The sensor was evaluated for its reproducibility and stability

TABLE II. Overview of proposed FinFET-engineered cancer detector vs other published cancer detectors.

References	Year	Platform device	Detection	Change in drain current, $\Delta I_{ds}$ ( $\mu A$ )	Drain current sensitivity, $S_{Id}$ (%)
[45]	2009	AlGaIn/GaN HEMT	c-erbB-2, a breast cancer marker	6	—
[73]	2011	rGO encapsulated NP-based FET	HER2 and EGFR, a breast cancer marker	—	3.9
[74]	2020	Apta-cyto-sensor	MDA-MB-231 breast cancer cells	3	—
[75]	2021	CNT FET biosensor	Breast cancer exosomal miRNA21	1.65	—
[44]	2022	DL-NC-FE-TFET	T47D, Hs578T, MDA-MB-231, and MCF-7 breast cancer lines	1.83	—
This work		GaAs-GS-GAA FinFET	MDA-MB-231 breast cancer cells	32.5	13.21



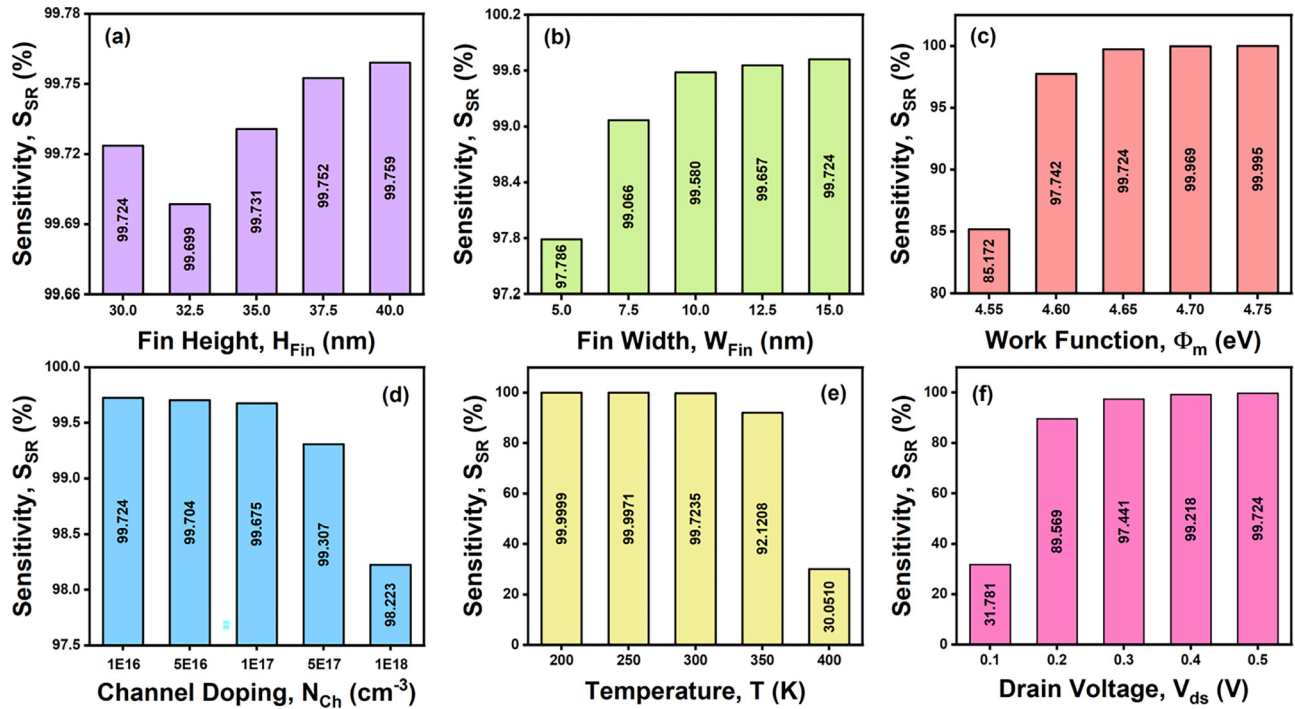


FIG. 11.  $S_{SR}$  of the proposed sensor against the deviation of mentioned parameters for the MDA-MB-231 cancerous cell.

and was found to be repeatable and adequately stable with settling times of 55.51 ps for the MDA-MB-231 cell, 60.80 ps for the MCF-10A cell, and 71.58 ps for air. Furthermore, the sensor is capable of distinguishing between viable and nonviable cells based on changes in their electrical response. The research also shows that breast cancer cells can be identified with the assistance of Bruggeman's model even when

present in a mixed solution of malignant and healthy cells, even though the quantity of cancerous cells is lower. The effect of biomolecule occupancy and frequency fluctuations on the device's sensitivity is also investigated. This research also describes how to enhance the sensing performance by altering the fin height, fin width, work function, channel doping, temperature, and drain voltage. The proposed sensor can better identify malignant cells when the levels of  $H_{Fin}$ ,  $W_{Fin}$ ,  $\phi_m$ , and  $V_{ds}$  increase and the  $N_{Ch}$  and  $T$  levels decrease. Finally, this work compared the suggested GaAs-GS-GAA FinFET sensor to previously published breast cancer sensors regarding the change in drain current and drain current sensitivity and found that the proposed sensor performed much better. Thus, the proposed GaAs-GS-GAA FinFET sensor may be considered an intriguing candidate for MDA-MB-231 breast cancer cell detection.

Parameters	MCF-10A		MDA-MB-231	
	200 MHz	13.6 GHz	200 MHz	13.6 GHz
$I_{on}$ ( $\mu A$ )	263.90	258.80	278.00	275.60
$I_{off}$ (pA)	0.081	0.078	16.25	4.62
SR ( $\times 10^9$ )	3.23	3.31	0.017	0.059
$S_{SR}$ (%)	47.78	46.42	99.72	99.03

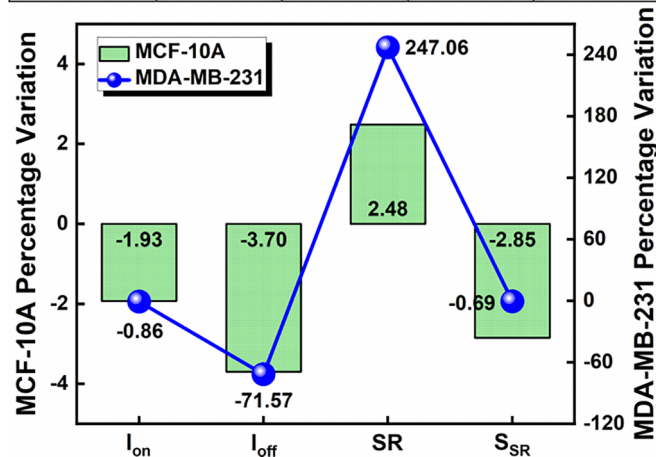


FIG. 12. Percentage change in each mentioned performance parameter for MDA-MB-231 and MCF-10A breast cells.

## ACKNOWLEDGMENTS

The authors express their gratitude to the Microelectronics Research Lab at DTU for providing all of the essential resources for this research. This research received no specific grant from the public, commercial, or not-for-profit funding agencies.

B.K. conceptualized the work and was responsible for methodology, software, analysis, data curation, and original draft preparation. R.C. conceptualized the work and was responsible for analysis, data curation, review and editing of the paper at different stages, and supervision.

The authors declare that they have no known conflict of interest or personal relationships that could have appeared to influence the work reported in this paper.

- [1] M. Hussein, F. Awwad, D. Jithin, H. El Hasasna, K. Athamneh, and R. Iratni, Breast cancer cells exhibits specific dielectric signature *in vitro* using the open-ended coaxial probe technique from 200 MHz to 13.6 GHz, *Sci. Rep.* **9**, 4681 (2019).
- [2] World Health Organization, Cancer, accessed Feb. 03, 2022, <https://www.who.int/news-room/fact-sheets/detail/cancer>.
- [3] Early Breast Cancer Trialists' Collaborative Group, Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: An overview of the randomized trials, *Lancet* **365**, 1687 (2005).
- [4] Cancer Research UK, Breast Cancer Incidence (Invasive) Statistics, 2015 (unpublished).
- [5] T. Gerecsei, I. Erdődi, B. Peter, C. Hős, S. Kurunczi, I. Derényi, B. Szabó, and R. Horvath, Adhesion force measurements on functionalized microbeads: An in-depth comparison of computer controlled micropipette and fluidic force microscopy, *J. Colloid Interface Sci.* **555**, 245 (2019).
- [6] R. P. Rand and A. C. Burton, Mechanical properties of the red cell membrane: I. Membrane stiffness and intracellular pressure, *Biophys. J.* **4**, 115 (1964).
- [7] F. Guilak, W. R. Jones, H. P. Ting-Beall, and G. M. Lee, The deformation behavior and mechanical properties of chondrocytes in articular cartilage, *Osteoarthr. Cartil.* **7**, 59 (1999).
- [8] M. Schmidt, M. K. Hourfar, S.-B. Nicol, H.-P. Spengler, T. Montag, and E. Seifried, FACS technology used in a new rapid bacterial detection method, *Transfus. Med.* **16**, 355 (2006).
- [9] I. Antoniadis, V. Skalický, G. Sun, W. Ma, D. W. Galbraith, O. Novák, and K. Ljung, Fluorescence activated cell sorting-A selective tool for plant cell isolation and analysis, *Cytom. Part A* **101**, 725 (2022).
- [10] X. Liao, M. Makris, and X. M. Luo, Fluorescence-activated cell sorting for purification of plasmacytoid dendritic cells from the mouse bone marrow, *J. Vis. Exp.* **117**, 54641 (2016).
- [11] D. L. Adams, P. Zhu, O. V. Makarova, S. S. Martin, M. Charpentier, S. Chumsri, S. Li, P. Amstutz, and C. M. Tang, The systematic study of circulating tumor cell isolation using lithographic microfilters, *RSC Adv.*, **4**, 4334 (2014).
- [12] P. Li, Z. Mao, Z. Peng, L. Zhou, Y. Chen, P. H. Huang, C. I. Truica, J. J. Drabick, W. S. El-Deiry, M. Dao, S. Suresh, and T. J. Huang, Acoustic separation of circulating tumor cells, *Proc. Natl. Acad. Sci. USA* **112**, 4970 (2015).
- [13] H. Song, J. M. Rosano, Y. Wang, C. J. Garson, B. Prabhakarpandian, and K. Pant, Continuous-flow sorting of stem cells and differentiation products based on dielectrophoresis, *Lab Chip* **15**, 1320 (2015).
- [14] Y. Zhou, Y. Wang, and Q. Lin, A microfluidic device for continuous-flow magnetically controlled capture and isolation of microparticles, *J. Microelectromech. Syst.* **19**, 743 (2010).
- [15] B. Aguilar-Bravo and P. Sancho-Bru, Laser capture microdissection: Techniques and applications in liver diseases, *Hepatol. Int.* **13**, 138 (2019).
- [16] J. A. Herrera, V. Mallikarjun, S. Rosini, M. A. Montero, C. Lawless, S. Warwood, R. O'Cualain, D. Knight, M. A. Schwartz, and J. Swift, Laser capture microdissection coupled mass spectrometry (LCM-MS) for spatially resolved analysis of formalin-fixed and stained human lung tissues, *Clin. Proteomics* **17**, 24 (2020).
- [17] J. G. Elmore, M. B. Barton, V. M. Moceris, S. Polk, P. J. Arena, and S. W. Fletcher, Ten-year risk of false positive screening mammograms and clinical breast examinations, *New England J. Med.* **338**, 1089 (1998).
- [18] Cancer Facts and Figures. [Online] American Cancer Society. (2007), available: <https://www.cancer.org>.
- [19] P. Skaane, S. Hofvind, and A. Skjennald, Randomized trial of screen-film versus full-field digital mammography with soft-copy reading in population-based screening program: Follow-up and final results of Oslo II study, *Radiology* **244**, 708 (2007).
- [20] J.-L. Gonzalez-Hernandez, A. N. Recinella, S. G. Kandlikar, D. Dabydeen, L. Medeiros, and P. Phatak, Technology, application and potential of dynamic breast thermography for the detection of breast cancer, *Int. J. Heat Mass Transf.* **131**, 558 (2019).
- [21] S. J. Lord, W. Lei, P. Craft, J. N. Cawson, I. Morris, S. Walleiser, A. Griffiths, S. Parker, and N. Houssami, A systematic review of the effectiveness of magnetic resonance imaging (MRI) as an addition to mammography and ultrasound in screening young women at high risk of breast cancer, *Eur. J. Cancer* **43**, 1905 (2007).
- [22] M. Säbel and H. Aichinger, Recent developments in breast imaging, *Phys. Med. Biol.* **41**, 315 (1996).
- [23] E. C. Fear and M. A. Stuchly, Microwave detection of breast cancer, *IEEE Trans. Microw. Theory Techn.* **48**, 1854 (2000).
- [24] S. C. Hagness, A. Taflov, and J. E. Bridges, Three-dimensional FDTD analysis of a pulsed microwave confocal system for breast cancer detection: Design of an antenna-array element, *IEEE Trans. Antennas Propag.* **47**, 783 (1999).
- [25] X. Li, E. J. Bond, B. D. V. Veen, and S. C. Hagness, An overview of ultra-wideband microwave imaging via space-time beam-forming for early-stage breast-cancer detection, *IEEE Antennas Propag. Mag.* **47**, 19 (2005).
- [26] S. Kwon and S. Lee, Recent advances in microwave imaging for breast cancer detection, *Int. J. Biomed. Imag.* **2016**, 5054912 (2016).
- [27] T. Kim, J. Oh, B. Kim, J. Lee, S. Jeon, and J. Pack, A study of dielectric properties of fatty, malignant and fibro-glandular tissues in female human breast, in *Proceedings of the Asia-Pacific and 19th International Zurich Symposium on Electromagnetic Compatibility, Singapore, 2008* (IEEE, Piscataway, NJ, 2008), pp. 216–219.
- [28] W. T. Joines, Y. Zhang, C. Li, and R. L. Jirtle, The measured electrical properties of normal and malignant human tissues from 50 to 900 MHz, *Med. Phys.* **21**, 547 (1994).
- [29] L. Chin and M. Sherar, Changes in dielectric properties of *ex vivo* bovine liver at 915 MHz during heating, *Phys. Med. Biol.* **46**, 197 (2001).
- [30] U. Andergassen, M. Zebisch, A. C. Kölbl, A. König, S. Heublein, L. Schröder, S. Hutter, K. Friese, and U. Jeschke, Real-time qPCR-based detection of circulating tumor cells from blood samples of adjuvant breast cancer patients: A preliminary study, *Breast Care* **11**, 194 (2016).
- [31] T. Y. Ryu, K. Kim, S.-K. Kim, J.-H. Oh, J.-K. Min, C.-R. Jung, M.-Y. Son, D.-S. Kim, and H.-S. Cho, SETDB1 regulates SMAD7 expression for breast cancer metastasis, *BMB Rep.* **52**, 139 (2019).
- [32] S. Wang, X. Chen, H. Luan, D. Gao, S. Lin, Z. Cai, J. Liu, H. Liu, and Y. Jiang, Matrix-assisted laser desorption/ionization mass spectrometry imaging of cell cultures for the lipidomic analysis of potential lipid markers in human breast cancer invasion, *Rapid Commun. Mass Spectrom.* **30**, 533 (2016).




- [33] L. C. Whelan, K. A. R. Power, D. T. McDowell, J. Kennedy, and W. M. Gallagher, Applications of SELDI-MS technology in oncology, *J. Cell. Mol. Med.* **12**, 1535 (2008).
- [34] R. Eghlimi, X. Shi, J. Hrovat, B. Xi, and H. Gu, Triple negative breast cancer detection using LC-MS/MS lipidomic profiling, *J. Proteome Res.* **19**, 2367 (2020).
- [35] N. Kanyo, K. D. Kovács, S. V. Kovács, B. Béres, B. Peter, I. Székács, and R. Horvath, Single-cell adhesivity distribution of glycocalyx digested cancer cells from high spatial resolution label-free biosensor measurements, *Matrix Biol. Plus* **14**, 100103 (2022).
- [36] V. K. Lam, T. C. Nguyen, B. M. Chung, G. Nehmetallah, and C. B. Raub, Quantitative assessment of cancer cell morphology and motility using telecentric digital holographic microscopy and machine learning, *Cytom. Part A* **93**, 334 (2018).
- [37] H. J. Pandya, K. Park, and J. P. Desai, Design and fabrication of a flexible MEMS-based electro-mechanical sensor array for breast cancer diagnosis, *J. Micromech. Microeng.* **25**, 075025 (2015).
- [38] K. Park, W. Chen, M. A. Chekmareva, D. J. Foran, and J. P. Desai, Electromechanical coupling factor of breast tissue as a biomarker for breast cancer, *IEEE Trans. Biomed. Eng.* **65**, 96 (2018).
- [39] N. Ayyanar, G. T. Raja, M. Sharma, and D. S. Kumar, Photonic crystal fiber-based refractive index sensor for early detection of cancer, *IEEE Sensors J.* **18**, 7093 (2018).
- [40] D. Sun, Y. Ran, and G. Wang, Label-free detection of cancer biomarkers using an in-line taper fiber-optic interferometer and a fiber Bragg grating, *Sensors* **17**, 2559 (2017).
- [41] M. A. Ahmad, A. Najjar, A. E. Moutaouakil, N. Nasir, M. Hussein, S. Raji, and A. H. Alnaqbi, Label-free cancer cells detection using optical sensors, *IEEE Access* **6**, 55807 (2018).
- [42] M. Verma, S. Tirkey, S. Yadav, D. Sharma, and D. S. Yadav, Performance assessment of a novel vertical dielectrically modulated TFET based biosensor, *IEEE Trans. Electron Devices* **64**, 3841 (2017).
- [43] A. Chhabra, A. Kumar, and R. Chaujar, Sub-20 nm GaAs junctionless FinFET for biosensing application, *Vacuum* **160**, 467 (2019).
- [44] S. Singh and S. Singh, Dopingless negative capacitance ferroelectric TFET for breast cancer cells detection: Design and sensitivity analysis, *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **69**, 1120 (2022).
- [45] K. H. Chen, B. S. Kang, H. T. Wang, T. P. Lele, F. Ren, Y. L. Wang, C. Y. Chang, S. J. Pearton, D. M. Dennis, J. W. Johnson, P. Rajagopal, J. C. Roberts, E. L. Piner, and K. J. Linthicum, c-erbB-2 sensing using AlGaIn/GaN high electron mobility transistors for breast cancer detection, *Appl. Phys. Lett.* **92**, 192103 (2009).
- [46] R. Ramesh, M. Madheswaran, and K. Kannan, Nanoscale FinFET sensor for determining the breast cancer tissues using wavelet coefficients, *J. Mech. Med. Biol.* **11**, 1295 (2011).
- [47] V. B. Sreenivasulu and V. Narendar, Junctionless SOI FinFET with advanced spacer techniques for sub-3 nm technology nodes, *AEU Int. J. Electron. Commun.* **145**, 154069 (2022).
- [48] B. Kumar and R. Chaujar, Numerical study of JAM-GS-GAA FinFET: A fin aspect ratio optimization for upgraded analog and intermodulation distortion performance, *Silicon* **14**, 309 (2022).
- [49] Y. C. Huang, M. H. Chiang, S. J. Wang, and J. G. Fossum, GAAFET versus pragmatic FinFET at the 5nm Si-based CMOS technology node, *IEEE J. Electron Devices Soc.* **5**, 164 (2017).
- [50] B. Kumar and R. Chaujar, Analog and RF performance evaluation of junctionless accumulation mode (JAM) gate stack gate all around (GS-GAA) FinFET, *Silicon* **13**, 919 (2021).
- [51] N. Gupta and R. Chaujar, Optimization of high-k and gate metal work function for improved analog and intermodulation performance of gate stack (GS)-GEWE-SiNW MOSFET, *Superlattices Microstruct.* **97**, 630 (2016).
- [52] A. Kerber, E. Cartier, L. Pantisano, R. Degraeve, T. Kauerauf, Y. Kim, A. Hou, G. Groeseneken, and H. E. Maes, Origin of the threshold voltage instability in SiO<sub>2</sub>/HfO<sub>2</sub> dual layer gate dielectrics, *IEEE Electron Device Lett.* **24**, 87 (2003).
- [53] K. Onishi, C. S. Kang, R. Choi, H. J. Cho, S. Gopalan, R. E. Nieh, S. A. Krishnan, and J. C. Lee, Improvement of surface carrier mobility of HfO<sub>2</sub> MOSFETs by high-temperature forming gas annealing, *IEEE Trans. Electron Devices* **50**, 384 (2003).
- [54] S.-H. Chen, W.-S. Liao, H.-C. Yang, S.-J. Wang, Y.-G. Liaw, H. Wang, H. Gu, and M.-C. Wang, High-performance III-V MOSFET with nano-stacked high-*k* gate dielectric and 3D fin-shaped structure, *Nanoscale Res. Lett.* **7**, 431 (2012).
- [55] B. Yang, P. D. Ye, J. Kwo, M. R. Frei, H.-J. L. Gossmann, J. P. Mannaerts, M. Sergeant, M. Hong, K. Ng, and J. Bude, Impact of metal/oxide interface on DC and RF performance of depletion-mode GaAs MOSFET employing MBE grown Ga<sub>2</sub>O<sub>3</sub>(Gd<sub>2</sub>O<sub>3</sub>) as gate dielectric, *J. Cryst. Growth* **251**, 837 (2003).
- [56] C. P. Chen, Y. J. Lee, Y. C. Chang, Z. K. Yang, M. Honga, J. Kwo, H. Y. Lee, and T. S. Lay, Structural and electrical characteristics of Ga<sub>2</sub>O<sub>3</sub> (Gd<sub>2</sub>O<sub>3</sub>) GaAs under high temperature annealing, *J. Appl. Phys.* **100**, 104502 (2006).
- [57] P. D. Ye, G. D. Wilk, B. Yang, J. Kwo, S. N. G. Chu, S. Nakahara, H.-J. L. Gossmann, J. P. Mannaerts, M. Hong, K. K. Ng, and J. Bude, GaAs metal-oxide-semiconductor field-effect transistor with nanometer-thin dielectric grown by atomic layer deposition, *Appl. Phys. Lett.* **83**, 180 (2003).
- [58] H. C. Lin, P. D. Ye, and G. D. Wilk, Leakage current and breakdown electric-field studies on ultrathin atomic-layer-deposited Al<sub>2</sub>O<sub>3</sub> on GaAs, *Appl. Phys. Lett.* **87**, 182904 (2005).
- [59] G. A. Niklasson, C. G. Granqvist, and O. Hunderi, Effective medium models for the optical properties of inhomogeneous materials, *Appl. Opt.* **20**, 26 (1981).
- [60] Wikipedia, Effective Medium Approximations, accessed Jan. 27, 2023, [https://en.wikipedia.org/wiki/Effective\\_medium\\_approximations](https://en.wikipedia.org/wiki/Effective_medium_approximations).
- [61] A. Razavieh, P. Zeitzoff, and E. J. Nowak, Challenges and limitations of CMOS scaling for FinFET and beyond architectures, *IEEE Trans. Nanotechnol.* **18**, 999 (2019).
- [62] B. Kumar, M. Sharma, and R. Chaujar, Gate electrode work function engineered JAM-GS-GAA FinFET for analog/RF applications: Performance estimation and optimization, *Microelectronics J.*, **135**, 105766 (2023).
- [63] Y. Omura, H. Konishi, and K. Yoshimoto, Impact of fin aspect ratio on short-channel control and drivability of multiple-gate SOI MOSFET's, *J. Semicond. Tech. Sci.* **8**, 302 (2008).
- [64] S. K. Mohapatra, K. P. Pradhan, D. Singh, and P. K. Sahu, The role of geometry parameters and fin aspect ratio of sub-20 nm SOI FinFET: An analysis towards analog and RF circuit design, *IEEE Trans. Nanotechnol.* **14**, 546 (2015).

- [65] *ATLAS User's Manual* (SILVACO International, Santa Clara, CA, 2016).
- [66] B. Kumar and R. Chaujar, TCAD temperature analysis of Gate Stack Gate All Around (GS-GAA) FinFET for improved RF and wireless performance, *Silicon* **13**, 3741 (2021).
- [67] P. D. Ye, G. D. Wilk, J. Kwo, B. Yang, H.-J. L. Gossmann, M. Frei, S. N. G. Chu, J. P. Mannaerts, M. Sergent, M. Hong, K. K. Ng, and J. Bude, GaAs MOSFET with oxide gate dielectric grown by atomic layer deposition, *IEEE Electron Device Lett.* **24**, 209 (2003).
- [68] B. Kumar and R. Chaujar, Numerical simulation of analog metrics and parasitic capacitances of GaAs GS-GAA FinFET for ULSI switching applications, *Eur. Phys. J. Plus* **137**, 110 (2022).
- [69] D. Martens and P. Bienstman, Study on the limit of detection in MZI-based biosensor systems, *Sci. Rep.* **9**, 5767 (2019).
- [70] A. Wei, M. J. Sherony, and D. A. Antoniadis, Transient behavior of the kink effect in partially-depleted SOI MOSFET's, *IEEE Electron Device Lett.* **16**, 494 (1995).
- [71] P. Dwivedi, R. Singh, B. S. Sengar, A. Kumar, and V. Garg, A new simulation approach of transient response to enhance the selectivity and sensitivity in tunneling field effect transistor-based biosensor, *IEEE Sens. J.* **21**, 3201 (2021).
- [72] Y. Zou, Q. Liu, X. Yang, H.-C. Huang, J. Li, L.-H. Du, Z.-R. Li, J.-H. Zhao, and L.-G. Zhu, Label-free monitoring of cell death induced by oxidative stress in living human cells using terahertz ATR spectroscopy, *Biomed. Opt. Express* **9**, 14 (2018).
- [73] S. Myung, A. Solanki, C. Kim, J. Park, K. S. Kim, and K.-B. Lee, Graphene-encapsulated nanoparticle-based biosensor for the selective detection of cancer biomarkers, *Adv. Mater.* **23**, 2221 (2011).
- [74] S. Akhtartavan, M. Karimi, N. Sattarahmady, and H. Heli, An electrochemical signal-on apta-cyto-sensor for quantitation of circulating human MDA-MB-231 breast cancer cells by transduction of electro-deposited non-spherical nanoparticles of gold, *J. Pharmaceutical Biomed. Anal.* **178**, 112948 (2020).
- [75] T. Li, Y. Liang, J. Li, Y. Yu, M.-M. Xiao, W. Ni, Z. Zhang, and G.-J. Zhang, Carbon nanotube field-effect transistor biosensor for ultrasensitive and label-free detection of breast cancer exosomal miRNA21, *Anal. Chem.* **93**, 15501 (2021).



# From methods to datasets: A survey on Image-Caption Generators

Lakshita Agarwal<sup>1</sup> · Bindu Verma<sup>1</sup> 

Received: 9 May 2023 / Revised: 22 July 2023 / Accepted: 18 August 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

## Abstract

Image - Caption Generator is a popular Artificial Intelligence research tool that works with image comprehension and language definition. Creating well-structured sentences requires a thorough understanding of language in a systematic and semantic way. Being able to describe the substance of an image using well-structured phrases is a difficult undertaking, but it can have a significant impact in terms of assisting visually impaired people in better understanding the images' content. Image captions has gained a lot of attention as a study subject for various computer vision and natural language processing (NLP) applications. The goal of image captions is to create logical and accurate natural language phrases that describes an image. It relies on the caption model to see items and appropriately characterise their relationships. Intuitively, it is also difficult for a machine to see a typical image in the same way that humans do. It does, however, provide the foundation for intelligent exploration in deep learning. In this review paper, we will focus on the latest in-depth advanced captions techniques for image captioning. This paper highlights related methodologies and focuses on aspects that are crucial in computer recognition, as well as on the numerous strategies and procedures being developed for the development of image captions. It was also observed that Recurrent neural networks (RNNs) are used in the bulk of research works (45%), followed by attention-based models (30%), transformer-based models (15%) and other methods (10%). An overview of the approaches utilised in image captioning research is discussed in this paper. Furthermore, the benefits and drawbacks of these methodologies are explored, as well as the most regularly used data sets and evaluation processes in this sector are being studied.

**Keywords** Image- Caption Generator · Natural language processing · Computer vision · Intelligent exploration · Deep learning

---

Bindu Verma contributed equally to this work.

---

✉ Bindu Verma  
[bindu.cvision@gmail.com](mailto:bindu.cvision@gmail.com)

Lakshita Agarwal  
[lakshitaagarwal\\_2k21phdit05@dtu.ac.in](mailto:lakshitaagarwal_2k21phdit05@dtu.ac.in)

<sup>1</sup> Department of Information Technology, Delhi Technological University, 110042 Delhi, India

# 1 Introduction

Image captions, also known as cut lines, are just a few lines of text used to describe and explain the objects used in an image. In some cases, captions and cut lines are separated, whereas captions are captions/short descriptions (usually one line) of the picture. While the cut line is long, a prose blog under captions, generally describing an image, giving context or linking to any relevant topic. Captions can also be obtained by few default image caption generating software that are present online [1].

Image caption generator is a function that consists of the computer-assisted imaging system as well as the natural language processing concepts for visualizing the images and to form sentences and expressions in a natural language such as English. It is a popular Artificial Intelligence (AI) research tool that works with image comprehension and language definition. Creating well-structured sentences requires an understanding of language in the form of structures and semantics. Describing the content of an image with the help of well-structured sentences is a very challenging task. Still, it can also have profound effects, like it can be used for helping the visually impaired people for better understanding of the content present in an image. This task is challenging when compared with the separation of images or the observations of a well-researched object. The biggest challenge is to be able to create a definition that should not only capture the content of the image but also highlight how these objects are related to each other [2].

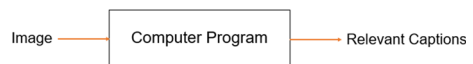
Caption generation in AI is also a challenge that is used to produce a readable textual meaning when given a picture. It requires both image comprehension in computer vision and a language model from NLP. It is also essential to consider and evaluate different ways in which a prediction model can be created as well as to create a problem-solving framework for generating captions for any given images. Photographic captions have steadily caught the attention of many researchers in recent years, because of the rapid development in the field of AI.

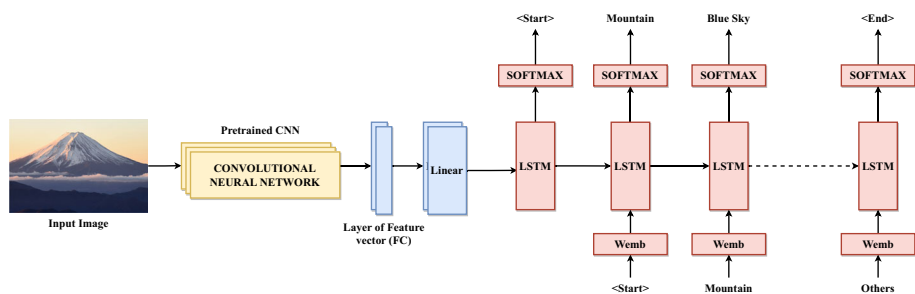
The goal of the research efforts for image captioning leads the researchers to find most efficient algorithms for processing the images and then representing its content as well as to translate those contents into the words by creating the links between all the visual elements and with the textual elements of the image, while also maintaining the language fluency of the captions. In the easier description, image captions are an image-to-sequence problem for the given pixel input of an image. And lastly, a sequence of words is being demonstrated according to a shared vocabulary of any language.

The following Fig. 1, displays the basic model for image caption generator, where an input image is taken from the given data-set and is further passed through the computer program. After passing it through the computer program, the relevant captions are then generated. The computer program mainly consist of different machine learning algorithms, traditional and deep network models that can be used for producing the captions.

Over the past few years, there have been many significant developments in the field of image captioning. Different models have been formed, for example, models that helped in the first in-depth study based on the Recurrent Neural Networks (RNNs) and producing global image definitions. There have been methods that have been developed with enhanced awareness and learning. Currently, work is being done with the help of the transformersmodel.

**Fig. 1** Basic Model of Image-Caption Generator





**Fig. 2** Image-Caption Generator using CNN model

Figure 2 depicts the proper procedure and functioning of the image caption generation model where an input image is passed through a CNN model, which is pre-trained to use different data sets. After this, the feature vectors are extracted from the fully connected layer of the model. Then these feature vectors are forwarded linearly towards the Long Short-Term Memory (LSTM) model of network, where the words and captions are classified using the activation function called SoftMax and are further forwarded to the embedding of words to form the sentences for the input image.

The area of NLP and computer vision have faced the challenges for developing appropriate assessment principles and testing metrics to compare the results with the basic facts. There are still many challenges that are being faced while comparing captions generated by the system with those done by a normal human being [3, 4]. To produce captions from a machine that match those of a human being, the first challenge is having thorough knowledge of how natural language is processed. How to create fully grammatically correct image captions for any given images is another challenge that frequently arises. Then finally, the issue is how to make an image's semantics as consistent as feasible while simultaneously creating captions that are as clear and understandable as possible for people.

The image caption generators can have various applications, such as the self-driving cars used for automatic driving. As we know, automated driving or self-driving cars face a lot of challenges, so if we know the exact location of the incidents happening near the car beforehand, then it can be helpful for improving the complete self-driving system. The image caption generator can also be implemented as the assistance for blind people. There can be a real-time system which will guide the person on the streets without the support of anyone else. It can be done by first translating an item into text and then the text into voice. It can also be used for the visually impaired people by combining NLP with Automatic Image Captioning to help them understand the nearby environment and surroundings. Another primary applications of image captioning is video surveillance. CCTV cameras are everywhere today, but with the world view, if we can generate the appropriate captions, we may be able to set alarms as soon as there is any serious malpractice that happens. This can help to reduce a lot of crime and also the occurring accidents. Automatic captions can also be helpful in terms of making google image search as all images can first be converted into captions and search can be done based on those generated captions. Image caption generators can also be used for annotation of different images and understanding the type of content used and displayed on different social media platforms.

Image captions, which automatically generate natural language meanings based on visual content, are an important part of the scenario, which includes computer visualization and NLP. This technique has recently got a lot of attention and is now one of the most important

area of research in computing. Image descriptions are hence obtained by predicting the most likely verbs, nouns, prepositions, etc., that will make up a different sentences for any set of images [5]. There has already been a lot of research done in this area, and some image caption survey studies are also available, but none of them address the significant models that have been developed and implemented between the year 2004 and 2022. In the developing field at this level, it is important to keep track of all the latest models and frameworks [3]. In addition, a number of domain-related image captioning processes have been studied. However, the results achieved so far are not the final solutions. There are still few research gaps and challenges that are needed to be solved for the novel idea of image caption generation.

In contrast to earlier survey papers on image captioning, this work offers a thorough and comparative analysis of both conventional and deep learning-based image captioning models. This paper also discuss a various image captioning methods over the last decades. As image captioning has various applications in the real-world scenario, the motivation behind this study is that, it surveys a much greater number of papers on the subject and offers a more in-depth examination of the suggested approaches. Additionally, the study discusses how to use the standard evaluation metrics to provide better results in image captioning. It also summarizes the broader set of available datasets that can be used. And finally, the study performs quantitative comparisons of all the approaches and provides us the information regarding the challenges and research gaps discovered till date. The contribution of the paper are:

- We thoroughly review the paper published in the duration 2004 – 2022 on image caption generators using traditional and deep learning-based methods.
- We discussed the various deep learning models with their advantages and disadvantages in Table 3
- The main datasets used for analyzing image captions are examined, including both standard domain benchmarks and specific domain databases.
- Standard and unconventional metrics for testing performance and analyzing different caption features are analyzed.
- The paper presents an overview of the diversity of image caption generation, its research gaps, open-ended challenges and future research directions.

The rest of the paper is organized as follows: Related work is discussed in Section 2. In Section 3, various methods and models are discussed in detail. Section 4 explains all the published benchmarks data sets in the field of image caption generation. In Section 5 various evaluation matrices are being discussed. Section 6 discusses the challenges and a few research gaps. Finally, Section 7 provides a conclusion with the future directions of the work.

## 2 Review of literature

A large portion of the language is used to describe the world around us, particularly the visual world in which we live or the world which is depicted in images, photographs or videos. Image captioning has received a lot of attention as an interdisciplinary study subject combining computer vision and natural language processing. The goal of image captioning is to create natural language sentences that are fair and accurate [2, 6]. The captioning model must recognize different objects and then appropriately represent relationships between those objects. Also, external knowledge is necessary for generating different informative sentences as well as for improving knowledge reasoning in the generated captions for given images. Natural language processing (NLP) applications include machine translation, summarizing,



dialogue systems and machine-assisted reviews. The majority of past NLP work on automatically creating captions or descriptions for photographs has relied on retrieval as well as have provided a summary for the images [7]. After so much research done till now, the main question that arises is where to put the images in the image caption generator? Deep learning, on the other hand, is currently considered to be the foundation for intelligent exploration for creating accurate image captioning models [4]. In this literature, we will discuss different approaches that have been developed for the image captioning generation. Following Fig. 3 displays few examples of image descriptions obtained from different approaches.

## 2.1 Automatic image captioning

Automatically describing an image with the help of words and sentences is known as image captioning. Image semantic information must be gathered and conveyed in the natural languages to meet the purpose of captioning of images. Image captioning is a complex topic connecting computer vision and natural language processing research fields. A few of the researches based on the automatic captioning of the image are presented below.



A baseball player is getting ready to hit the ball.  
The man at bat reads to swing at the pitch while the umpire looks on.



A bus is parked on the side of a street.  
A large bus sitting next to a very tall building.



A herd of giraffes walk down the street in the middle of some trees.  
A horse carrying a large load of hay and two people sitting on it.



A kitchen with a sink and mirror next to a wall.  
Bunk bed with a narrow shelf sitting underneath it.



A white teddy bear sitting on top of a laptop  
A woman is typing on a laptop on a wooden table.



A piece of luggage sitting on top of a counter.  
A faucet running next to a dinosaur holding a toothbrush.



A fire hydrant on a lush green field.  
A fire hydrant is placed in a wooded area.



A group of people standing around a table.  
A very pretty lady eating a big pizza.



A man with a colorful umbrella walking down a street.  
A bald man holding a blue umbrella on a street.



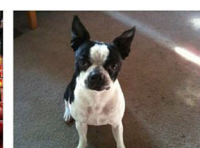
Amazing colours in the sky at sunset with the orange of the cloud and the blue of the sky behind.



A female mallard duck in the lake at Luukki Espoo



Fresh fruit and vegetables at the market in Port Louis Mauritius.



Street dog in Lijiang



Tree with red leaves in the field in autumn.



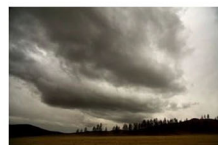
One monkey on the tree in the Ourika Valley Morocco



Clock tower against the sky.



The river running through town I cross over this to get to the train



Strange cloud formation literally flowing through the sky like a river in relation to the other clouds out there.



The sun was coming through the trees while I was sitting in my chair by the river

Fig. 3 Sample for image captioning [8, 9]

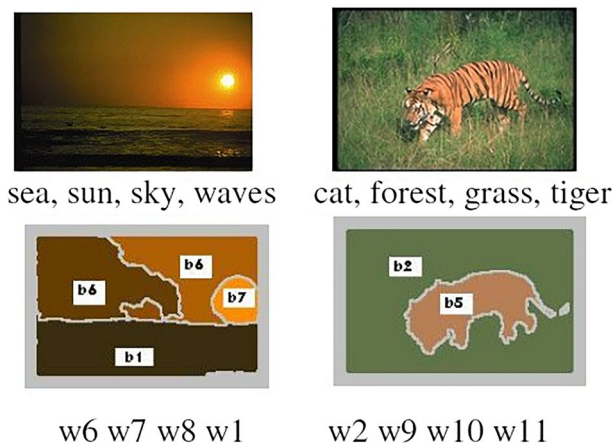
Authors J. Jeon et al. [10] offered an intuitive way to annotate and retrieve photographs. They proposed that a tiny vocabulary of blobs can be used to characterize regions in an image. Cross-Media Relevance Models have been proved to be a valuable choice for annotating and retrieving photos. They demonstrated how to use several models for the ranked retrieval. They also anticipated that adopting formal information retrieval models to this field of research would be fruitful. Although it is challenging, the authors think that labelled training and testing data are essential for the performance and evaluation of the algorithms that are suggested. The outcomes will probably be improved by better feature extraction or the usage of continuous features for an image. The use of real captions (rather than keywords) is another area that could be the subject of research; in their opinion, this is a promising field for the use of formal models of information retrieval.

A new term for automatic linguistic indexing of images was introduced by Jia Li et al. [11]. The authors presented a statistical modelling approach to this problem in this work. They focused on a specific class of stochastic processes with the help of two-dimensional multi-resolution hidden Markov models (2D MHMMs). The following were the significant advantages of this approach: 1) distinct models for different concepts can be separately trained and retrained, 2) it is possible to train and store a huge number of concepts and 3) the spatial link between image pixels is taken into consideration utilising probabilistic likelihood as a universal metric inside and across resolutions. Experiments have shown that the approach was accurate and had a lot of potential for linguistic indexing of photographic images. But, there were a number of restrictions with the way the current system is implemented and evaluated. For instance, training with 2D still images may make it harder to understand concepts later on. Also, because some categories of images are visually widely dispersed, it is hard to teach a notion using just a small number of these images. The presented evaluation results should be treated with caution until this constraint is fully explored. In general, it took more time and experience to understand more complicated topics.

In image database management, image annotation is often utilized. Therefore, Patrick Hède et al. [12] offered a new method for automatically describing images in natural language for images without occlusion concerns. This strategy was based on a combination of picture indexing and natural language processing along-with creation expertise using NLP techniques. Furthermore, by decreasing the ambiguity in a keywords index, indexing in natural language sentences increased the quality of the results. Finally, this research was able to lay a solid foundation for future work on more complicated images.

According to [13], Jia-Yu Pan et al., in their research work, they focused mainly on the automatic image captioning. They have worked upon a training set consisting of captioned images and have developed a relationship between different keywords and features of the image. Their approach achieved around 45% accuracy for providing the captions of the image, especially on the larger dataset consisting of various styles of content. Following that, it was determined that the suggested techniques may be used in other contexts, such as creating an image vocabulary of various cell kinds using illustrations from medical journals. Further, new methods of Corr, SvdCorr, SvdCos and Cos were being proposed. It was also observed that the consistent generation in the improvement of the 'adaptive' blob-tokens has led to the incrementation of the performance rate inside the system. Figure 4 displays the word tokens and blobs generated for the annotated images.

Author Siming Li et al. [14], offered a simple but successful method for automatically composing image descriptions utilizing web-scale n-grams and computer vision-based inputs. In this research, they developed a unique surface realization technique for automatically generating descriptions of the images that were based on the n-gram data of the web-scale. This system was extremely good at producing largely pleasant and presentable language



**Fig. 4** Annotated Images with their captions and there corresponding Blob and word tokens [13]

sentences while also allowing for some creative writing. Furthermore, they showed how implicitly encoded world information in natural language might aid image content detection. This method uses far more basic ways to handle photos in the open-domain in a more natural way. They employed the same vision-based modifier classifiers, inputs-object detectors and the prepositional functions. But they concentrated primarily on refining the sentence production process to produce descriptions that are more similar to those written by humans.

Author Vicente Ordonez et al. [8] used a huge annotated photo collection to create and showcase automatic image captioning systems. The method for automatically collecting the new dataset of Flickr was solving many queries and was then filtering the noisy results down to 1 million photographs with visually interesting captions. They also proposed approaches that included a variety of state-of-the-art, yet noisy, image content estimated to produce even more appealing outcomes. Finally, a new objective of performance measure for image captioning was introduced. They also described two versions of their approach: one that composes captions solely using global picture descriptors and another that includes estimations of image content for caption synthesis. Author Rebecca Mason and Eugene Charniak et al. [15] demonstrated a data-driven framework for creating image captions that included visual and linguistic characteristics with variable degrees of spatial structure. They suggested the task of domain-specific image captioning [16]. They also employed a hybrid visual and textual bag-of-words approach to estimate individual word accuracy, pulling previously written descriptions from a database and adjusting them to new query images. They demonstrated that their captioning method effectively deletes terms that have errors in the extracted captions while preserving a high level of details in the resulting output using automatic and human evaluations.

Another approach was being introduced by Jacob Devlin et al. [17] that was based on exploring the nearest neighbours for image captioning. For image captioning, they investigated a number of nearest neighbour baseline techniques. When tested using automatic assessment metrics on the evaluation server of the MS COCO caption dataset, this technique surpassed many contemporary algorithms that generated innovative captions. The success of nearest neighbour captioning methods highlighted the need for better dataset evaluation and testing. In this study, they just looked at basic Neural Network (NN) techniques to give the picture captioning research work, a base. Authors Jack Hessel et al. [18] investigated whether recent promising outcomes in automatic caption synthesis can be attributed only to language

models or there can be any other ways for image captioning? They discovered that a neural captioning system could produce high-quality captions even when the image quality was extremely low. They also ran additional tests to see if datasets were suitable for captioning automatically and also discovered that having many captions for any given image can be beneficial but it cannot be necessary. They demonstrated a relationship between the accuracy of classification of the CNN model and for the caption quality which is generated by the neural captioning approach. In the most recent approach, the authors Rashid khan et al. [19] used numerous pre-trained convolutional neural networks to encode a picture into a feature vector as graphical properties. To increase the performance, the researchers combined the attention model of Bahdanau with the GRU. This allowed the researchers to be focused on a specific portion of the image only. On the dataset of MSCOCO, the experimental results found in this approach were better when compared with the pre-existing state-of-the-art approaches. This model was a single model that combined CNN and GRU with an attention network for automatic image captioning. Experiments have proven that the proposed approach can generate meaningful descriptions for the images automatically.

## 2.2 Human-like image caption generation techniques

Human ratings are considered to be the most accurate approach for assessing the accuracy of an image captioning model. Evaluating captions on the basis of human feed-backs as well as generating captions exactly like the phrases of the humans is mostly considered to be most accurate. Hence, multiple researchers demonstrated different approaches to generate captions that are more in human-like form or we can say in the natural language.

P. Kuznetsova et al. [20] described a data-driven and holistic method to image description generation that takes advantage of the large amount of (noisy) parallel image data and natural language descriptions available on the web. They obtained existing human-composed phrases that were used for describing the visually comparable photos and then selectively mixing those phrases for constructing a novel description for the image being queried. Their method was the first to systematically combine cutting-edge computer vision application to obtain visually relevant candidate sentences and then generate image descriptions that are significantly more complex and human-like than earlier attempts. M. Mitchell et al. [21] described a new generation technique for creating human-like descriptions of images based on computer vision detection. They have created a new generation system that combines a syntactic model with computer vision detection to generate language. They developed a framework Midge that creates a well-formed description of an image. Midge's output is considered by humans to be the most natural description of images generated thus far. They further planned to improve and refine the system to capture more linguistic phenomena.

In the next paper, authors P.H. Seo, et al. [22], worked upon the image caption generators that were based upon the signals obtained from the ratings of human captions at instance-level. They have created a method which was used for maximizing the ratings of humans using different samples for the datasets. This method compared two policies for developing positive/negative feedback for getting better captions for any given image inside the dataset. It also concluded that, when the researchers used the reinforcement learning (RL) technique on off-policy, it showed successful results for dealing with information regarding the rating. This work could be considered as the base work for working upon the techniques based on RL technique as well as for considering human feed backs and ratings for generating image captions. Similarly, authors Yue Zheng et al. [23] proposed a novel approach that was used for generating the captions with guiding objects (CGO). These guiding objects of the captions

were mainly related to the objects which were useful for the humans inside any image as shown in Fig. 5.

In this work, the authors designed a framework that consisted of two models of LSTMs but combined in opposite directions using the images of MS-COCO dataset. The CGOs were also able to generate captions with great advancement in terms of describing the objects that were novel. The results were being displayed on datasets of two platforms: ImageNet and MS-COCO. The evaluation of this method showed better result in terms of both: the content fluency as well as the accuracy of the descriptions generated. Only one chosen object was guaranteed to be stated in the CGO strategy, but this did not hinder its viability. To include more objects in the outputs, CGO can be used along with other techniques like CBS. Additionally, it guarantees that the descriptions are include the guiding items.

## 2.3 Role of RNNs, CNNs and LSTM in an image caption generator

Deep learning based framework are being tremendously used in almost every field of the computer vision and human-computer interaction. In automatic image caption generation authors X. Chen et al. [9] explored a new technique of the bi-directional mapping between photographs and their sentence-based descriptions. They presented the first bi-directional model which was capable of producing smooth image descriptions as well as visual attributes. Unlike many earlier RNN-based techniques, this model can learn long-term interactions, to recreate visual features as new words. In the year 2014, a new deep captioning approach was introduced which used the multi-modal recurrent neural networks, known as the m-RNN [24]. A deep RNN model for texts and a deep convolutional network for images were included in the model. The entire m-RNN model was made up of these two sub-networks consisting of the interacting multi-modal layer. This model's effectiveness was verified using four benchmark datasets which were Flickr 8K, IAPR TC-12, MS COCO and Flickr 30K. Furthermore, they applied the model of m-RNN for the task of retrieving words or images, and found that it outperforms all the previous state-of-the-art algorithms. It was also concluded that, this model was capable of connecting different images and the sentences to accommodate complicated representations of the images and the advanced models of language. But, if the dictionary size is big, the embedded layer of the model has more parameters than the original m-RNN.

Several papers have mostly looked into how to learn joint feature spaces for images and their descriptions. These methods combines picture and language information into a single space that could be utilized for image search or image caption ranking. RNNs and the deep neural networks were among the methods which were used to train the projection. These methods were capable enough of projecting both semantics and visual traits to a common embedding, but not the inverse projection. Another bi-directional representation approach

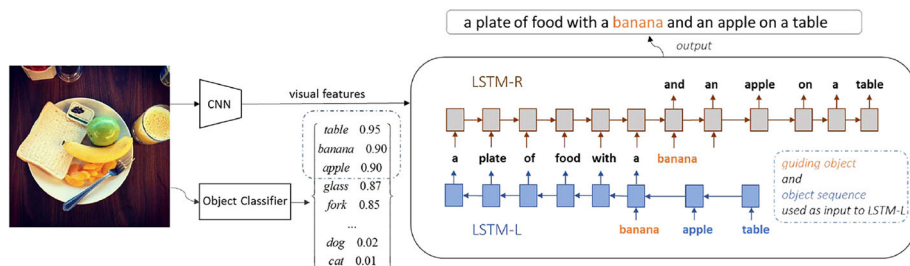


Fig. 5 Architecture for caption generation using Guiding Objects [23]



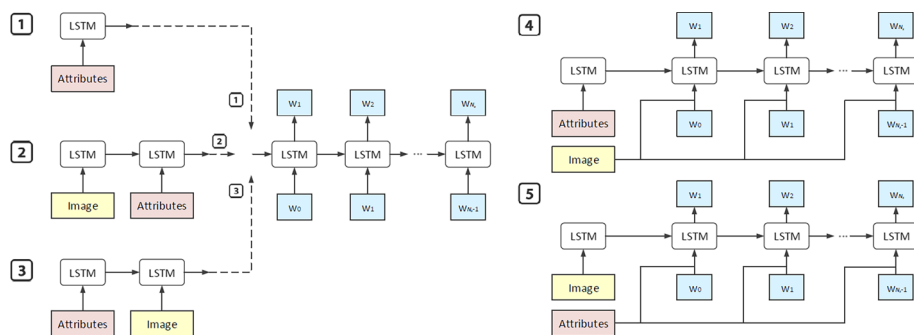
was proposed in [25] that generated unique descriptions from images as well as visual representations from descriptions. They tested their algorithm's bidirectional capabilities on both picture and text retrieval tasks. They also produced results that were better than or comparable to previous state-of-the-art results. But, RNNs have a history of having trouble remembering concepts after a few iterations of recurrence. For instance, RNN language models frequently struggle to learn distant relationships without the use of specialised gating units. The innovative dynamically updated visual representation serves as a long-term recall of the previously discussed notions during sentence production. This made it possible for the network to choose important ideas to communicate without speaking them. They showed that a visual representation of a written description may be produced using the same representation.

In the year 2015, researchers Andrej Karpathy and Li Fei-Fei [26] presented a model which generated descriptions of images in the form of the natural language. To discover inter-modal correspondences between linguistic and visual data, this method used image databases and sentence descriptions. They presented the architecture of a multi-modal RNN technique that learned to provide novel descriptions of the regions that were visually represented using the inferred alignments. They demonstrated that the descriptions that are generated outperformed the baselines of retrieval on the entire images as well as for a new dataset of annotation at region-level. Their method included a novel ranking algorithm that used a shared embedding of the multi-modal system to match elements of linguistic and visual modalities. They also tested the performances on two different tests of region-level and full-frame level, which further concluded that this model outperformed retrieval baselines in both the circumstances. In December, 2016 authors A. Mathews et al. [27], introduced a new RNN model namely, SentiCap that was used for generating descriptions of images with the help of sentiments. They created a system for characterizing a given image using the emotions, as well as created a model that generates captions that shows positive as well as negative feelings automatically. They then used various algorithmic and crowd-sourcing metrics to assess the captions. The crowd-sourced workers certified that 88% of the positive captions had the acceptable sentiment. It was a switching RNN model for creating sentiment-based image captions and was able to generate the caption which contains emotions for over 90% of the images. The creation of generative models for a wider range of emotions, such as pride, humiliation and wrath, as well as models for linguistic styles (including sentiments) beyond the level of words, are possible future research works.

The further work was done by researchers Ting Yao, et. al [28]. They presented the LSTM-A (Long Short-Term Memory with Attributes) model, which was considered to be a novel architecture that was used for integrating different types of attributes into the Convolutional Neural Networks(CNNs) and Recurrent Neural Networks(RNNs) inside the framework of image captioning architecture, as shown in Fig. 6. These models were then trained from beginning to end in the form of end-to-end manner.

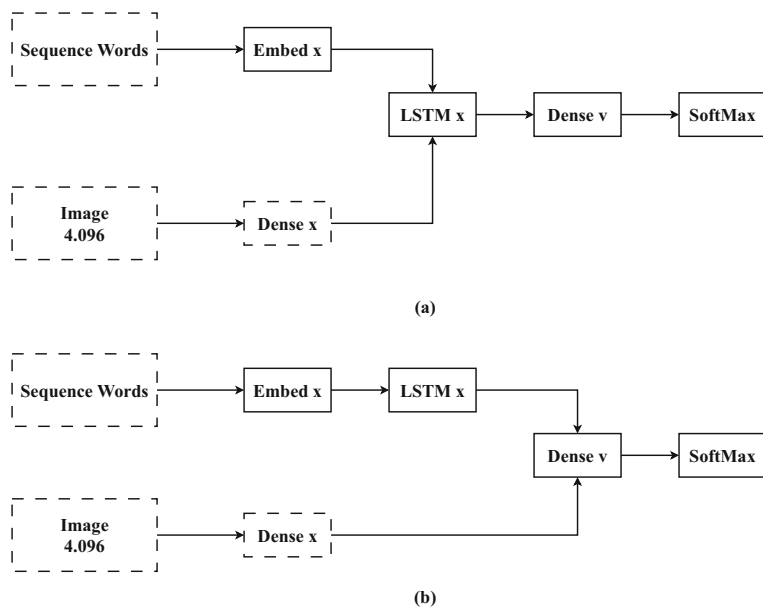
Another work also detected multiple attributes by further exploring the correlation of symbols within the Multiple Instance Learning(MIL) framework [29] and learning the problem of adding high-quality attributes from pictures to completing pictorial presentations to improve sentence production. They suggested a permutation-invariant aggregating operator based on neural networks that correlates to the attention process. Experiments were conducted on the MS-COCO dataset for validation of the work. In the next approach, authors Marc Tanti et al. [30] discussed about two architectures that were based on the RNNs in generating image captions. They further described that features of the image must be considered along the RNN as shown in Fig. 7.



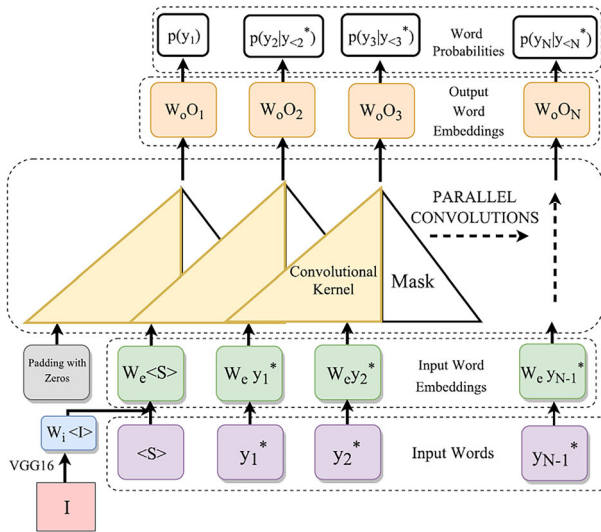


**Fig. 6** Variants of LSTM-A framework [28]

The paper also discussed about the three main classes in which the approaches to image captioning can be defined: the first class of approach was based on the systems that were dependent on the techniques of computer vision for the extraction of object detection and features from a source image. The second approach was mostly focused on systems that visualized the task as the problem of retrieval, which was characterised on the basis of string proximity and relevance utilising training data for any particular image. The third class of the approach was based on the system that is highly dependent on the neural models and can perform partial as well as complete retrieval of the captions using RNN along with LSTM. Jyoti Aneja et al. [31] in the year 2018, worked upon the convolution image captioning. The main contribution of this work was: A method that includes CNN based captioning of images when compared with LSTM based methods in terms of the performance. Using spatial image features for improving the performance of CNN based model for image captioning as shown in Fig. 8.



**Fig. 7** The diagrammatic representation of the two approaches investigated in the paper: a) Merge Architecture b) Inject Architecture



**Fig. 8** Model for convolutional image captioning [31]

The following paper resulted on the convolutional approach that was performed with the approach based on the LSTM for the captioning of the images. They also said that this approach has better improvement for different metrics and was also better in terms of the performance for the baseline of LSTM+Attn. The approach also showed that when they trained the CNN with more parameters (approximately 1.5 times), then the training of the whole system was done within the comparable time, because the system avoided using the RNNs in the sequential processing.

The process of captioning an image attempts to construct a sentence from a group of linguistic words that are used for describing the features, interactions as well as objects inside the given image, referred to as visual semantic units (VSUs). Based on this viewpoint, in the year 2019, the authors L. Guo et al. [32] proposed the Graph Convolutional Networks (GCNs) to explicitly represent object interactions in semantics and geometry, as well as to fully leverage the alignment between linguistic words and VSUs for image captioning. They proposed the concept to bridge the information gap between the content of visual representation and the description of language by employing VSUs. The VSUs were the components of visuals addressing objects which tells the relation of the images and the captions with their attributes and the interaction between different objects. The researchers used structured graphs to represent the VSUs consistently and GCNs to contextually embed them.

## 2.4 Attention-based image captioning approaches

The image is divided into  $n$  parts by an attention mechanism, which then computes an image representation for each portion of the image. Whenever RNN creates a new word, the models' mechanism concentrates with the appropriate portion of the given image, allowing the decoder to use only certain sections of the given image for implementation. Traditionally, image captioning has been accomplished using the encoder-decoder architecture using a convolutional neural network (CNN) serve as the encoder, and a recurrent neural network

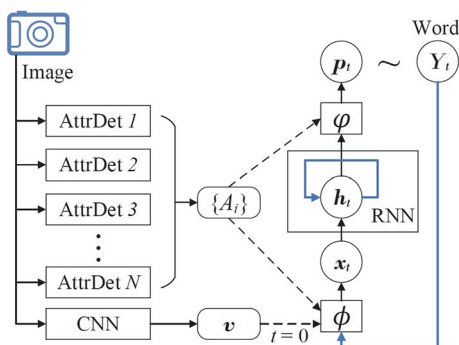
serves as the decoder (RNN). Both the generally categorised fields of machine learning, namely CV and NLP, are involved in image captioning. Large image classification datasets have made it possible for a rise in study on the subject, and as interest has grown. The decoder can more accurately translate the important features of the images into natural language when they are brought to the forefront by attention. An algorithm is tasked with creating a logical caption for an image when image captioning is being done. It is difficult to complete for a number of reasons, not the least of which is that it incorporates the idea of salience or relevance. Due to this, most contemporary deep learning systems contain an “attention” mechanism to aid in the emphasis on appropriate image elements.

In March 2016, Quanzeng You, et al. [33] worked upon the image captioning that can be done with the help of the semantic attention. In this paper, the researchers developed a new algorithm that was used in combining different approaches with the help of the model used in semantic attention as shown in Fig. 9. In this model, the CNN response visual features ( $v$ ) and attribute detection ( $A_i$ ) were introduced into the RNN (shown with dashed arrows) and then they were combined together with the help of a feedback loop (displayed with blue arrows). Here, both the input and output models required the qualities that needed more attention for generating the image captions. They further evaluated the algorithm which was based on two public benchmarks namely, MS-COCO and Flickr30K datasets. The results showed that this algorithm significantly outperformed different approaches with better consistency along with different metrics of evaluations.

It was also concluded that, this method was useful in exploiting the overview of input image as well as to review abundant fine-grain visuals, they were being observed semantically.

Existing methods focused solely on visual material, leaving the question of whether textual context might boost attention in image captioning or not. To investigate this issue authors L. Zhou et al. [34] offered the text-conditional attention. This was a unique attention mechanism that allowed the caption generators to focus on specific image features that were based on previously generated text. They used the guided Long Short-Term Memory (gLSTM) captioning architecture [35] with CNN fine-tuning to get text-related image characteristics for their attention model. Using the MS-COCO dataset, this strategy allowed for end-to-end learning of picture embedding, text-conditional attention, text embedding and finally a language models with a single network architecture. The results showed that the method outperformed all the other captioning systems on a range of quantitative and human evaluation parameters, showing that text-conditional attention could be applied in image captioning. Visual attention is critical for understanding images and it has been shown to be helpful in generating natural language descriptions of them. In the next approach, authors J. Mun et al. [36] provided a text-guided attention model for image captioning that learnt to drive

**Fig. 9** Framework of the image captioning model using semantic attention [33]



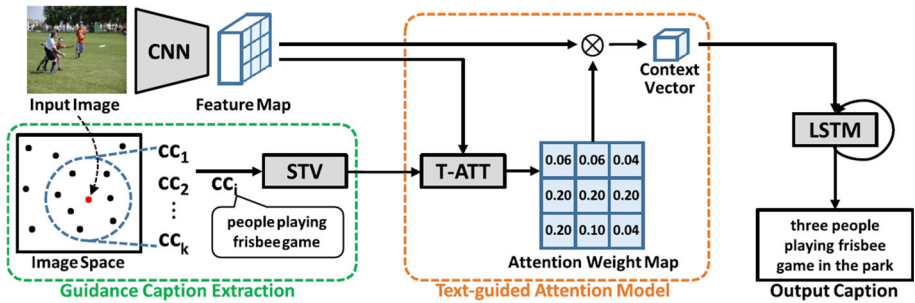
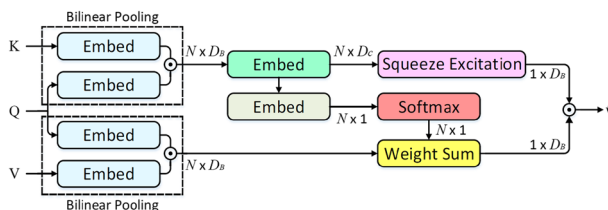


Fig. 10 Image captioning using architecture of text-guided attention. [36]

visual attention using linked captions. They presented an exemplar-based learning technique as shown in Fig. 10. This model collected relative captions with each of the given image from training data and used them for learning the attention on various visual attributes. They also provided a robust strategy for dealing with the noise present in the guided captions that further uses a set of captions in the form of the guidance captions with the methods of training and testing. They also used the model of text-guided attention on the dataset of MS-COCO captioning to get best-in-class results.

The authors K. Xu et al. [37] introduced an attention-based model that automatically learns to characterise the content of images, based on current work in machine translation and object detection. They demonstrated the effectiveness of attention using three datasets: Flickr30k, Flickr8k and MS COCO. They offered an attention-based strategy that used the METEOR and BLEU metrics to provide basic performance on these three datasets. They also demonstrated how learnt attention could be used for improving the interpretability of the process of model generation and that the alignment of learning which is closely related to human intuition. Further, authors Peter Anderson et al. [38] proposed a mechanism that was a combination of both bottom-up and top-down attention. This mechanism was used for enabling the attention that needed to be calculated at different object levels as well as for different salient regions of an image. The results of this approach were observed on the test server of MSCOCO which was able to create a better state-of-art for different tasks and functions. The approach reached the scores of CIDEr as 117.9, SPICE as 21.5 and of BLEU-4 as 36.9. The benefits of this approach can be taken by replacing the CNN's pre-trained features with advanced features of the bottom-up attention approach.

Authors Ting Yao et al. [39], have designed a method which can be used for exploring different connections that occur between the objects inside the framework of encoder-decoder. The method was also based on attention mechanism for the process of captioning an image. They mainly presented an architecture that was known as GCN-LSTM i.e., 'Graph Convolutional Network plus Long-Short Term Memory' which was useful for integrating spatial as well as semantic relationships of the objects of an image into the encoder. The large-scale experiments were done on the dataset of MS-COCO and then results were being evaluated. Further those results were compared with previously obtained results from different algorithms and methods. According to the authors, the architecture of GCN-LSTM had incremented the performance of CIDEr-D from the percentage value of 120.1% to the value of 128.7% when being tested on the dataset. Another approach of image captioning is encoder-decoder approach where the sentence generation for the hidden states is done by giving attention to the frameworks of image captioning [40–42]. In this method, the input image was being used with the help of the framework of encoder-decoder so that the process of



**Fig. 11** Architecture of X-Linear Attention block [43]

decoding was able to focus on different and particular aspects of the image and that would further help in generating description for the input image at every step.

In the next approach, authors have introduced a unique block for attention inside the image captioning model, known as the X-Linear Attention block [43]. This block was used to completely work on the bi-linear pools that could be further used for capitalizing selectively on the information provided visually as well as for working upon the reasoning on the multi-modals as shown in Fig. 11.

The experiments were performed on the COCO framework and it was observed that the performance of the CIDEr was found to be 132.0%, which was the best one till date. This model resulted up to this value of CIDEr when different transformations were done with X-Linear attention blocks in the model. They also included an ELU (Exponential Linear Unit) block which helped in using infinite interactions of features. This model showed better results in terms of efficacy with the integrated X-LAN and the X-Linear Attention block. As well as, it created a performance with better state-of-art for captioning the dataset that was used. The authors noted that sometimes Up-Down approach concentrates on an area of the image that was irrelevant and whose associated object shouldn't be formed at that instance of time. Instead, the X-LAN continually concentrated on the appropriate regions for captioning by taking advantage of higher order feature interactions for multi-modal reasoning via X-Linear attention block.

The ability to generate accurate words for any particular objects inside an image using the techniques of visual attention is known as the grounded image captioning. In the next paper, researchers Yuanen Zhou et al. [44], worked upon the improvement of the accuracy of the generated image captions with the help of grounded image captioning models. The authors proposed a model that was based on enhancing the matching ability of the given images and texts using the POS i.e., part-of-speech. Further, this model was named as POS-SCAN (where SCAN stands for Stacked Cross Attention Network) [45]. When POS-SCAN was used with conventional methods, it significantly showed better results in terms of accuracy without much interference. The researchers also worked upon the SCST method i.e., Self-Critical Sequence Training method [46], for generating captions for grounded images. The authors considered this technique to be an interesting one for designing different metrics of the machine for generating the captions, specifically the n-gram based metrics. The authors Z. Song et al. [47], proposed a mechanism of CAAG i.e., Context-Aware Auxiliary Guidance which helped the image captioning models to obtain the sentences using the contexts which are used globally, shown in Fig. 12.

The method proposed here was generic in nature and was used for improving the performance of the already existing models for reinforcement learning. CAAG also improved the performance of the CIDEr-D model based on the LSTM decoder to 128.8% value from the already evaluated value of 123.4%. It was found that when CAAG was utilised, a substantially larger percentage of generated captions were more descriptive than the other baseline model.

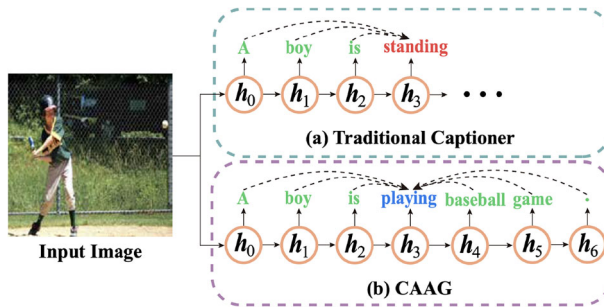


Fig. 12 Architecture of CAAG [47]

## 2.5 Image captioning based on policy-gradients

In December 2016, researchers Rennie et al. [46] worked upon the creation of a self-critical sequence training that was used for the process of image captioning. In this paper, the researchers have represented an efficient and simple approach that was used to compare more effectively supporting approaches. This approach was also known as the Reinforce algorithm that was used for providing the policy-gradient which were based upon the reinforcement learning (RL), as depicted in Fig. 13.

The results that were obtained on the MSCOCO dataset test server that further introduced modern performance as well as helped in enhancing the best CIDEr results from 104.9% to 114.7%. On the MS-COCO dataset, they have tried with both the TD-SCST and the “True” SCST as stated, however it was discovered that the methods did not produce any appreciable additional gain. Additionally, they tried unsuccessfully to train a control variable for the SCST baseline using MS-COCO. However, they also believed that these extensions will be crucial for other tasks involving sequence modelling and policy-gradient based RL in general.

In the next work, researchers Siqu Liu et al. [48] worked upon an improved method of image captioning using the optimization of policy gradients of SPIDER. They showed how to use the method of policy gradient (PG) to directly optimise a combination of SPICE and CIDEr linearly (which they named SPIDER). The SPICE score ensured that the captions are semantically correct for the image, whereas the CIDEr score ensured that the captions were

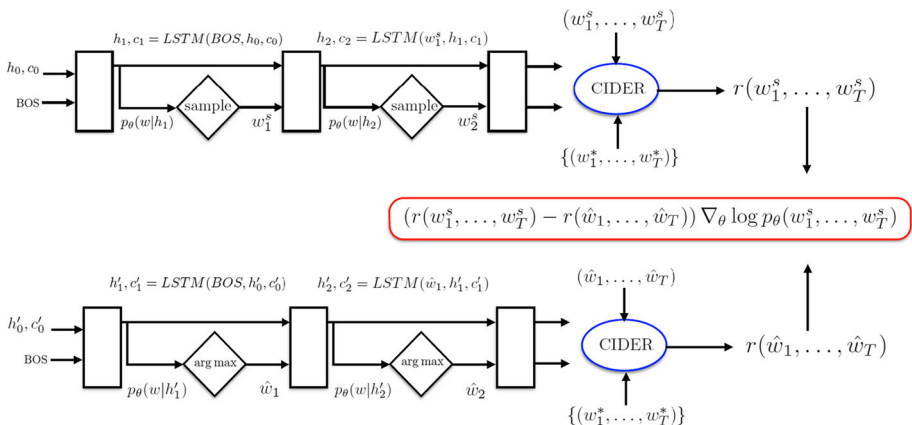


Fig. 13 Depiction of policy gradient based on CIDEr [46]



fluent enough to read and understand. To maximise the range of captioning parameters, they proposed and effectively implemented a robust and efficient policy gradient approach. They also demonstrated that by optimising the conventional COCO measures, they might reach to better state-of-the-art outcomes. This new metric, SPIDeR, was optimised for yielding qualitatively improved results as judged by human raters. They discovered that PG-SPICE frequently produces sentences that are illegible and contain numerous repetitions. This is thus because SPICE, although measuring how closely a sentence's scene graph resembles the scene graph of the real world, is comparatively insensitive to syntactic quality. However, they achieve significantly better outcomes when they mix d SPICE with CIDEr. Thus, the rest of the paper ignores pure SPICE.

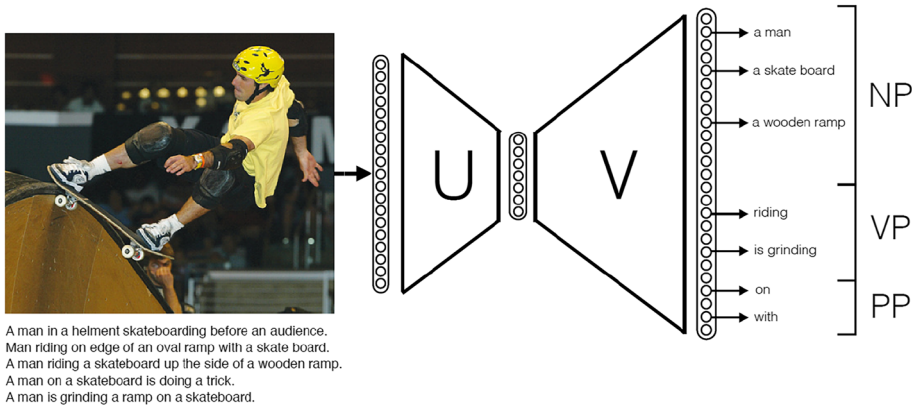
## 2.6 Image caption generator with refined descriptions based on regions

The authors of this study presented visual dependency models to capture the links between the items in an image [49]. They used a new data set of region-annotated images with visual dependency representations and best level descriptions to test this hypothesis. In both automatic and human evaluations, they discovered that image description models based on visual dependency representations beat rival models, significantly. They demonstrated that using a standard dependency parser, visual dependency representations can be persuaded automatically, and also that the descriptions generated from these representations are as excellent as those created from high level representations. In the next approach, a method was introduced by Philip Kinghorn et al. [50] which was based on the deep learning architecture for image captioning. The proposed system mainly focused on the approach that were local based and that were used to improve the existing methods which were related to different regions of people as well as different objects in an image. The images inside the system were being evaluated with the help of the datasets of IAPR TC-12 [51]. The system proposed was being evaluated with the dataset of IAPR TC-12 and then with several methods, like, NeuralTalk, Show, NTC, etc., for further comparison. It was concluded that the proposed system produced more informative and descriptive captions for the given set of images. They also aimed to use the system they proposed for the transfer learning that could be used for dealing the image descriptions for images and other various things like oil paintings, cartoons, etc.

## 2.7 Image captioning using phrase-based, template-based and retrieval-based methods

In the first approach, the authors presented a basic approach for generating descriptive sentences from a sample image [52]. They then developed a bi-linear model which created a relationship between an image representation and the sentences that were used to describe it. They proposed a simple language model that was based on the statistics of the syntax of the captions, that were useful for generating meaningful descriptions for any given test image using the phrase identification method as shown in Fig. 14.

On MS-COCO dataset, this technique was found to be significantly simpler than other state-of-the-art models, that were mainly dependent on complex RNNs. This model also produced better results when compared to previously derived methods. They also concluded that future work would include applying the model to different datasets and experimenting them with the ranking of image-sentence and also for improving the language model that was being employed.



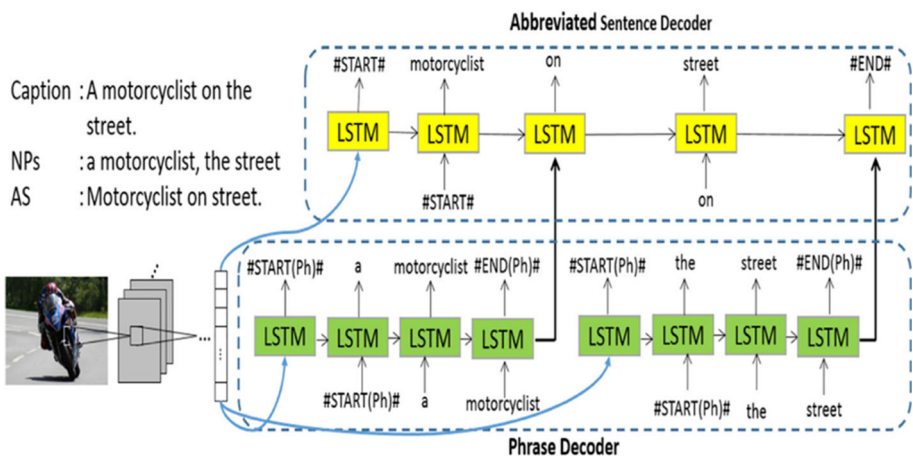
**Fig. 14** Phrase-based paradigm for image descriptions is depicted in this diagram. [52]

In the year 2018, authors Chan et al. [53] proposed a technique for image captioning that was based on phrases which used the LSTM network that was hierarchical in nature which was known as the phi-LSTM [54] as shown in Fig. 15. The authors claimed that their proposed model has shown better results for the datasets of Flickr30K images [55], MS-COCO dataset [56] as well as for Flickr8K dataset [57], when being compared to other models.

In this approach, the image captions were generated by LSTM models in which the natural processing of the language takes place initially. This processing mainly described the salient features of the objects in an image and then was used to form the complete caption for any given image.

## 2.8 Deep learning mechanism for generating visual image captions

In image captioning for using the deep neural networks, the researchers have started using the retrieval-based methods [58] that were formulated like the embedding for multi-modality techniques as well as for solving the problems of ranking. These methods were used for



**Fig. 15** Basic structure of LSTM network architecture [53]

retrieving descriptive sentences for an image and were used like the dependency- trees that were recursive in nature using the neural networks. In these technique, the software were used for extracting the phrases and sentences from a training set and then these phrases were being represented by HD vectors using different approaches like, representation through word-vector approaches. The images were further represented with the help of deep convolutional neural networks (CNNs). It was concluded that, using deep natural networks for image captioning has improved the overall performance of the system significantly. But there were still few limitations, that were needed to be solved.

In a survey done by author Shan et al. [40] on the automatic image caption generation, it was represented that there can be different ways which can be adopted for classifying the caption generating approaches for an image. This work mainly aimed at the methods that were based on the neural networks and not like the ones used earlier which were based on retrieval and templates. Here, the authors also defined few subcategories of the neural network-based methods that were mostly based on the frameworks that were specific for their use. The process of image captioning is a task of multiple modals and requires the whole machine to understand the image as well as to describe it efficiently with proper natural language [59]. Authors Feng et al. [60] have attempted to develop a model for image captioning that would be training images in the manner which is unsupervised. This model just required three things: (i) sentence corpus: for training the model for captioning (ii) image set: image datasets (iii) detector for visual concept: used for detecting visual concepts inside the model for an image. The following Fig. 16 displays the architecture of the unsupervised model for image caption generation.

The proposed model consisted of: an encoder for image, a generator of sentences and the discriminator.

- The encoder has the input of the CNN images which represented the features of the image.
- The generator used for sentences is LSTM, that decoded the image and generated the sentences naturally as well as described the content of the image.
- The discriminator was mainly used for differentiating that whether the generated sentence was obtained from the model or it was a real one.

The authors claimed this approach to be the first attempt of investigation for this problem using unsupervised methods [61]. Further, they worked upon the corpus for image description that were of large scale by using around 2 million sentences with the help of the Shutterstock

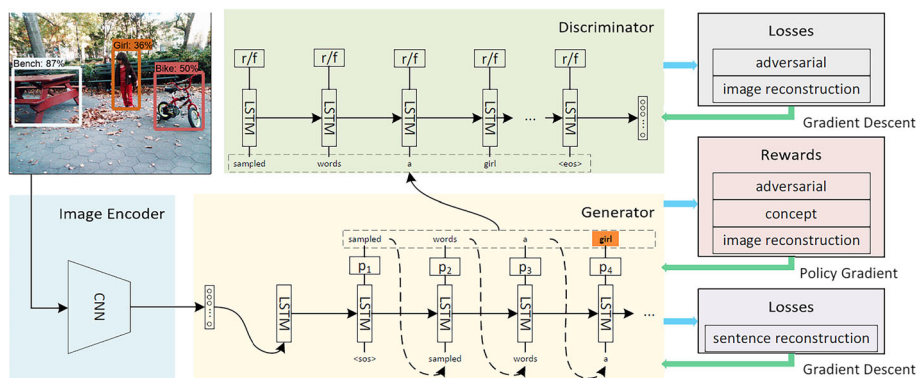


Fig. 16 Architecture for Unsupervised model of image captioning [60]

and then facilitated the method of image captioning using the unsupervised method. They finally concluded that their experiment showed better results with unsupervised data rather than using the labeled pairs of images and sentences.

Image captioning is considered an inherently challenging task for AI, because it combines challenges from both NLP and computer vision. In recent study, authors M. Stefanini [62] discussed about the various deep learning mechanisms that could be used for image captioning. To determine the most significant technical breakthroughs in architectures and training methods, they quantitatively analysed numerous pre existing state-of-the-art approaches. Additionally, other variations of the issue and there unresolved solutions were presented.

The ultimate objective of this work was to provide a resource for comprehending the body of literature and outlining potential future possibilities in a field of study where Computer Vision and NLP can function together. The following discussion traced three major possibilities for future progress in the field of image captioning.

- Challenges of the procedures and architectures: Promoting open access to various datasets will be essential for fair comparisons. In order to advance research area, it will be necessary to look at less computationally demanding alternatives due to the expanding size of pre-training models [63]. In contrast, one of the major unresolved problems in architecture is the rising conflict between early-fusion methods and the encoder-decoder paradigm [64]. On the other hand, the dominance of detection features leaves room for a number of visual encoding techniques, all of which seem to function equally well.
- Specializing and generalizing the captions: One of the major challenge is for specialising in specific fields and producing captions with various intents and styles. Furthermore, to create models that are appropriate for use in real-world applications, more research is required [65]. Improvements in image captioning variations like novel object captioning or adjustable captioning may be able to aid with this unresolved problem.
- Designing AI solutions that are trustworthy: Dataset bias and over-represented visual concepts are significant problems for any vision-and-language task because the majority of vision-and-language datasets contain common patterns and regularities. In this way, an important challenge in image captioning is the construction of acceptable and repeatable evaluation methodologies and meaningful metrics.

Lastly, since present image captioning algorithms lack trustworthy and understandable ways to pinpoint what led to a specific output, more research is required to clarify model explain-ability, concentrating on how these algorithms handle various modalities or innovative concepts.

## 2.9 Transformer-based image captioning

When used for image captioning, the transformer model takes an image as input and produces a textual description of the image's visual information that is relevant and cohesive. The transformer is highly suited for this task because it can properly model sequential data and capture long-range dependencies [66].

While transformer-based architectures have excelled in sequence modeling tasks, their potential in multi-modal contexts like image captioning has not been fully explored. Cornia et al. [67] introduces M2, a Meshed Transformer with Memory, as a novel architecture for image captioning. Experimental evaluations compare M2 with fully-attentive and recurrent models, showcasing its superior performance on the MSCOCO dataset.

Another concern was the challenge of the semantic gap between vision and language in image captioning, where traditional attention mechanisms struggle to align visual signals with

highly abstract words. To overcome these limitations, Li et al. [68] proposed an attention-based transformer framework called EnTangled Attention (ETA) Transformer'. This model achieved state-of-the-art performance on the MSCOCO image captioning dataset. Additionally, thorough comparisons with traditional techniques and sufficient ablation trials showed that this model was quite effective.

Yu et al. [69] proposed a Multi-modal Transformer (MT) model inspired by the success of Transformers in machine translation. The model captured both intra and inter modal interactions within a unified attention block, enabling complex multi-modal reasoning and accurate caption generation. The proposed approach is evaluated on the MSCOCO image captioning dataset, demonstrating superior results compared to previous state-of-the-art methods.

Transformers have proven to be successful in various NLP tasks and their capabilities can be utilized for the specific domain of medical imaging. To address one such application, Xiong et al. [70] presented a hierarchical Transformer-based model for medical imaging report generation. It generated a coherent paragraph of the medical imaging report. The model was trained using self-critical reinforcement learning. It significantly outperforms other state-of-the-art image captioning methods, achieving more than a 50% improvement in BLEU-1 scores in the medical imaging domain.

Transformer-based image captioning models have produced accurate and descriptive captions with outstanding outcomes. They have proven to be capable of capturing minute details, interpreting complicated visual scenarios and producing various and semantically significant descriptions.

## 2.10 Dense captioning

The procedure of creating thorough and useful captions for images or videos is referred to as dense captioning. Dense captioning goes beyond the scope of traditional image captioning by giving multiple when compared to the standard image methods that creates a single sentence to describe the content of an image. It mainly consists of two components: a) object detection and b) module for generating the captions. There are several applications of dense captioning for example, visual question answering, retrieval and indexing of images, accessibility to blind people, etc. [71]

A research based on dense captioning suggested that, there is a need to introduce a dense relational captioning-based model which seeks to produce several captions in response to the relationships between the objects present in a visual scene. Researchers Kim et al. [72] proposed a framework which consisted three units of recurrences that were responsible to generate sentences based on the parts-of-speech (POS). The model was coined as 'Multi-task triple-stream network (MTTSNet)'. The model deals with visual relationship detection (VRD) performed by various other methods. The model shows better results in terms of recalls and is also useful for transferring sufficient amount of data to the required algorithm. This work is open to new application-based implementations that requires combinations related to the subject, prediction and objects [73]

Further a new architecture that combines the task of localization and image description was being proposed by Johnson et al. [74]. The model was named as 'Fully Convolutional Localization Network (FCLN)'. The FCLN architecture processes an image in a single, efficient forward pass, so that it can be trained end-to-end in a single round of optimisation. The model was evaluated on Visual Genome dataset [75], that consists of 94K images and around 41 lakhs grounded captions based on the regions. The common failures that occurred in this work were repeated detection of the objects. The future work of the paper suggested

the relaxation in the number of proposed regions and to reduce the testing time of the model that could favour the trainable layer in terms of spatial suppression.

Despite numerous encouraging leads, current techniques still have two main drawbacks:

1. The majority of previous research just take visual hints from the environment into consideration while captioning, ignoring possibly significant textual context;
2. the wide range of terminology learned from the dictionary is restricted by present uneven learning methods, leading to low language acquisition and less efficiency.

In order to overcome these gaps, authors Shao et al. [76] suggested an end-to-end enhanced dense captioning architecture called Enhanced Transformer Dense Captioner (ETDC), which gathers contextual text from nearby areas and dynamically expands the vocabulary library while captioning. The paper proposes a module called, Textual Context Module (TCM) for implementation which showed better result than state-of-art methods. But the drawback of the work suggested that, without the TCM module, the model cannot fix the mistakes of the hints given in the context of a nearby foreground objects known as RoIs, that leads in the creation of incorrect words.

Although there have been some recent achievements, there are still two major limitations: 1) The majority of available approaches use an encoder-decoder framework, sequentially encoding contextual information using LSTM and 2) the large majority of state-of-art methods do not pay attention to more informative areas since they view all regions of interests (RoIs) as equally essential. To overcome these limitations, authors Shao et al. [77] presented a novel approach by introducing a groundbreaking dense image captioning framework called the Transformer-based Dense Captioner (TDC). Because of its adaptability, this framework can be easily adapted to other applications that are similar in measuring visual features. They concluded that this proposed model could be implemented on dense video captioning by changing the introduced modules according to the requirements of the users for any specific task.

### 3 Various methods and models for image captioning

As we all know, deep learning is becoming an important field of study in today's time, therefore there are few recent works that are being done for the automatic image captioning using deep learning mechanism. For defining such methods, there have been a lot of subcategories that are being developed on the basis of different frameworks. Image captioning has been a matter of great interest for researchers in the field of AI. It is also being used in various applications in the field of business, robotics, computer vision, etc. [5]. The researchers Sharma et al. [78] have presented various models for generating the image captions that were based on the mechanism of deep learning and deep neural networks. They mainly focused on different techniques that were based on RNNs and analyzed their usage for the generation of sentences. Further, they have taken few sample images and have generated captions for them as well as they compared features for different extraction methods and have analyzed the models of encoders for better accuracy and then have obtained the results that were desirable.

There are different models that different authors have optimized in various papers for the captioning of images, which are described in Fig. 17.



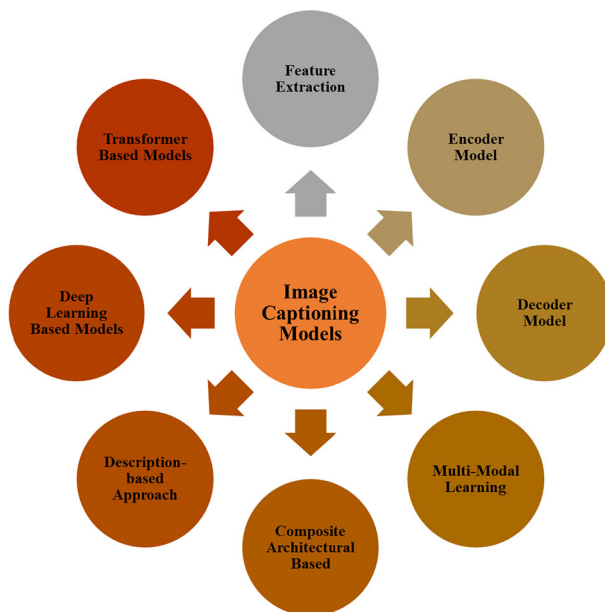


Fig. 17 Various Image Captioning Models

### 3.1 Feature extraction model

This model is mainly used for extracting variety of features from any image with the help of the training datasets. These features which are extracted from the image, acts as the input to this particular model [78].

The model employs a VGG16 architecture, as depicted in Fig. 18, to extract features from images quickly using a combination of numerous  $3 \times 3$  convolution layers and maximum pooling layers. The VGG16 network produced vectors with a size of  $1 \times 4096$ , which were used to represent the image features. To reduce over-fitting, a dropout layer with a value of 0.5 was added to the model. The ideal value was between 0.5 and 0.8 which represented

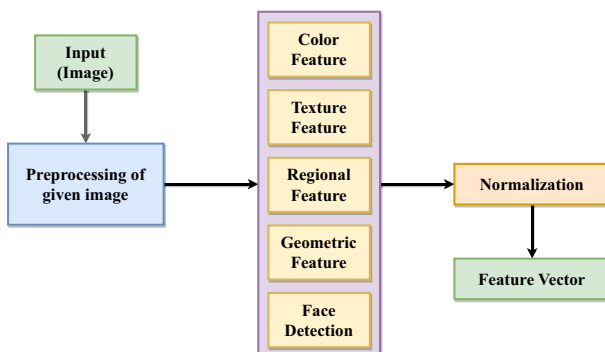


Fig. 18 Feature Extraction Model

the likelihood of the layer's outputs being dropped out. After the dropout layer a dense layer was added which effectively applied the activation function to the input and a kernel with a bias. The 'ReLU' (Rectified Linear Units) activation function was further utilized, and the output space size was set at 256. The feature extraction model then produced 256-dimensional vectors, which were then employed in the decoder model. Likewise any deep learning model such as variants of VGG, AlexNet, variants of ResNet, variants of InceptionNet and EfficientNet could be used to extract the features from images and generate the required feature descriptor.

### 3.2 Encoder model

This model was mainly used for captioning the image that has been used during the training of datasets. The encoder model generated the output in the form of a vector that is of  $1 \times 256$  size and this output further becomes the input to the decoder model [78].

In the starting the captions were being tokenized which meant that word inside the sentence were being converted into the integers. This was done so that the neural network could easily and efficiently process the data. The captions that were tokenized are then padded because of which the size of the sentence which is longest becomes equal to the length of the token. This happens because all the sentences were then being processed at the length which is equal. After this, the tokenized captions were being embedded using an embedded layer so that the fixed vectors can use them as an output. The vectors that were used then help in easing out the way of processing by using an easier way for the representation of the word inside the vector space provided. Further, the LSTM layer was considered to be the most important layer inside the encoder model. It was used for helping the model to generate the words which have highest occurrence possibilities as well as for generating sentences that were valid. The output hence generated by this LSTM layer is the final output of the encoder model layer. Whole architecture of the model is shown in the Fig. 19.

### 3.3 Decoder model

This model is the concatenation of both the encoder model as well as the feature extraction model. The output of this model are the words which were being predicted for any given image and also the sentences that are being generated at that time.

The input of this model is being obtained from the outputs of both feature extraction and the encoder model. These obtained outputs were then passed through the layer which was dense in nature and uses the activation function of 'ReLU'. After this, one more layer was added inside this model that consisted of different sizes of the vocabulary and considered them as the output space. The architecture for the decoder model is shown in Fig. 20.

The words that are then predicted are the output for the decoder model layer [78]. This model was mainly used to obtain image captions automatically by describing the surroundings as well as it helps the visually-impaired people for understanding the nearby environment in a better way.

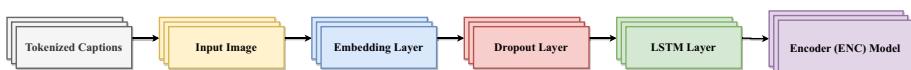
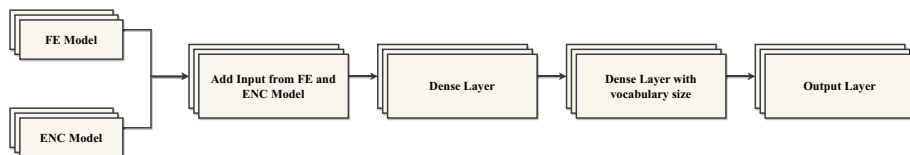


Fig. 19 Encoder Model



**Fig. 20** Decoder Model

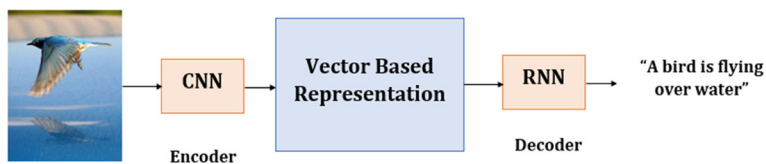
Therefore, it can be concluded that the authors have presented a model that combined the three models: firstly the model that is based on the extraction of features and uses CNN, secondly the encoder model that was used for converting the image into the representation of vectors and finally the third model, the decoder model which used the techniques of RNN and generated sentences for the given images. They have also compared different models of encoders and decoders so that they can obtain different captions with the help of different use cases.

It was further observed that the model of LSTM worked better than the model of GRU under more complexities. It can be also concluded that the performance of this model can increase when used for a larger dataset of the images after training. This work can also be of great help for the visually impaired people, because of its accuracy and it can help them to sense their surroundings easily with the help of the technology like, text-to-speech, etc. In future, the authors plan to work on different models of feature extraction so that they will be able to analyze the effect of different components of CNN inside the whole network.

With the increment in the research work in the field of translation in neural machines, the captions are being generated with the help of the frameworks of encoder-decoder [40] as shown in Fig. 21.

This framework was usually designed for translating the sentences from one language to another one. In the Encoder-Decoder framework, the encoder's neural network first seeks to encode the image into an intermediate representation, after which the decoder takes the above created representation as input and produces output in the form of sentences consisting of a sequence of words as shown in Fig. 22.

As we all know, deep learning is becoming an important field of study in today's time, therefore there are a lot of research that is being done for the automatic image captioning, which are dependent on the mechanism of deep learning. For defining such methods, there have been a lot of subcategories that are being developed on the basis of different frameworks.



**Fig. 21** Encoder-Decoder Model

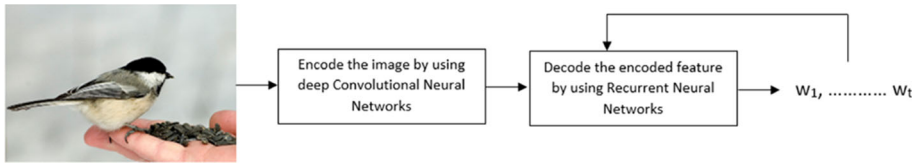


Fig. 22 Framework for encode-decoder method

### 3.4 Multi-modal learning used for image captioning

This method of image captioning was one of the most important methods which was used for generating captions for image that are purely dependent on deep learning mechanism [40].

In these methods, the extraction of the features of image was done first using the tool of feature extractor, like the deep CNN as shown in Fig. 23. After this, the obtained features were being forwarded to the model of neural language. This model further mapped the features of images inside a common space using the features of words and finally the prediction of words was done.

### 3.5 Image captioning using compositional architectures

As we have seen that all the previously mentioned methods were mostly end-to-end captioning methods, there was a need to develop a method that was useful in generating image captions using the compositional architectures [40]. Figure 24 shows the basic architecture of the compositional model being developed.

This mechanism was used for combining the individual building blocks in a sequence for the generation of the captions for an image. These methods generally used the models that were visual for the detection of the concepts that appear inside the input image. Then, these visual concepts that were detected were further sent to the model of language for generating the descriptions that could be of multiple types for an image.

### 3.6 Image captioning using description-based approaches

Till now we have seen that the methods that were being used for image captioning have the limitation to use the already existing and pre-defined words inside the dictionary and because of which they were unable to generate the image description for the concepts and datasets that were not trained before [79]. It was also possible that there may be a situation that we get

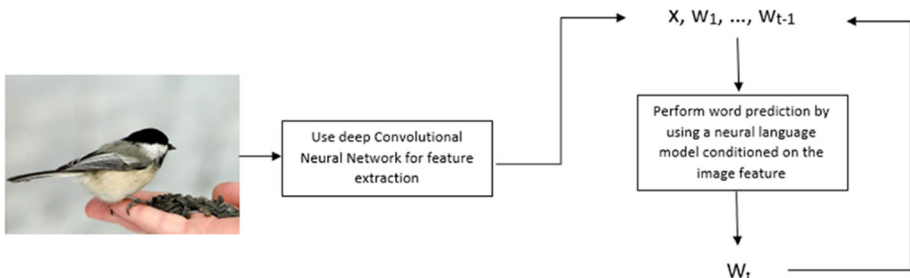
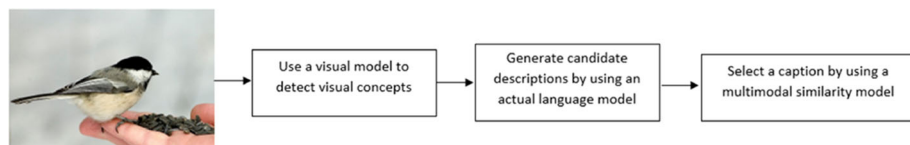


Fig. 23 Basic structure of image captioning based on multi-modal learning



**Fig. 24** Basic structure of image captioning with compositional architecture

to come across some new objects that have not been trained earlier or were not present in the pre-defined vocabulary, in such cases there was a need to define a system of image captioning that would be used for these new object and it would be generating captions for images with more efficiency. There are many approaches that were being developed for image captioning, like using the visual features along with the linguistic contexts using the semantic sentences that were hypothetical in nature [6]. For the description of the new objects that were not used inside the pairs of the training datasets, there were approaches that were developed using the method of Deep Compositional Captioner [79].

Recent captioning methods have limited ability to scale and express topics that were not demonstrated in the datasets of paired images and texts. The Novel Object Captioner (NOC), a deep visual semantic captioning model that can explain a significant number of object categories not found in existing image-caption datasets was proposed by authors S. Venugopalan et al. [80]. This approach made use of the external sources such as tagged photographs from object recognition datasets and semantic data extracted from un-annotated text. They proposed a method for minimising the combined objective that were able to learn from the multiple data sources, while also utilising distributional semantic embedding, which would allow the model to generalise and characterise novel things outside of the image-caption data-sets. They also demonstrated that their model generated captions for hundreds of item categories in the ImageNet object identification dataset that weren't encountered in MSCOCO dataset for image-caption of training data, as well as many that were only seen occasionally. Automatic and human evaluations revealed that this model beats previous work in terms of being able to characterize many more kinds of objects. One promising future direction is to develop a model that could learn from new image caption data after it had been trained. The researchers might wish to add more objects to the vocabulary and train it on a few image-caption pairings as now an initial NOC model has been developed. The main work for innovation would be to enhance the captioning model by retraining only on fresh data rather than starting from scratch.

It is essential to balance visual input with past language knowledge from pretraining in order to use a pre-trained language model (PLM) efficiently. To quickly adjust the pre-trained language, the researchers Jun Chen et. al [81] suggested VisualGPT, which used a revolutionary self-resurrecting encoder-decoder attention technique with a modest amount of image-text data from the domain. By producing sparse activations, the suggested self-resurrecting activation unit avoided unintentionally overwriting language knowledge. On MS COCO [56] and Conceptual Captions [82], VisualGPT outperforms the best baseline by up to 10.0% CIDEr and 17.9% CIDEr, respectively, when trained on 0.1% 0.5% and 1% of the corresponding training sets. Additionally, VisualGPT gets the most advanced result on the IU X-ray dataset [83] for medical report production. This experiment had a drawback, as demonstrated in the experiments, the distance between the baseline models and VisualGPT rapidly closes as the amount of in-domain training data increases with time. In COCO, the phenomena was more obvious than in Conceptual Captions, which used a wider language.

The researchers proposed that when the training data are small and do not adequately cover the vocabulary, linguistic information from pre-trained models is most helpful.

### 3.7 Recent methodologies on deep learning for image captioning

Following we have reviewed different methodologies and techniques used in captioning the image. We have reviewed methods of OSCAR, UpDown, VIVO, model using generative adversarial network (GAN) and model for Meta learning. The approaches for recent methods are discussed as follows:

#### 3.7.1 OSCAR: object-semantics aligned pre-training

This method uses faster values of R-CNN for detecting the objects inside the images and consists of mainly two types of views. The first view is called the modality view which defines the features, languages and tags present inside the image. Second, is the dictionary view which represents the semantics, tokens, spaces, etc., inside the image [84].

#### 3.7.2 UpDown attention model

This mechanism uses a language LSTM as well as a visual attention LSTM for creating captions for the image. It also uses salient features of the regions of the images and also allows those features to be used as important value as the objects of the image. This method also uses all the distributions that are conditional to generate captions completely [38].

#### 3.7.3 VIVO: visual vocabulary pretraining model

VIVO basically stands for Visual Vocabulary and uses the SOFTMAX and linear layer of an image to generate captions for images inside a transformer that consists of multiple layers. Different features of image regions are being extracted during the pre-training using the transformers consisting of tags of images. After this the captions are generated in a manner which is auto-regressive until and unless maximum length of the captions is reached [85].

#### 3.7.4 Conditional model based on GAN [86]

This model consists the combination of both CNN model as well as the RNN-LSTM model for generating the captions. These models are used for overcoming the problem of hacking using different networks and then finally generating the captions for the set of images. During all these cases, there is a need to pre-train the discriminators as well as the generators so that the caption generated are tuned finely [87].

#### 3.7.5 Meta-learning

This model firstly optimizes the tasks of reinforcement and then performs the tasks of supervision using different steps of gradients. It also consists of the “meta update” which is used for optimizing the tasks for generating the captions [46, 88]. It is therefore concluded that the ongoing research work in image captioning uses mainly methods that are based on



deep-learning. This paper discussed the recent techniques that used deep learning mechanisms for image caption generations.

### **3.8 Transformer-based model**

A neural network called a transformer model mainly follows the relationships in sequential manner, such as the words in the sentence, to predict the context and adding subsequent meaning. The transformer architecture consists of a structure based on encoders and decoders.

Luo et al. [89] introduced a novel ‘Dual-Level Collaborative Transformer (DLCT) network’ that utilizes the importance of both contextual information as well as the fine-grained details. To validate this model, they conducted extensive experiments on MS-COCO dataset [56] and achieved better CIDEr results. This approach has been extensively evaluated, and the results consistently show its superiority, surpassing previous state-of-the-art methods on both offline and online test splits. In future, the researchers intend to expand the application of the collaborative features to other multimedia domains that demands a higher comprehensive and contextual information.

## **4 Datasets**

The objective of image captioning is to describe the content of an image in words. This task falls between between computer vision and natural language processing. An encoder-decoder structure is used by most image captioning systems, in which an input image is encoded into an intermediate representation of the image’s content and then decoded into a descriptive text sequence. Models are often evaluated using a Bilingual Evaluation Understudy Score: BLEU or Consensus-based Image Description Evaluation (CIDEr) measure, with NoCaps and COCO being the most prominent benchmarks. There are few datasets that are used thoroughly for the image caption generation. The following are the descriptions of those datasets:

### **4.1 COCO (Microsoft Common Objects in Context)**

The MS COCO dataset is a large-scale dataset for object detection, segmentation, key-point detection, and captioning. The dataset consists of around 328K images [56].

### **4.2 Flickr8k dataset**

There are a total of 8092 JPEG photos in this collection, all of which are of different shapes and sizes. Around 6000 images are for training, 1000 are for testing and remaining 1000 are for development.

### **4.3 Flickr30k**

The Flickr30k dataset contains 31,000 photos gathered from Flickr, as well as 5 human-annotated reference sentences [90].

#### 4.4 Google's Conceptual Captions (GCC) datasets

It offers almost 3 million photos with captions in natural language. Conceptual Captions photographs and their raw descriptions are taken from the web, in contrast to the curated style of MS-COCO images, and hence it reflects a larger variety of styles. The raw descriptions come from the Alt-text HTML property attached to web photographs [82].

#### 4.5 Nocaps

The nocaps benchmark is made up of 166,100 captions created by humans that describe 15,100 photos from the OpenImages validation and test sets [91].

#### 4.6 SentiCaps

The SentiCap collection comprises thousands of photos with positive and negative sentiment descriptions. The authors created these emotive labels by rewriting factual descriptions. There are around 2000 emotive captions in total [27].

#### 4.7 STAIR captions

STAIR Captions is a big dataset with 820,310 captions in Japanese. This dataset can be used to create captions, retrieve multimodal data, and create images [92].

#### 4.8 SBU captions

It is a collection that enables academics to tackle the extremely difficult topic of description creation using relatively easy non-parametric methods while achieving surprisingly successful outcomes [8].

#### 4.9 SciCap

SciCap is an image captioning collection including real-world scientific figures and captions. SciCap was built with over two million photos gathered from over 290,000 papers and is provided via arXiv [93].

#### 4.10 TextCaps

There are 145,500 captions for 28,500 photos. The dataset requires spatial, semantic and visual reasoning between many text tokens and visual things such as objects in order for a model to detect texts that are related to its visual surroundings and also to decide which part of the text to copy or paraphrase [94].

#### 4.11 Google Refexp

Based on MS-COCO, a new large-scale dataset for referencing phrases has been created in the form of Google Refexp [95].

## 4.12 CC12M (Conceptual 12M)

The Conceptual 12M (CC12M) collection contains 12 million image-text pairs that were created primarily for vision and language pre-training [96].

The following Table 1 displays the overview of the datasets that are mainly used for the models of image caption generators:

## 5 Evaluation metrics

To measure the effectiveness of the models in terms of generating image captions, results of various models were tested in using the different types of evaluation metrics. The results of the system-generated captions and those described by humans must be examined. The globally

**Table 1** Overview of Data-sets in image captioning

Data-Set Name	Size	Number of captions per image
MS-COCO [56] <a href="https://cocodataset.org/">https://cocodataset.org/</a>	330K images (>200K Labeled)	5
Flickr8K [57] <a href="https://www.kaggle.com/adityajn105/flickr8k">https://www.kaggle.com/adityajn105/flickr8k</a>	8000 images	5
Flickr30K [90] <a href="https://www.kaggle.com/hsankesara/flickr-image-dataset">https://www.kaggle.com/hsankesara/flickr-image-dataset</a>	31.8K images	5
GCC [82] <a href="https://github.com/google-research-datasets/conceptual-captions">https://github.com/google-research-datasets/conceptual-captions</a>	3.3 Million annotated images	—
NoCaps [91] <a href="https://nocaps.org/">https://nocaps.org/</a>	15,100 images	11
SentiCaps [27] <a href="http://users.cecs.anu.edu.au/u4534172/data/Senticap/senticapdataset.zip">http://users.cecs.anu.edu.au/u4534172/data/Senticap/senticapdataset.zip</a>	1000+ images	2000+ emotive captions
TextCaps [94] <a href="https://paperswithcode.com/dataset/textcaps">https://paperswithcode.com/dataset/textcaps</a>	28k images	5
SBU Caption [8] <a href="https://datasets.bifrost.ai/info/1620/">https://datasets.bifrost.ai/info/1620/</a>	1 Million images	1
CC12M [96] <a href="https://github.com/google-research-datasets/conceptual-12m">https://github.com/google-research-datasets/conceptual-12m</a>	12 Million images	1
STAIR [92] <a href="https://stair-lab-cit.github.io/STAIR-captions-web/">https://stair-lab-cit.github.io/STAIR-captions-web/</a>	164062 images	5
Google Refexp [95] <a href="https://www.tensorflow.org/datasets/catalog/gref">https://www.tensorflow.org/datasets/catalog/gref</a>	30K images (MSCOCO)	
SciCap dataset [93] <a href="https://aclanthology.org/2021.findings-emnlp.277">https://aclanthology.org/2021.findings-emnlp.277</a>	2 Million images	—

accepted and available evaluation measures are given in depth in the following paragraphs for this purpose [61].

### 5.1 Bilingual evaluation understudy (BLEU 1-4) [97]

A simple technique is used in this metric, in which the generated caption is matched against a set of pre-determined sentences that are provided by the people or humans. The primary purpose of this metric is to calculate a score based on the degree of similarity between the system-generated captions and the captions that are provided by the humans. Finally, an average score is used to measure the overall quality of the system-generated captions. The system-generated captions and the predicted interpretations (i.e. reference captions provided by the humans) have most of the impact on the BLEU statistics.

### 5.2 Meteor [98]

METEOR is a unique measure that aids in the computation and analysis of system-generated language. Under a generalised unigram, the system-generated captions and human interpretations are both matched. The similarity between the two counterparts is then used to calculate a score. When there are several interpretations or possibilities, the best score from the distinctly calculated ones has to be chosen.

### 5.3 Rouge-L [99]

ROUGE is a set of metrics that calculates a score by matching pairs and sequences of words (essentially, n-grams) with human-generated summaries and reference texts. Depending on the task, several ROUGE metrics are available. ROUGE-N, ROUGE-W, ROUGE-S, ROUGE-L, and ROUGE-SU are a few of them. Each of the metrics listed above will be used to assess a separate set of features within a sentence. ROUGE-L, which is based on the Longest Common Subsequence and evaluates the score by detecting the longest co-occurring sequence of n-grams in the sentence, may be employed in any of the studies.

### 5.4 Consensus-based image description evaluation (CIDEr) [100]

CIDEr are four regularly used evaluation metrics that are used to evaluate the quality of generated sentences using the publicly available MS-COCO tools to quantitatively measure the performance of our suggested approaches.

### 5.5 Between-Set CIDEr (CIDErBtw) [101]

The proposed metric demonstrated that the human annotations for each image in the MS-COCO dataset [56] are not equally distinctive. However, previous approaches typically treated all human annotations equally during training, which may contribute to the generation of less distinctive captions. During training, the weight of each ground-truth caption based on its distinctiveness was adjusted. And then the researchers of this metrics incorporated a long-tailed weight strategy that emphasized on rare words, as they often convey more meaningful information. To promote uniqueness in the generated sentence, they also sampled

captions from the similar image set as negative examples. Therefore, it can be concluded that CIDErBtw is introduced as an evaluation metric for measuring distinctiveness, offering a quick and easily implementation for calculation method.

## 5.6 Semantic propositional image caption evaluation (SPICE) [102]

This evaluation metric is based on a consensus that is relevant to the greatest number of applicants, as the name implies. This measure will require a set of human interpretations for each image that will serve as a caption for that image. With a large number of human descriptions for a single image, this metric will compare the closeness or similarity of these references to the system-generated caption and assign a score based on the reached consensus, i.e. similarity with the majority of the human references.

These metrics were presented in order to analyse the similarity of scene graphs built from candidate and reference phrases and to find a better correlation with human judgments. All of these measurements are used to compare the consistency of n-grams in generated and reference phrases. To create appropriate comparisons with various ways of image captioning.

The following Table 2 depicts the comparative analysis of the different approaches that are being considered for the process of image caption generation. There are different methods that are taken into account and the techniques that are used to implement those methods. Further it also gives us the information regarding the datasets being used for performing image captioning in all the methods and then the accuracy of different approaches is being compared on the basis of different evaluation metrics.

The first paper in the comparison table by Jia-Yu Pan et al. [13], describes the automatic keyword generation technique that uses 10 Corel images and has around 45% of the accuracy. The next paper is regarding the use of semantic attention described by Quanzeng You et al. [33], that used both MS-COCO and Flickr30k dataset for its performance. The accuracy of this model was 0.030% better than the previously derived methods. After this, Steven J. Rennie et al. [46], described a new self-critical sequence method that has a better value of CIDEr when compared to other methods. The improved value was 0.098% better than previous one. Next approach was region-based approach that used the IAPR TC-12 dataset. This method shown 9% improved results than the approach of NeuralTalk. Talking about the other methods, the retrieval-based method, the GCN-LSTM method, the unsupervised method, X-linear attention-based technique and the CAAG mechanism showed improved CIDEr values of 1.8%, 8.6%, 5.9%, 0.8% and 8.7% when implemented on the MS COCO dataset. Other methods like convolution-LSTM, phrase-based LSTM has improved values of 12.1% and 17.5% from the baseline LSTM. Also, the technique of intention oriented with CGO constraint was considered to be 0.20% better than the LSTM-C i.e., the convolution LSTM. Method of POS-Scan was considered to display an accuracy of 28.58% when implemented on MSCOCO dataset. Bottom-up and top-down technique depicted 3 – 8% relative increment in their approach.

The offline-human feedback technique that uses human feed backs in their approach implemented their method on the conceptual captions. This technique approximately resulted in 8% accuracy in information, 6% in correctness and 1.7% in the fluency of the image captioning. Lastly, the end-to-end attention-based approach introduced by Carola S. et al. [103], depicted the larger positive impact on the performance on the model when implemented on the MS COCO dataset. According to the observations from the Table 3, the most suitable methods for image caption generation can be based on the intention-oriented with CGO constraints and the end-to-end attention based model, because they have better accuracy results as well

**Table 2** Comparative analysis of the Image-Caption Generating Approaches

Methods	Technique Used	Data-set Observed	Accuracy
Jia-Yu Pan et al. [13]	Automatic keywords generation	10 Corel image	45% accuracy
Quanzeng You et al. [33]	Using semantic attention	MSCOCO, Flickr30K	0.030% better from previous methods
Steven J. Rennie et al. [46]	Self-critical sequence	MSCOCO	0.098% improved CIDEr
Philip K. et. al [50]	Region based	IAPR TC-12	9% better from NeuralTalk
Min Yang et. al [59]	Retrieval based method	MSCOCO, Flickr-30K	1.8% improved CIDEr
J. Aneja et. al [31]	Convolution-LSTM based	MSCOCO	12.1% improved from Baseline (LSTM)
P. Anderson et. al [38]	Bottom-up and Top-down	MSCOCO	3-8% relative gains
Ting Yao et. al [39]	GCN-LSTM	COCO	8.6% gain in CIDEr
Ying H. T. et. al [53]	Phrase-based + LSTM	MS-COCO	17.5% improved from Baseline (LSTM)
Yue Zheng et. al [23]	Intention oriented	MSCOCO	0.201% with CGO constraints better from LSTM-C
Yang Feng et. al [60]	Unsupervised method	MSCOCO	5.9% improved CIDEr
Yuanen Zhou et. al [44]	POS-SCAN method	MSCOCO	28.58% accuracy
P.H. Seo et. al [22]	Off-line Human Feedback	Conceptual Captions	Information 8% (app.) Correctness 6% (app.) Fluency 1.7% (app.)
Yingwei Pan et. al [43]	X-Linear Attention	COCO	0.8% improved CIDEr
Zeliang Song et. al [47]	Context-Aware Auxiliary Guidance (CAAG) mechanism	MS COCO	7.7% , 8.7% , 2.4% improved CIDEr-D.
Carola S. et. al [103]	End-to-end attention based	MSCOCO	larger positive impact on model performance

as have larger value of positive impact factor when implemented for the generation of image captioning.

The most relevant approaches were reviewed in this survey paper. On the various datasets, we also acknowledged their success in terms of BLEU- 4, METEOR, and CIDEr scores, as well as their key characteristics in terms of visual encoding, training approaches and language modelling. In a short period of time, image captioning models have achieved amazing performance. There are different methods that have achieved an average BLEU-4 score of 25.1 for techniques using global CNN features. As well as there has been an average BLEU-4 scores of 35.3 and 39.8, with a peak of value of 41.7 for approaches that were based on attention and self-attention mechanisms. Additionally, it was noted that the addition of region-based visual encodings significantly improved both standard and embedding-based measures. The inter-object interactions inside self attention mechanisms were also responsible for further development. Notably, CIDEr and SPICE scores largely mirrored the advantages of



**Table 3** Descriptive Summarization of Image Captioning Approaches

Approach	Model Proposed	Advantages	Disadvantages
Automatic Image Captioning	Cross-Media Relevance Models [10]	1. Proposed that a tiny vocabulary of blobs can be used to characterize regions in an image. 2. Provided valuable choice for annotating and retrieving photos	1. Does not extract features of an image.
	2D MHMMs [11]	1. Based on a specific class of stochastic processes for image captioning. 2. A statistical modelling approach.	2. Cannot be implemented on a larger dataset. 1. Hard to teach a notion using just a small number of these images.
	Automatic image annotation method [12]	1. Based on a combination of picture indexing and NLP techniques. 2. Laid a solid foundation for complicated images.	2. Took more time and experience to understand more complicated topics. 1. This approach needed a lot of research in the field of NLP as well as was not so accurate.
	Corr, SvdCorr, SvdCos and Cos models [13]	1. Developed a relationship between different keywords and features of the image. 2. Around 45% accuracy for larger dataset.	1. Requirement of more accuracy as well as time consuming since it uses blobs and not whole sentences.
	Domain-specific image captioning model [15]	1. Effectively deletes terms that have errors in the extracted captions while preserving a high level of details. 2. Resulting output is generated using automatic and human evaluations.	1. Data-driven framework. 2. Only limited to a specific field of work.
	Neural captioning system: Attention model of Bahdanau with the GRU [19]	1. This approach was better when compared with the pre-existing state-of-the-art approaches.	3. More complicated to understand. This allowed the researcher to be focused on a specific portion of the image only.

Table 3 continued

Approach	Model Proposed	Advantages	Disadvantages
Human-like captioning techniques	Midge Framework [21]	2. Can generate meaningful descriptions for the images automatically. Midge's output is considered by humans to be the most natural description of images generated thus far.	To improve and refine the system to capture more linguistic phenomena.
	Reinforcement learning (RL) technique on off policy [22]	This work could be considered as the base work for working upon the techniques based on RL technique as well as for considering human feed-backs and ratings for generating image captions.	Still needs some improvement in terms of accuracy.
	Captions with guiding objects (CGO) [23]	The model provided content fluency as well as the accuracy of the descriptions.	1. Only one chosen object was guaranteed to be stated in the CGO strategy. 2.Hence was difficult to work with images of multiple objects.
	Bi-directional model [9]	This model can learn long-term interactions, to recreate visual features as new words	Not that accurate as well as requires more study in terms of RNN based approach.
CNN, RNN and LSTM based approaches	Multi-modal recurrent neural networks model (m-RNN) [24]	This model was capable of connecting different images and the sentences to accommodate complicated representations of the images and the advanced models of language	If the dictionary size is big, the embedded layer of the model has more parameters than the original m-RNN.
	SentiCap RNN model [27]	1. Generates descriptions of images with the help of sentiments.  2. 88% of the positive captions had the acceptable sentiment.	Does not work for wider range of emotions, such as pride, humiliation and wrath.
	LSTM-A (Long Short-Term Memory with Attributes) model [28]	1. Integrates different types of attributes into the CNNs) and RNNs framework.	More work needs to be done on the basis of accuracy of the model.

Table 3 continued

Approach	Model Proposed	Advantages	Disadvantages
Attention based image captioning	Multiple Instance Learning (MIL) framework [29]	<p>2. These models were then trained from beginning to end in the form of end-to-end manner.</p> <p>This work suggested a permutation-invariant aggregating operator based on neural networks that correlates to the attention process.</p>	Validation of the work is not so accurate on multiple datasets.
	Convolution image captioning [31]	1. CNN based captioning of images.	Not a sequential based approach as it does not uses any parameter of RNN based models.
	Graph Convolutional Networks (GCNs) models [32]	<p>2. Has better improvement for different metrics and was also better in terms of the performance for the baseline of LSTM+Attn.</p> <p>1. Proposes alignment between linguistic words and visual semantic units (VSUs) for image captioning.</p>	Quite a complex model to perform image captioning as VSUs were not available in all the datasets that are being used for image captioning.
	Graph Convolutional Network plus Long Short Term Memory (GCN-LSTM) architecture [39]	<p>2. The researchers used structured graphs to represent the VSUs consistently and GCNs to contextually embed them.</p> <p>1. Useful for integrating spatial and semantic relationships of the objects of an image into the encoder.</p> <p>2. Incremented the performance of CIDEr-D from 120.1% to 128.7% when tested on MS-COCO dataset.</p>	<p>Attention mechanism still needs to be studied and verified thoroughly.</p> <p>2. Performance on larger dataset is quite time consuming.</p>
	X-Linear Attention (X-LAN) block architecture [43]	1. Work was done on the bi-linear pools.	Sometimes Up-Down approach concentrates on area of image that were irrelevant and whose associated object shouldn't be formed at that instance of time

Table 3 continued

Approach	Model Proposed		Advantages	Disadvantages
Policy- Gradients based image captioning	Context-Aware Auxiliary Guidance (CAAG) model [47]		2. The performance of the CIDEr was found to be 132.0%. 1. Obtained the sentences using the contexts which are used globally. 2. Improves the performance of already existing models for reinforcement learning.	CIDEr-D value was 128.8% that was not better than obtained by X-LAN model i.e., 132.0%.
	Self-critical sequence training (SCST) model [46]		This approach was also known as the Reinforce algorithm that was used for providing the policy-gradient which was based upon the reinforcement learning (RL).	1. Methods did not produce any appreciable additional gain.
	Optimization of policy gradients using SPIDEr model [48]		1. A combination of SPICE score and CIDEr linearly. 2. Qualitatively improved results as judged by human raters.	2. Unsuccessful to train a control variable for the SCST baseline using MS-COCO. 2. Frequently produces sentences that are illegible and contain numerous repetitions. 2. Insensitive to syntactic quality.
Phrase- based Image captioning	Bi-linear model [52]		1. Created a relationship between an image representation and the sentences that were used to describe it. 2. Useful as phrase identification method.	Experiments need to be done for ranking of image-sentence.
	phi-LSTM Architecture [53]		Described the salient features of the objects in an image.	2. Improving the language model that was being employed should be done. Does not produce relevant captions for an image.

Table 3 continued

Approach	Model Proposed	Advantages	Disadvantages
Image captioning using Deep Learning Approaches	Retrieval-based methods [58]	<ol style="list-style-type: none"><li>1. Multi-modality techniques used for solving the problems of ranking</li><li>2. Were used like the dependency- trees that were recursive in nature using the neural networks.</li><li>3. Software was used for extracting the phrases and sentences.</li><li>4. Overall performance of the system was improved.</li></ol>	Still few limitations were needed to be solved in terms of accuracy as well as models based on deep convolutional neural networks.
	Unsupervised model [60]	Image captioning	Not so accurate results for the labeled pairs of images and sentences.

pre-training in vision and language. It was also observed that, one of the most important metrics for evaluating the effectiveness of image captioning systems is the CIDEr score.

We present a distribution of image captioning approaches discussed in this paper. The pie chart in Fig. 25, showcases the outcome of various techniques employed for generating image captions.

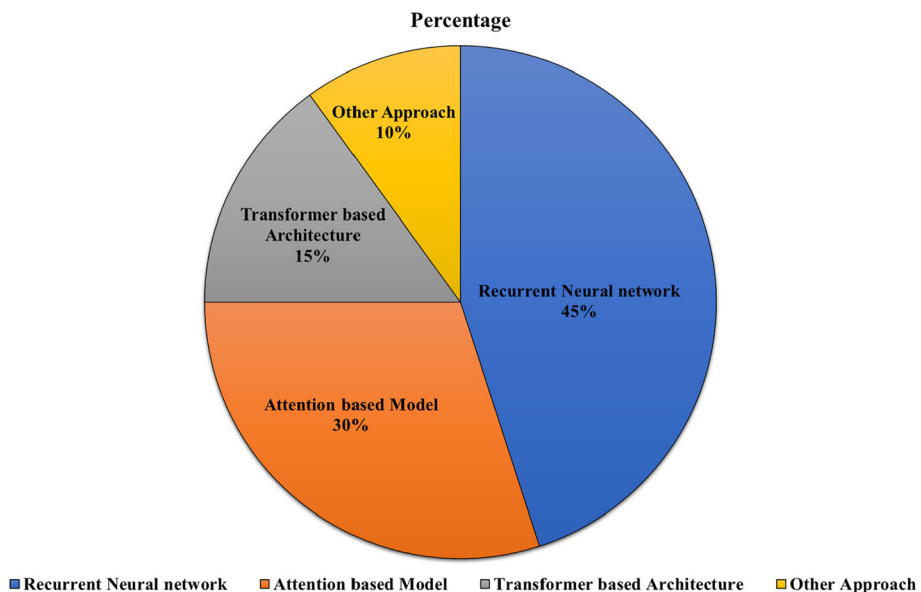
Recurrent neural networks (RNNs) are used in the bulk of publications (45%), followed by attention-based models (30%), transformer-based models (15%), and other methods (10%). An overview of the approaches utilised in image captioning research is shown in this graph, which also emphasises the dominance of recurrent neural networks as the most widely applied method.

Deep-learning methods appear to be the main focus of the present research on image captioning and with good reason. Because it mixes computer vision and natural language processing, as image captioning is an extremely hard operation that requires strong techniques which can handle its level of complexities. As demonstrated in this study, attention mechanisms, adversarial learning and deep reinforcement, all seem to be actively studied approaches in this research area. Also, a well-liked network option is the LSTM, along with faster R-CNN method.

## 6 Challenges and research gaps

As we already know that there has been a thorough development in the field of image captioning, video captioning as well as in the retrieval of the images, but there are few challenges that occur while generating the captions:

- The first challenge that occurs is to have the complete knowledge of the processing of the natural language for generating human-like captions from a machine.



**Fig. 25** Distribution of approaches used in image captioning



- Second challenge that occurs mainly is how to generate the completely grammatically image captions for any given images.
- And lastly, whenever any image is given, how to make the semantics of that image as consistent as possible and also for generating the captions that are most understandable and clear for the human beings.

In addition there is a necessity to pre-train the language we use to generate captions, therefore approaches with various procedures and architectural alterations, that can be used for large-scale work, must be developed. There is a need to solve the problem of generalization inside different domains of the datasets of the image that are been pre-trained. This basically means that the image captioning models are unable to solve all the possible scenarios of real-life problems, hence there must me some novel approaches that should be developed that could help in solving this issue.

There is a need to develop more AI solutions that have to be trustworthy as well as easy to interpret by the users. There must be a solution for tackling the bias databases also, that can provide better descriptions for the images and with more fairness. Researchers may be able to improve the system in the future, making it a valuable tool for extracting precise information from images via audio output, as well as making it a comprehensive guide for users and the broader public.

An image might have a lot of information in it. Instead of only describing a single target item, a model should be able to generate description sentences for many primary objects for images with numerous target objects. A universal image description system capable of handling many languages should be developed for corpus description languages of various languages. It is tough to assess the results of natural language generation systems. Subjective evaluation by languages is the best approach to judge the quality of any machine generated texts, but this is difficult to obtain. Hence, the evaluation indicators should be improved to make them more in line with human expert assessments in order to improve system performance. The speed of training, testing and producing words for the model should be improved in order to increase the performance of the caption generating model.

## 7 Conclusion and future direction

Image captioning is one of an important application that is being used in the field of AI. For future technologies, there is an increasing demand of generating image-captions using deep learning techniques. We have already studied about various techniques and methods that are used for creating image captions but still there are few directions of development which are needed to be done in this field.

Because this technology can be used to automate machines and achieve outcomes that are comparable to those produced by the human mind, the scope of this discipline is enormous. The goal could be to improve the system's accuracy in the future in order to make it more human-like. The system's accuracy can also be improved by incorporating datasets with a substantial amount of relevant data that will become available in the future. Eventually, the system developed must be able to train and collect domain-specific outputs, allowing it to enhance accuracy and give field-specific outcomes. Image captioning can also help in a variety of fields, such as medicine: where it can aid doctors in analysing x-rays or MRI images, human-computer interactions, traffic and surveillance: where it can assist the visually impaired people for understanding their surroundings as well as environment with the help

of images, applications of computer vision and so on. Future work could include creating a single model for positive and negative sentiment, modelling linguistic patterns (including feelings) beyond the word level and creating generative models for a broader spectrum of emotions like pride, shame and wrath.

For generating captions there can be work done on creating the models that consider human based interactions and their feed backs, so that a better design of the architecture can be obtained for enhancing the strategies of pre-training the datasets involved. Developing different strategies for pre-training the data can be done to increase the availability of the datasets by helping the inputs to get reconstructed. This can also help in improving the performance of the process of generating the captions for the images.

Captioning techniques have the potential to offer valuable insights from NLP and bridge the gap between visual and textual aspects. This can result in enhanced performance across diverse computer vision tasks and enable more engaging interactions between humans and computers. By offering more relevant and semantically rich representations of images, models of image captioning can enhance image retrieval. In visual question answering (VQA) tasks, captioning techniques can be used to produce answers based on visual input.

Visual grounding tasks, which include finding and recognising specific objects or areas in images, can be done easily with image caption generation techniques. It can also enhance the image understanding part and can provide important contextual descriptions based on the information using object detection, scene understanding and image classification, whenever needed. Captioning techniques can also improve the interaction between humans and computers by enabling more natural and intuitive communication. It can be used in various domains like robotics, assistive technologies and autonomous vehicles.

Additionally, it may be said that this study discusses current methodologies and how they are put into practise. Updown, Meta Learning, OSCAR, GAN-based and VIVO models, are some state-of-art methods. The most useful models are VIVO and OSCAR, the GAN-based models are mainly used in terms of better performance and the UpDown models have the greatest influence.

As almost all the approaches work only on the semantic identities of the images, there must be techniques that should be developed which also focuses on the naturalness as well as on the diversity of the image captions using auto-encoders, latent spaces in words, etc. And lastly, proper evaluation metrics should be developed for enhancing the captions that are generated with better understanding of the visual contents. Since accuracy, robustness, and generalisation results are far from ideal, there are still a lot of unsolved problems. Likewise, the criteria for faithfulness, naturalness, and diversity have not yet been satisfied. Lastly, since present image captioning algorithms lack trustworthy and understandable ways to pinpoint what led to a specific output, more research is required to clarify model explainability, concentrating on how these algorithms handle various modalities or innovative concepts.

**Acknowledgements** The author would like to acknowledge Department of Information Technology, Delhi Technological University, New Delhi, India for providing me necessary resources to carry out the research.

**Data Availability** All the dataset link is provided in the paper.

## Declarations

**Conflict of Interest** The authors declare that they have no conflict of interest.

## References

1. Wikipedia contributors (2022) Photo caption - Wikipedia, The Free Encyclopedia. [Online; accessed 28-February-2022]
2. Chen, F., Li, X., Tang, J., Li, S., Wang, T.: A survey on recent advances in image captioning. In: *Journal of Physics: Conference Series*, vol. 1914, p. 012053 (2021). IOP Publishing
3. Elhagry, A., Kadaoui, K.: A thorough review on recent deep learning methodologies for image captioning. *arXiv preprint arXiv:2107.13114* (2021)
4. Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., Cucchiara, R.: From show to tell: A survey on image captioning. *arXiv preprint arXiv:2107.06912* (2021)
5. Wang, H., Zhang, Y., Yu, X.: An overview of image caption generation methods. *Computational intelligence and neuroscience* 2020 (2020)
6. Mao, J., Wei, X., Yang, Y., Wang, J., Huang, Z., Yuille, A.L.: Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2533–2541 (2015)
7. by Saheel, S.: Baby talk: Understanding and generating image descriptions
8. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems* 24 (2011)
9. Chen, X., Zitnick, C.L.: Learning a recurrent visual representation for image caption generation. *arXiv preprint arXiv:1411.5654* (2014)
10. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pp. 119–126 (2003)
11. Li J, Wang JZ (2003) Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on pattern analysis and machine intelligence* 25(9):1075–1088
12. Héde, P., Moëlle, P.-A., Bourgeois, J., Joint, M., Thomas, C.: Automatic generation of natural language description for images. In: *RIAO*, pp. 306–313 (2004). Citeseer
13. Pan J-Y, Yang H-J, Duygulu P, Faloutsos C (2004) Automatic image captioning. In: 2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763), vol. 3, pp. 1987–1990. IEEE
14. Li S, Kulkarni G, Berg T, Berg A, Choi Y (2011) Composing simple image descriptions using web-scale n-grams. In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pp. 220–228
15. Mason R, Charniak E (2014) Domain-specific image captioning. In: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pp. 11–20
16. Han S-H, Choi H-J (2020) Domain-specific image caption generator with semantic ontology. In: 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 526–530. IEEE
17. Devlin J, Gupta S, Girshick R, Mitchell M, Zitnick CL (2015) Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*
18. Hessel J, Savva N, Wilber MJ (2015) Image representations and new domains in neural image captioning. *arXiv preprint arXiv:1508.02091*
19. Khan R, Islam, MS, Kanwal K, Iqbal M, Hossain M, Ye Z et al (2022) A deep neural framework for image caption generation using gru-based attention mechanism. *arXiv preprint arXiv:2203.01594*
20. Kuznetsova P, Ordonez V, Berg A, Berg T, Choi Y (2012) Collective generation of natural image descriptions. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 359–368
21. Mitchell M, Dodge J, Goyal A, Yamaguchi K, Stratos K, Han X, Mensch A, Berg A, Berg T, Daumé III, H (2012) Midge: Generating image descriptions from computer vision detections. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 747–756
22. Seo PH, Sharma P, Levinboim T, Han B, Soricut R (2020) Reinforcing an image caption generator using off-line human feedback. *Proceedings of the AAAI Conference on Artificial Intelligence* 34:2693–2700
23. Zheng Y, Li Y, Wang S (2019) Intention oriented image captions with guiding objects. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8395–8404
24. Mao J, Xu W, Yang Y, Wang J, Huang Z, Yuille A (2014) Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*
25. Chen X, Lawrence Zitnick C (2015) Mind’s eye: A recurrent visual representation for image caption generation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2422–2431

26. Karpathy A, Fei-Fei L (2015) Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137
27. Mathews A, Xie L, He X (2016) Senticap: Generating image descriptions with sentiments. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30
28. Yao T, Pan Y, Li Y, Qiu Z, Mei T (2017) Boosting image captioning with attributes. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4894–4902
29. Ilse M, Tomczak J, Welling M (2018) Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136. PMLR
30. Tanti M, Gatt A, Camilleri KP (2017) What is the role of recurrent neural networks (rnns) in an image caption generator? arXiv preprint [arXiv:1708.02043](https://arxiv.org/abs/1708.02043)
31. Aneja J, Deshpande A, Schwing AG (2018) Convolutional image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5561–5570
32. Guo L, Liu J, Tang J, Li J, Luo W, Lu H (2019) Aligning linguistic words and visual semantic units for image captioning. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 765–773
33. You Q, Jin H, Wang Z, Fang C, Luo J (2016) Image captioning with semantic attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4651–4659
34. Zhou L, Xu C, Koch P, Corso JJ (2017) Watch what you just said: Image captioning with text-conditional attention. Proceedings of the on Thematic Workshops of ACM Multimedia 2017:305–313
35. Jia X, Gavves E, Fernando B, Tuytelaars T (2015) Guiding the long-short term memory model for image caption generation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2407–2415
36. Mun J, Cho M, Han B (2017) Text-guided attention model for image captioning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31
37. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057. PMLR
38. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6077–6086
39. Yao T, Pan Y, Li Y, Mei T (2018) Exploring visual relationship for image captioning. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 684–699
40. Bai S, An S (2018) A survey on automatic image caption generation. *Neurocomputing* 311:291–304
41. Janakiraman J, Unnikrishnan K (1992) A feedback model of visual attention. In: [Proceedings 1992] IJCNN International Joint Conference on Neural Networks, vol. 3, pp. 541–546. IEEE
42. Spratlting MW, Johnson MH (2004) A feedback model of visual attention. *Journal of cognitive neuroscience* 16(2):219–237
43. Pan Y, Yao T, Li Y, Mei T (2020) X-linear attention networks for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10971–10980
44. Zhou Y, Wang M, Liu D, Hu Z, Zhang H (2020) More grounded image captioning by distilling image-text matching model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4777–4786
45. Lee K-H, Chen X, Hua G, Hu H, He X (2018) Stacked cross attention for imagetext matching. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 201–216
46. Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V (2017) Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7008–7024
47. Song Z, Zhou X, Mao Z, Tan J (2021) Image captioning with context-aware auxiliary guidance. Proceedings of the AAAI Conference on Artificial Intelligence 35:2584–2592
48. Liu S, Zhu Z, Ye N, Guadarrama S, Murphy K (2017) Improved image captioning via policy gradient optimization of spider. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 873–881
49. Elliott D, Keller F (2013) Image description using visual dependency representations. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1292–1302
50. Kinghorn P, Zhang L, Shao L (2018) A region-based image caption generator with refined descriptions. *Neurocomputing* 272:416–424
51. Escalante HJ, Hernández CA, Gonzalez JA, López-López A, Montes M, Morales EF, Sucar LE, Vil-lasenor L, Grubinger M (2010) The segmented and annotated iaprr tc-12 benchmark. *Computer vision and image understanding* 114(4):419–428

52. Lebre R, Pinheiro PO, Collobert R (2014) Simple image description generator via a linear phrase-based approach. arXiv preprint [arXiv:1412.8419](https://arxiv.org/abs/1412.8419)
53. Tan YH, Chan CS (2019) Phrase-based image caption generator with hierarchical lstm network. *Neuro-computing* 333:86–100
54. Tan YH, Chan CS (2016) Phi-lstm: a phrase-based hierarchical lstm model for image captioning. In: *Asian Conference on Computer Vision*, pp. 101–117 Springer
55. Van Miltenburg E (2016) Stereotyping and bias in the flickr30k dataset. arXiv preprint [arXiv:1605.06083](https://arxiv.org/abs/1605.06083)
56. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: *European Conference on Computer Vision*, pp. 740–755. Springer
57. Anitha Kumari K, Mouneeshwari C, Udhaya R, Jasmitha R (2019) Automated image captioning for flickr8k dataset. In: *International Conference on Artificial Intelligence, Smart Grid and Smart City Applications*, pp. 679–687. Springer
58. Socher R, Karpathy A, Le QV, Manning CD, Ng AY (2014) Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics* 2:207–218
59. Yang M, Liu J, Shen Y, Zhao Z, Chen X, Wu Q, Li C (2020) An ensemble of generation-and retrieval-based image captioning with dual generator generative adversarial network. *IEEE Transactions on Image Processing* 29:9627–9640
60. Feng Y, Ma L, Liu W, Luo J (2019) Unsupervised image captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4125–4134
61. Kumar D, Gehani S, Oza P (2020) A review of deep learning based image captioning models
62. Stefanini M, Cornia M, Baraldi L, Cascianelli S, Fiameni G, Cucchiara R (2022) From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*
63. Zhou L, Palangi H, Zhang L, Hu H, Corso J, Gao J (2020) Unified visionlanguage pre-training for image captioning and vqa. *Proceedings of the AAAI Conference on Artificial Intelligence* 34:13041–13049
64. Zhang P, Li X, Hu X, Yang J, Zhang L, Wang L, Choi Y, Gao J (2021) Vinvl: Revisiting visual representations in vision-language models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5579–5588
65. Hu X, Gan Z, Wang J, Yang Z, Liu Z, Lu Y, Wang L (2022) Scaling up visionlanguage pre-training for image captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17980–17989
66. He S, Liao W, Tavakoli HR, Yang M, Rosenhahn B, Pugeault N (2020) Image captioning through image transformer. In: *Proceedings of the Asian Conference on Computer Vision*
67. Cornia M, Stefanini M, Baraldi L, Cucchiara R (2020) Meshed-memory transformer for image captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10578–10587
68. Li G, Zhu L, Liu P, Yang Y (2019) Entangled transformer for image captioning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8928–8937
69. Yu J, Li J, Yu Z, Huang Q (2019) Multimodal transformer with multi-view visual representation for image captioning. *IEEE Trans Circ Syst Video Technol* 30(12):4467–4480
70. Xiong Y, Du B, Yan P (2019) Reinforced transformer for medical image captioning. In: *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings* 10, pp. 673–680. Springer
71. Xiao X, Wang L, Ding K, Xiang S, Pan C (2019) Dense semantic embedding network for image captioning. *Pattern Recognition* 90:285–296
72. Kim D-J, Choi J, Oh T-H, Kweon IS (2019) Dense relational captioning: Triple-stream networks for relationship-based captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6271–6280
73. Kim D-J, Oh T-H, Choi J, Kweon IS (2021) Dense relational image captioning via multi-task triple-stream networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44(11):7348–7362
74. Johnson J, Karpathy A, Fei-Fei L (2016) Densecap: Fully convolutional localization networks for dense captioning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4565–4574
75. Li L, Gan Z, Cheng Y, Liu J (2019) Relation-aware graph attention network for visual question answering. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10313–10322
76. Shao Z, Han J, Debatista K, Pang Y (2023) Textual context-aware dense captioning with diverse words. *IEEE Transactions on Multimedia*
77. Shao Z, Han J, Marnerides D, Debatista K (2022) Region-object relation-aware dense captioning via transformer. *IEEE Transactions on Neural Networks and Learning Systems*

78. Sharma G, Kalena P, Malde N, Nair A, Parkar S (2019) Visual image caption generator using deep learning. In: 2nd International Conference on Advances in Science & Technology (ICAST)
79. Hendricks LA, Venugopalan S, Rohrbach M, Mooney R, Saenko K, Darrell T (2016) Deep compositional captioning: Describing novel object categories without paired training data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-10
80. Venugopalan S, Anne Hendricks L, Rohrbach M, Mooney R, Darrell T, Saenko K (2017) Captioning images with diverse objects. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5753-5761
81. Chen J, Guo H, Yi K, Li B, Elhoseiny M (2022) Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18030-18040
82. Sharma P, Ding N, Goodman S, Soricut R (2018) Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2556-2565
83. Demner-Fushman D, Kohli MD, Rosenman MB, Shooshan SE, Rodríguez L, Antani S, Thoma GR, McDonald CJ (2016) Preparing a collection of radiology examinations for distribution and retrieval. *J Am Med Inf Assoc* 23(2):304-310
84. Li X, Yin X, Li C, Zhang P, Hu X, Zhang L, Wang L, Hu H, Dong L, Wei F et al (2020) Oscar: Object-semantics aligned pre-training for visionlanguage tasks. In: European Conference on Computer Vision, pp. 121-137. Springer
85. Hu X, Yin X, Lin K, Wang L, Zhang L, Gao J, Liu Z (2020) Vivo: Visual vocabulary pre-training for novel object captioning. *arXiv preprint arXiv:2009.13682*
86. Gonog L, Zhou Y (2019) A review: generative adversarial networks. In: 2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 505- 510. IEEE
87. Chen C, Mu S, Xiao W, Ye Z, Wu L, Ju Q (2019) Improving image captioning with conditional generative adversarial nets. *Proceedings of the AAAI Conference on Artificial Intelligence* 33:8142-8150
88. Li N, Chen Z, Liu S (2019) Meta learning for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence* 33:8626-8633
89. Luo Y, Ji J, Sun X, Cao L, Wu Y, Huang F, Lin C-W, Ji R (2021) Dual-level collaborative transformer for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence* 35:2286-2293
90. Young P, Lai A, Hodosh M, Hockenmaier J (2014) From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2:67-78
91. Agrawal H, Desai K, Wang Y, Chen X, Jain R, Johnson M, Batra D, Parikh D, Lee S, Anderson P (2019) Nocaps: Novel object captioning at scale. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8948-8957
92. Yoshikawa Y, Shigeto Y, Takeuchi A (2017) Stair captions: Constructing a largescale japanese image caption dataset. *arXiv preprint arXiv:1705.00823*
93. Hsu T-Y, Giles CL, Huang T-H (2021) Scicap: Generating captions for scientific figures. *arXiv preprint arXiv:2110.11624*
94. Sidorov O, Hu R, Rohrbach M, Singh A (2020) Textcaps: a dataset for image captioning with reading comprehension. In: European Conference on Computer Vision, pp. 742-758. Springer
95. Mao J, Huang J, Toshev A, Camburu O, Yuille AL, Murphy K (2016) Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11-20
96. Changpinyo S, Sharma P, Ding N, Soricut R (2021) Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3558-3568
97. Papineni K, Roukos S, Ward T, Zhu W-J (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311-318
98. Denkowski M, Lavie A (2014) Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 376-380
99. Lin C-Y (2004) Rouge: A package for automatic evaluation of summaries. In: *Text Summarization Branches Out*, pp. 74-81
100. Vedantam R, Lawrence Zitnick C, Parikh D (2015) Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566-4575
101. Wang J, Xu W, Wang Q, Chan AB (2022) On distinctive image captioning via comparing and reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45(2):2088-2103



102. Anderson P, Fernando B, Johnson M, Gould S (2016) Spice: Semantic propositional image caption evaluation. In: European Conference on Computer Vision, pp. 382-398. Springer
103. Sundaramoorthy C, Kelvin LZ, Sarin M, Gupta S (2021) End-to-end attentionbased image captioning. arXiv preprint [arXiv:2104.14721](https://arxiv.org/abs/2104.14721)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



# Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities

Anjum<sup>1</sup> · Rahul Katarya<sup>1</sup>

Accepted: 2 September 2023

© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE 2023

## Abstract

Information and communication technology has evolved dramatically, and now the majority of people are using internet and sharing their opinion more openly, which has led to the creation, collection and circulation of hate speech over multiple platforms. The anonymity and movability given by these social media platforms allow people to hide themselves behind a screen and spread the hate effortlessly. Online hate speech (OHS) recognition can play a vital role in stopping such activities and can thus restore the position of public platforms as the open marketplace of ideas. To study hate speech detection in social media, we surveyed the related available datasets on the web-based platform. We further analyzed approximately 200 research papers indexed in the different journals from 2010 to 2022. The papers were divided into various sections and approaches used in OHS detection, i.e., feature selection, traditional machine learning (ML) and deep learning (DL). Based on the selected 111 papers, we found that 44 articles used traditional ML and 35 used DL-based approaches. We concluded that most authors used SVM, Naive Bayes, Decision Tree in ML and CNN, LSTM in the DL approach. This survey contributes by providing a systematic approach to help researchers identify a new research direction in online hate speech.

**Keywords** Deep learning · Natural language processing (NLP) · Machine learning · Online hate speech (OHS) · Social media · Toxicity detection

## 1 Introduction

Social media sites like Facebook, WhatsApp, Instagram and Twitter are easy to use, a free source that provides advantages to people to air their voices. Now people can easily exchange their views and information from anywhere, anytime. According to a Global Digital Report [1], the world's total number of internet users in 2019 was 4.388 billion, among which 3.484 billion were online social media users. Also, according to the World Bank Report (2017), 241 million users on Facebook are Indians [1]. In Fig. 1, we summarize the total number of users on different online social media platforms with reference to the Global Social Networks [2]. Among all the social networking websites,

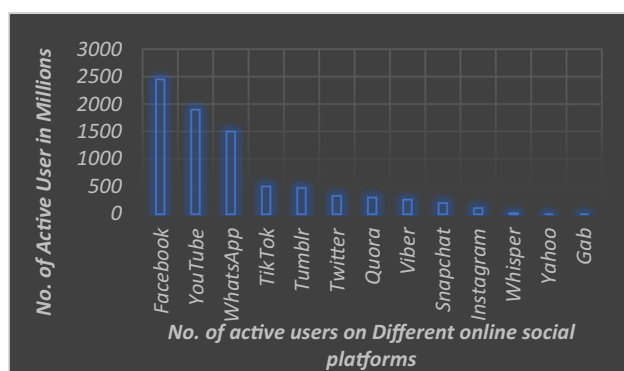
Facebook has the maximum number of users. In today's scenario, massive amounts of data are shared online every day which makes social media the most significant medium of communication. Besides these excellent features, these sites, however, have downsides as well. In the absence of meaningful restrictions or procedures, anybody can make detrimental and untrue comments in abusive or offensive language against anybody with an intention to spoil one's image and status in the community. Also, since many people around the globe during the COVID-19 pandemic were working from home and staying indoors, internet usage has risen sharply. Though many people using social media platforms can communicate virtually with their friends and relatives, there is also a spread of frustration, anger and anxiety online. These negative feelings can easily lead to hatred toward someone else. So, it becomes a huge concern for the government and for all social media sites to detect hate content before it spreads into public in general.

Also, in the present scenario, more people are using social networking websites resulting in the generation of a massive amount of data. Handling such a large amount of information is a crucial and non-trivial task since there are several target

✉ Rahul Katarya  
rahuldu@gmail.com

Anjum  
anjum\_2792@yahoo.com

<sup>1</sup> Big Data Analytics and Web Intelligence Laboratory,  
Department of Computer Science and Engineering, Delhi  
Technological University, New Delhi, India



**Fig. 1** Number of active users on social media in 2019

groups and each group is exposed to particular hate-related words that complicate the task of automated classification [1]. Example: 1. "Queers are an abomination and need to be helped to go straight to hell." 2. "Wipe out the Muslims." Both sentences are hate speech toward a particular group. The primary reason for the increase in aggressive behavior and the generation of hate speech is the anonymity provided by the social media platforms [2]. Therefore, many social websites need to develop online hate speech detection tools to control the online circulation of toxic messages [5]. The social networking websites like Twitter, Facebook, etc., are developing artificial intelligence techniques to stop the dissemination of online hate speech and toxicity on their public network. For the detection of online hate/toxicity, there already exists a web browser plugin called "Hate Speech Blocker," which flags the user that the expression could be construed as hate speech [6].

## 1.1 Problem statement

The literature in computer science on online hate speech detection concentrates on a few languages: flaming, aggressive, offensive, toxicity and cyberbullying. All of these languages are compared, with a focus on their most prevalent manifestation. To increase the quality and applicability of automated solutions, we believe that a study on one language may be useful for research on another language. We also believe that precise and ordered terminology is necessary. We referred to the broad category of research papers and weblinks and google search that includes all of these forms: "Hate Speech, toxicity, flaming, cyberbullying, aggressive". We used the term "online hate speech (OHS)" as the phrase has never been used in linguistics or computer science before, to eliminate confusion and misinterpretation. Numerous social and computer disciplines, including psychology, political science and law, have examined the manifestation, dynamics and consequences of hate speech. The literature assessment reveals that a significant amount

**Table 1** Research questions

RQ1:	"What are the primary sources of articles for OHS detection?"
RQ2:	"What is hate speech and how it originated in online social media?"
RQ3:	"What are the available OHS datasets for different languages?"
RQ4:	"What are the extracted features and most used in the traditional machine learning algorithm for OHS?"
RQ5:	"What are the trends of Traditional machine learning for classifying an online hate speech?"
RQ6:	"What are the trends of Deep learning for classifying an online hate speech?"

of study has been done on how to identify different types of hateful content. The reported publications have concentrated more on the many components of manual moderation and the difficulties that AI-based techniques should address. Fewer research articles concentrate on fully automated strategies for filtering harmful content on social networking sites. This article mainly focuses on the identification of hate speech using various artificial intelligence approaches because it offers precise definitions and solutions to the problem. Although some of the research issues (shown in Table 1) are addressed by our work, our study of the computer science literature enables us to provide additional recommendations and directions for future research.

This paper presents a survey of online hate speech identification using different Artificial Intelligence techniques. This review study looks into a number of research questions shown in Table 1 that will help us to learn about the most recent trends in online hate speech in the field of artificial intelligence. It also includes an overview of recently used machine learning and deep learning algorithms for evaluating data used by the proposed research problem.

This manuscript offers the following four contributions in greater detail:

1. Presented a framework of the online hate speech (OHS) manuscript given in Fig. 3.
2. Identified the most used traditional machine learning classifier with handcrafted features.
3. Compared different approaches of OHS detection including their advantages and disadvantages.
4. This paper provides an organized review to examine how hate speech and toxicity are incorporated into deep learning and machine learning algorithms.

This paper provides an organized review to examine how hate speech and toxicity are incorporated into deep learning and machine learning algorithms. In Sect. 1, we briefly explained the problem statement and the implication of the

study. To answer **RQ1**: we presented the OHS methodology and paper organization in Sect. 2. The previous reviews of online hate speech in the domain of AI are discussed in Sect. 3. We answer the **RQ2**, by discussing the fundamentals of hate speech, how it is originated in online social media and laws that are adopted to combat it in Sect. 4. To answer **RQ3**, we compared and discussed all the available online datasets in Sect. 5. Section 6 aims to answer **RQ4** by discussing the types of features and those that are most used in the domain of hate speech. The traditional machine learning (ML) framework, models and earlier OHS work advantages and disadvantages are discussed in Sect. 7, which aims to answer **RQ5**. Section 7 holds the answer of **RQ6**, where deep learning framework and models and types of features used in OHS detection are presented. Section 8 covers all the evaluation metrics that are used by the researchers to evaluate the results of OHS. In Sect. 9 we concluded the findings of this survey, research opportunities and future steps.

## 2 Methodology and paper organization

This section outlines the processes taken to compile the prior contributions and to gather the computer science literature that will be the subject of our analysis.

In order to answer the RQ1: "What are the primary sources of articles for OHS detection?". We tried to find all the sources for the detection and analysis of OHS. We have found approximately 200 research papers and other documents from the Google search engine, ACM Digital Library, IEEE Xplore Digital Library, Springer Link, google scholar, Science Direct, Research Gate and Wiley Online Library. We shortlisted the most relevant 136 papers suitable for this research from the above set. The complete search methodology is shown in Fig. 2 using the PRISMA diagram [7].

We consistently gathered pertinent terms by scanning cited literature in order to discover the most detailed hate speech and other related surveys. Following that, we coined the terminology "hate," "hateful," "toxic," "aggressive," "abusive," "offensive," and "damaging speeches," as well as "cyberbullying," "cyberaggression," "flaming," "harassment," "denigration," "outing," "trickery," "exclusion," "cyberstalking," "flooding," "trolling". We utilize our proposed term, "online hate speech," to refer to the combination of all these concepts in the survey's remaining questions (abbreviated to OHS). We have also taken the papers which had "hate speech," "cyberbullying", "OHS detection using deep learning", "toxicity in online social media", "OHS detection using machine learning" and "OHS detection using natural language processing" as the search keywords. The distribution of articles on online hate speech is shown in Table 2.

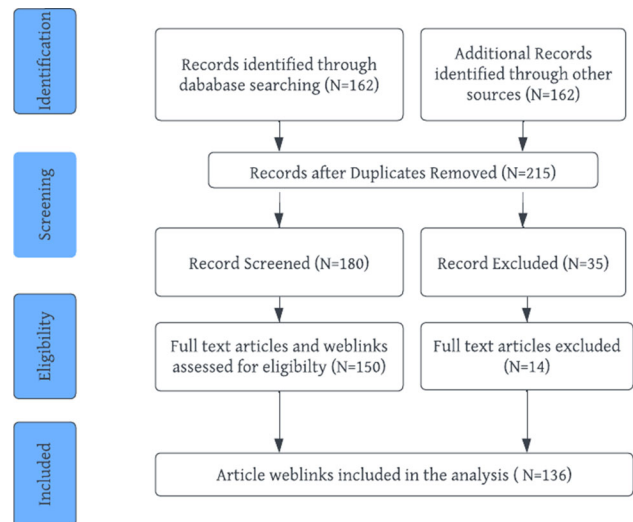


Fig. 2 Evidence synthesis for the literature survey

This review considers a broad perspective of the researchers and our analysis of toxicity detection. The flow of information in this review is presented in Fig. 3. The year-wise classification of the online hate speech article is shown in Fig. 4a, and the content-wise distribution of the referred articles is shown in Fig. 4b.

It can be inferred from Fig. 4a that hate speech has been an area of focus (computer science and engineering) from 2016 onward and is now becoming a popular research area among researchers. Also, from Fig. 4b we can see that only four survey papers have been published on Online Hate Speech as a subject of research [4, 8] in computer science.

1. **Identification** We searched all the papers on online hate speech detection tasks, such as OHS datasets, different organization contributions, proposed OHS detection models and different feature extraction techniques by including each above-mentioned keyword as the search query. All the extracted papers were taken from several journals and websites, as mentioned in Table 2.
2. **Screening** After collecting all the related information. We removed duplicates and redundant searches.
3. **Eligibility** 46 records were present from psychology, law and social science backgrounds. So, in this phase we took only 15 relevant papers from them, which were required for the problem statement. Furthermore, only the relevant search concerning the research problem has been taken. We selected total 136 articles and weblinks on which we performed this survey.

**Table 2** Amount of research contribution per source

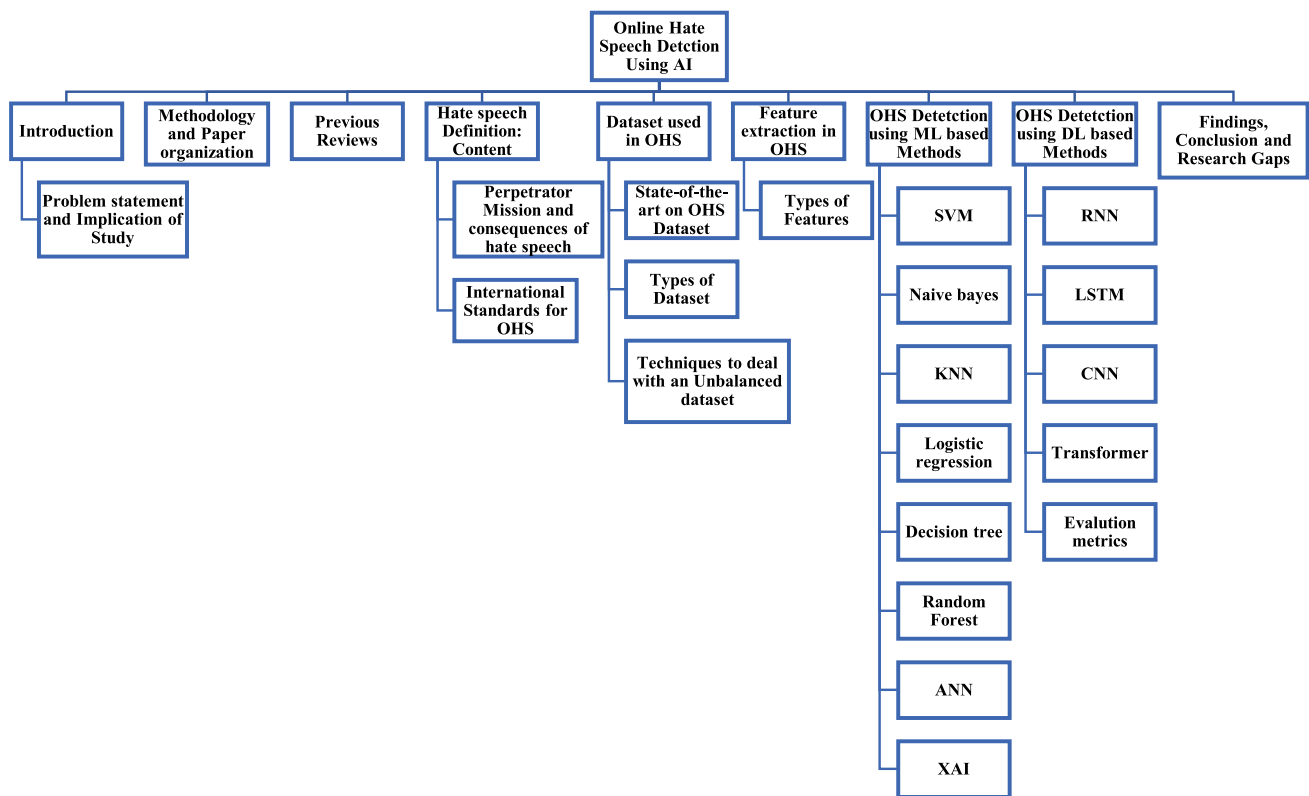
S. no	Main source of articles	Journal Name	Number of articles
1	Elsevier	Information processing and management	1
		Expert system and application	2
		Data in brief	1
		Online Social Networks and Media Journal	1
		Computing	1
		Telematics and Informatics	1
		Applied Intelligence	1
		Computers in Human Behavior Journal	1
		Aggression and Violent Behavior	1
		Interacting with Computers	1
2	Springer Link	Multimedia tools and application	2
		Crime science	1
		SN Computer Science	1
		Human-centric Computing and Information Sciences	1
		Multimedia Systems	1
		Cognitive Computation	1
3	ACM Digital Library	ACM Transactions on Internet Technology	2
		Proceedings of the ACM on Human-Computer Interaction	1
		ACM Transactions on The Web	1
		ACM Transactions on Management Information Systems	1
4	IEEE Xplore Digital Library	IEEE Access	3
		IEEE Transactions on Computational Social Systems	1
5	Other Journals	Indonesian Journal of Electrical Engineering and Computer Science	1
		Journal of Artificial Intelligence Research	1
6	Google Scholar	-	11
7	Wiley Online Library	Periodicals (Policy and Internet)	2
8	Other	Total springer and including other journals proceedings like AIS eLibrary	92
9	Weblinks	-	20
	Total journal and weblinks		136

### 3 Previous review

In recent years, few survey papers have been published in the domain of OHS using artificial intelligence techniques. The authors of the paper [2–6] present the study of OHS. These works are mainly focused on the concept of online hate speech, techniques, features and datasets published in the area of OHS. In one of the paper [2], the authors establish the basic definition of hate speech by taking into consideration different connotations and concepts this phenomenon might occur. Then the authors provide a comparative analysis of the

resources available for the research on hate speech and the pre-existing research from a computer science perspective. They deduce a lack of public datasets and metrics to establish and compare results in this field. But the author focused on the traditional machine learning approaches and did not compare different author work's limitations and advantages.

Similarly, the survey paper [4] explains the short, structured overview of hate speech using NLP. This survey compares different studies done on online hate speech from a natural language processing perspective. The review mainly focuses on comparing different types of features that are used



**Fig. 3** Systematic representation of the manuscript

to classify hate speech. It compares features like basic syntactic features, character-level features, sentiment features and more. It argues that information from features based on the text may not alone be accurate enough and researchers shall also consider multimodal and meta information features for a more accurate result and judgment. It also addresses the issue of lack of public open sources resources like datasets. The survey paper [6] presented the meta-analysis of cyberbullying papers using soft computing techniques, but the author did not present the advantages and disadvantages of the previous literature. Furthermore, the survey was limited to the cyberbullying area only. This paper [7] aims to map different themes, concepts, stakeholders and research hotspots in the field of Online Hate Research. On the basis of this analysis, the authors deduce trends and patterns in OHR like what type of countries invest in it more and change of focus in the field with time. Moreover, they try to cluster the main focal points of the research field to understand what parts are predominantly taken up by researchers, namely cyberbullying, sexual solicitation and intimate partner violence, deep learning and automation and extremist. This study is constricted to the web of science core database and shall be expanded to more databases of papers. Very few survey papers have been seen in the area of online hate speech using artificial intelligence techniques, which covered all of the information in one place.

Our survey significantly differs from earlier efforts by examining the OHS problem using AI techniques. New conceptual elements that are crucial for autonomous detection tasks are brought to light, such as integrated definitions of OHS, datasets, various kinds of features and models that affect the outcomes. It also identifies deficiencies in the way detection tasks are currently designed, notably in terms of accounting for context and individual subjectivity.

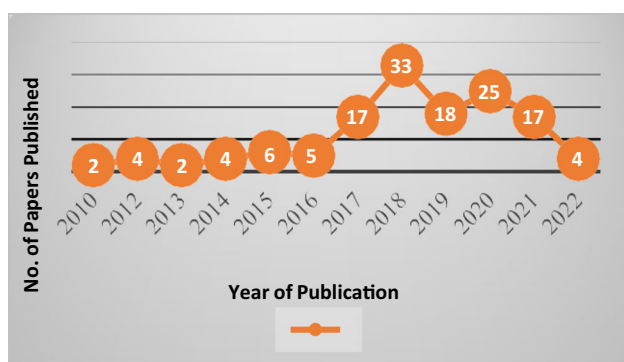
The proposed review overcomes the shortcoming of the existing surveys by providing limitations of the existing techniques and a systematic review of the online hate speech problem.

## 4 Hate speech definition: Content

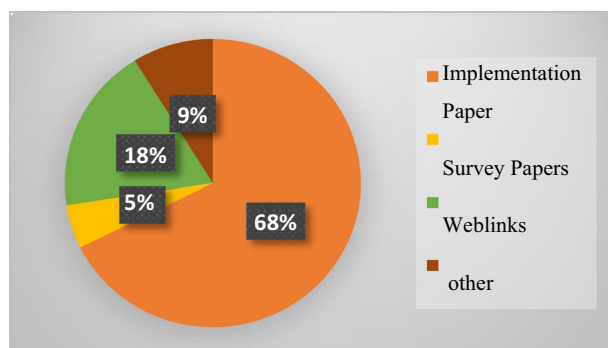
### 5 RQ2: "What is hate speech and how it originated in online social media?"

With the advent of social media and internet, we found OHS and toxicity present on every social networking website in the form of images, text and videos. With the recent advantage of mobile computing and the internet, social media provides a platform to share views and exchange information from anywhere anytime. Social media plays an essential role in the





(a)



(b)

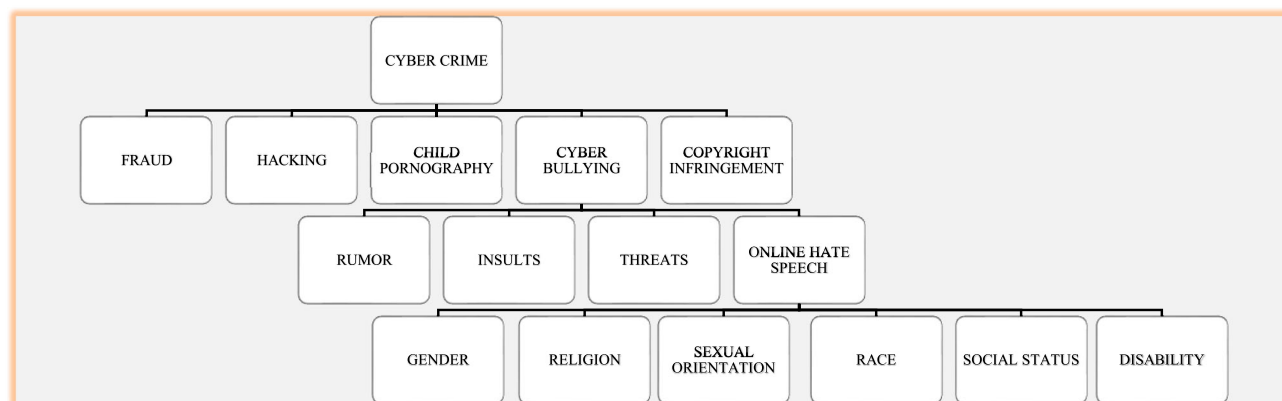
**Fig. 4** a: Year-wise classification of the referred "related papers". b: Content-wise distribution of OHS article

origin of online hate speech. On sites like Facebook, Instagram and Twitter, users can hide their identity or can bully or use toxic thoughts without being noticed. The anonymity of the user on these social platforms provides the user to conceal their identity and say and do whatever atrocious they want [9]. The origin of OHS is the class of cybercrime. So, we proposed the Taxonomy of cyber-crime to understand the origin of OHS in a more transparent way. So, we classify



**Fig. 6** Hate speech content on Twitter

the Hate problem in its various forms shown in Fig. 5. We have shown that hate speech is a part of the cybercrime and cyberbullying problem. Different authors define hate speech in different ways. The author [10] defines hate speech "The use of harsh and abusive words on online platforms to propagate immoral ideas such as communal or political polarity is called Online Hate Speech". In this paper [11] "The speech which use of offensive and hateful language to target specific characteristics of a person or a community is found to be hate speech". The author defines hate speech as when insulting and derogatory language is used to target certain people with the intend to humiliate them or condescend them [12]. Hate speech is an expression that vilifies and disparages a group of people or a person on the basis of the congregation in a social group recognized by attributes such as mental disability, race, religion, sexual orientation, or gender inequality and others [13]. Typically, hate speech promotes malevolent stereotypes and encourages savagery against people or a group. With this concept, we assume that "hate speech is any speech, which attacks an individual or a group intending to hurt or disrespect based on the identity of a person". For example, in the COVID-19 pandemic, the communal harmony between Hindus and Muslims got deteriorated due to a maligning campaign carried out on Twitter shown in Fig. 6,



**Fig. 5** Taxonomy of cyber crime

which describes the religious hate speech content and anti-social elements that exist in our society. Certain applications of detecting hate speech content are in politics, terrorism, casteism, religion. Various types of hate speech content are shown in Fig. 7. Most of the work in OHS using artificial intelligence has been done in *racism, sexism and religious* areas. Other areas of hate speech are untouched or either classified in the field of hate or non-hate category. We also surveyed five practical ways to deal with OHS in online social networking platforms like Instagram, Twitter, Facebook, that is:

- o *Report it* Hate speech violates most site's terms of service; people can report it anonymously.
- o *Block it* Block abusive users
- o *Do not share it* Forwarding any type of hate speech is wrong because offensive content can be traced back to them.
- o *Call it out* Understand how other people feel, and find ways to nurture empathy and compassion.
- o *Learn more* Hate often stems from ignorance, so learn from other's experiences.

The *consequences* of OHS can be low self-esteem, anxiety, depression, and in some cases, a victim can commit suicide. Therefore, the analysis and detection of online hate speech in social media is an area of concern.

### 5.1 Perpetrator mission and consequences of hate speech: a brief analysis

In the USA, the Federal Bureau of Investigation finds that almost all crimes, including hate speech crimes, are based on four factors [14, 15], explained in Table 3. In the manual of Ontario [16], they identify some consequences of hate crimes. Also, adolescents play an important role for being a bystander who does not participate in online hate, but they observe all things, by being a victim who suffer from online hatred and being perpetrators who do hate crimes by posting, replying and forwarding toxic content [17].

Studies show if offline aggression increases, then online hate crime also increases. There can be various consequences of online hate speech for a victim and others as well. A victim can experience anxiety, depression and in the worst case can commit suicide [18]. We categorize the repercussion of hate speech on society in Fig.8. Hate speech impacts the victim and sometimes the whole community. A person can be inflicted with psychological harm like low self-esteem. Sometimes, it also affects the target group from which the victim belongs to and makes the group or community vulnerable.

**Table 3** Perpetrator motive

Thrill-seeking	Where some people do hate crimes to make themselves happy, or they were enjoying themselves by seeing their victim sensitive to their religion, ethnicity, gender, or background
Defensive	Hate crimes arise when perpetrators are defensive about their community and to protect their society
Retaliatory	The motive of the perpetrators here is revenge
Mission offenders	Ideological reasons of the criminal such as "terrorism" where innocent people prey to perpetrators

### 5.2 International standards for OHS

We found cyberbullying has been a long-studied terminology that threatens the individual, whereas hate speech is an unpleasant language addressed to the individual or a group of people. Figure 9 shows registered cyberbullying cases along with the country of origin. Because of these high number of cases on online social media like Twitter, Facebook, etc., needs to share the responsibility to intercede and quarantine the toxic content, which is widespread on their platforms [19], hate speech on online platforms can lead to violence and is a general threat to peace and social harmony. To discourage the use of toxic language, some popular social media websites like Facebook, Twitter, Instagram and YouTube have framed new policies and guidelines [19–22]. From Fig. 9, we can conclude that India has the maximum number of reported cyberbullying cases [23] in 2019, then Brazil and the USA.

We found two bodies that make laws for the OHS: UDHR, Universal declaration of human rights, is an international body for human rights that stands for freedom of speech and expression given in article 19. To use this law appropriately, Article 29(2) established some restrictions [24]. Similarly, the European Convention on Human Rights, the International Covenant on Civil and Political Rights [25], broadens the restriction on hate speech. The government has a right and responsibility to intercede when there is a high probability of imminent harm and then take preventive policing.

### 6 Datasets used in OHS

Input data play an essential role in machine learning; therefore, it is important to use the relevant and correctly annotated data.

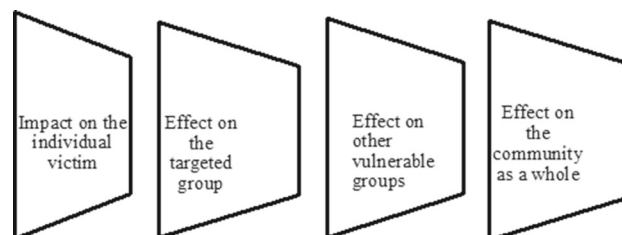
**Fig. 7** Types of hate speech on online social media



## 6.1 State of the art on OHS dataset

RQ4: "What are the available OHS datasets for different languages?"

In this study, we collected datasets from various reliable sources, and almost all the available datasets are explained in Table 4. Many researchers have used different types of hate speech datasets which are based on language, race, ethnicity, etc. Most of the datasets are available on the GitHub website. To collect data from Twitter, many researchers have used Twitter's Streaming API for analysis of hate speech as a data source, where researchers can have free access to 1% of all the data. The collected data always have metadata and are downloaded in the JSON format. Later, we need to convert it into a CSV file. The author provides an unbalanced 16 k annotated dataset collected from Twitter [8], which classify as racist, sexist and neither. In paper [9] a Facebook crawler was used to retrieve the comment from the Facebook post and five volunteered students annotated 6502 comments as no hate, strong hate or weak hate. In this [10] author used the Tumblr search APIs to get the data from Tumblr. Two–three experienced annotators performed the annotation of 2456 posts as racist, radicalized or unknown. HatEval dataset is available from collab website [11]. Whisper is an anonymous app that does not store old data, so the author [12] collected the data in real time using a distributed web crawler. Most of the authors used the kappa and Interrater agreement to capture



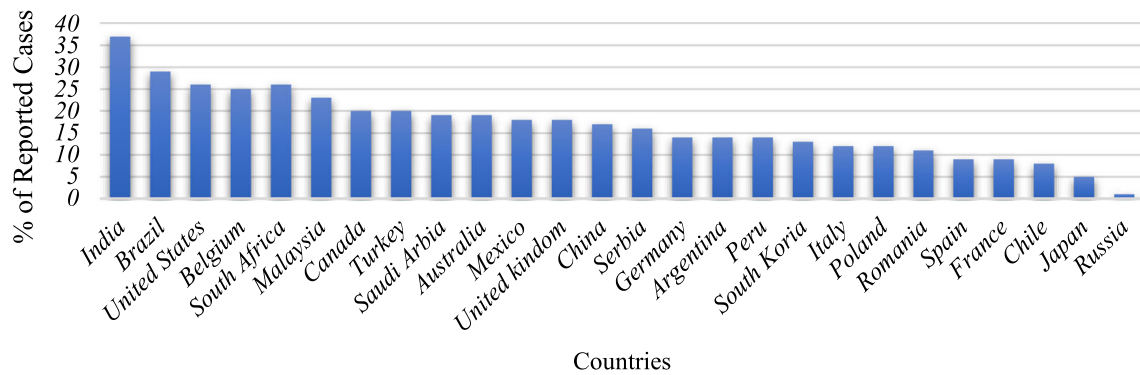
**Fig. 8** Repercussion of hate speech

the quality of the dataset. Cohen kappa is a statistical measure of inter-rater agreement of the agreement between the two raters for categorical items. Suppose we have a bunch of people and two and more raters have to find out whether each individual in this group is able to his job not. So the experts have to evaluate the group of people independently and to find out whether each individual is able to perform the job [13]. In Table 4 we have also discussed relevant details of the given datasets.

Only a few prior surveys included an in-depth examination of OHS databases. We attempted to cover practically all of the accessible datasets in our work, and scholars can also refer to the hate speech databases for extra information.<sup>1</sup>

We investigated most of the datasets that are used in the detection of OHS is imbalanced. So, to use these datasets for

<sup>1</sup> <https://hatespeechdata.com/>.



**Fig. 9** Registered cyberbullying case

classification, the researcher adopted oversampling or under-sampling techniques. In the next section, we have discussed a few techniques for the sampling purpose and their associated advantages and disadvantages.

## 6.2 Types of datasets

This section discussed the datasets used in the previous papers for OHS detection. In supervised machine learning, we deal with the labeled dataset, but in unsupervised machine learning, we deal with the unlabeled dataset. Few labeled data with a high amount of unlabeled data are used in semi-supervised learning. The labeling of data is labor-intensive and high-cost associated work. So, in this section, we explored the type of dataset which can be further classified as balanced and unbalanced dataset and we found that mostly all the given datasets in Table 4 are an unbalanced form. Therefore, for better results, different sampling techniques are taken into consideration by the authors.

- Labeled dataset and unlabeled dataset** The labeled datasets are the one in which we have both the parameters that are input and output. The author [49] collected the unlabeled multilingual data from Twitter. Thereafter keyword-based approach is used to annotate the data and then, transfer learning is used to cluster the data into hate and non-hate. To manually tag the dataset is a very time-consuming and labor-intensive job. Therefore, tools development that can automatically label the text is a very interesting area to work on. On the other hand, in the unlabeled data, we do not have the output parameter, which means that the tag is not attached to the data. We only have raw data that we fed into the classifier, which finds the hidden parameters within a dataset. The author [27] has used labeled and unlabeled dataset for training and testing the classifier, respectively. To work with an unlabeled dataset is less costly as compared to labeled dataset and is therefore used in unsupervised machine learning.

- Balanced dataset and unbalanced dataset:** When all the dataset are almost equally distributed among all the classes, then it is known as a balanced dataset. Example: suppose we have two classes as hate and non-hate, and the dataset contains 10 k tweets. Hate: 4.5 k and non-hate:5.5 k. But in a real-time scenario, we have some degree of imbalance like medical diagnosis, fraud detection, etc. If this degree of imbalance is low, then it is still called a balanced dataset. However, if this degree of imbalance is high, then this will impact the performance of the model [55]. So, when almost all the dataset belongs to one class only, it is called an imbalanced dataset. Example: From the total 10 k tweets, we have 2000 for hate and 8000 for non-hate. The author [56] used an imbalanced dataset in their work, but the classifier falsely classifies new observations to the majority class. In Sect. 2.2.1, we explore some majorly used sampling algorithms that are used in the previous work.

### 6.2.1 Techniques to deal with an unbalanced dataset

The term "class imbalance problem" in machine learning refers to categorization issues where groups of data are not separated equally. Sometimes considerable skew in the classification process of a binary or multi-class classification task is indicated by the nature of the problem in many application areas.

**Under-sampling** To mitigate the effect of an imbalanced dataset, the author [57] has used the under-sampling technique, in which random samples have been chosen from the majority class data present in training set to balance with the minority class. But this technique might discard some crucial information because it reduces the samples from the majority class, which can lead to the loss of some relevant information. Becoming more selective with the examples from the majority class that are eliminated can be an extension of under-sampling strategy. The Heuristics approaches [32] are frequently used in this process, which tries to find redundant

**Table 4** A detail list of online hate speech dataset

References	Source	Language	Size	Summary
[14] [15] [16] [17]	GitHub [ by zeerak W] <sup>a</sup>	English	6655 Tweet Dataset Distribution: NAACL_SRW_2016 None: 11,559 Racism: 1969 Sexism: 3378	The dataset is annotated by only 1 Expert and three amateur annotators The author found $k = 0.57$ Cohen kappa value for the proposed dataset
[9]	GitHub [keras-team] <sup>b</sup> And WaCky corpora <sup>c</sup>	English	17,567 Comments three classes as strong hate, weak hate, and No hate	Three different annotators are used to annotate the dataset, and the comment is collected from the Facebook pages To find the level of an agreement, the author computed the Fleiss' kappa( $k = 0.19$ ) inter-annotator agreement
[18]	GitHub [ by zeerak W] <sup>d</sup>	English	6909 Tweets Dataset Distribution: NLP + CSS_2016 Neither: 5263 Racism: 207 Sexism: 1269 Both: 52 Link: 118	Twitter API is used to collect data To find the reliability of the dataset, the author calculated the Fleiss' kappa( $k = 0.74$ )
[19] [20]	GitHub [ by T Davidson] <sup>e</sup>	English	25,000 Tweets Dataset Distribution: Hate: 1430 Offensive: 19,190 Neither: 4163	Three crowdflower workers coded the tweets manually The author used Flesch Reading Ease scores and Flesch-Kincaid Grade Level to capture the quality of each tweet The author found a 92% intercoder-agreement score
[21]	WebScope Dataset <sup>f</sup>	English	2000 Comments Two classes as clean or abusive	All the comments were collected from yahoo's new posts page The agreement rate of annotated data is 0.922, and Fleiss's Kappa is 0.843
[22]	Stormfront And crowdflower <sup>g</sup>	English	10,568 Two classes as Racist and religion	Sentence-level annotator from Stormfront and to find the hate or non-hate they used the crowdflower dataset

**Table 4** (continued)

References	Source	Language	Size	Summary
[10]	Tumblr dataset <sup>h</sup>	Arabic	5,569 comments. Dataset Distribution: Hate: 2512 Non_hate: 3057	To annotate the data, they arrange tasks on crowd flower websites only who speak Arabic Two annotators were participated to annotate the data The proposed dataset found highly imbalanced, and the inter-annotator agreement and Cohen's Kappa coefficient was 0.95
[11]	HatEval <sup>i</sup>	English and Spanish	9 k Dataset Distribution: English_train Non_hate: 5217 Hate: 3783	The data are collected from the Twitter page Hate against immigrants and women taken into consideration only
[23]	Kaggle <sup>j</sup>	English	Dataset Distribution: Neutral: 2898 Insulting: 1049	The data are collected from Twitter
[24]	TRAC(Facebook) <sup>k</sup>	English and Hindi	Non-aggressive: 69% Overtly aggressive: 16% Covertly aggressive: 16%	The data are collected using Facebook API from Facebook
[25]	Hatebase database <sup>l</sup>	All languages	N/A	Hatebase consists of all the hate words that are present in almost all languages. Example: Gender Sexual-orientation, disability class
[1]	HASOC (2019) <sup>m</sup>	Hindi, German and English	5983- Hindi 7005-English 4649- German	The dataset is classified into non-hate, offensive and hate and offensive. Also, the data were collected from Twitter and Facebook websites
[26]	Zenodo <sup>n</sup>	English, German, Spanish, French and Greek	Approx 90 k-English 62 k-German 38-Spanish 39 k-French 62 k-greek	Each dataset contains a tweet id and their annotation To access the dataset pre-request to the zenodo is required



**Table 4** (continued)

References	Source	Language	Size	Summary
[27, 28]	Github <sup>o</sup>	Arabic	Total 6000 text Tweets 2,526- hate. Which is divided as: [Jews-33% Shia-32% Christians-25% Atheists-24% Muslims-9% Sunnis-7%]	The author collected the data from Twitter The dataset is classified into hate and non-hate class and contains religious hate speech The dataset gives an accuracy of 0.79 while experimenting on GRU-based RNN with pre-trained embeddings
[29]	Github <sup>p</sup>	English, French, and Arabic tweets	English-5647 French -4014 Arabic-3353	The dataset was collected based on Directness, Hostility, Target, Group and Annotator attributes The annotator agreement scores for labeling the dataset are 0.153, 0.244, and 0.202 for English, French, and Arabic, respectively,
[30]	Github <sup>q</sup>	Arabic	Total 5,846 tweets Abusive- 1728 Normal- 3650 Hate-468	The author collected the data from Twitter, which was Group-directed and Person-directed Tweets
[31]	File <sup>r</sup>	Arabic	Total 1,100 tweets Percentage abusive: 0.59	The Twitter platform is used for the dataset collection
[31]	Github <sup>s</sup>	English	Total 33,776 posts Hate-14,614 Non-hate-19,162	The author collected the dataset from the Gap website

<sup>a</sup><https://github.com/ZeeraKW/hatespeech/><sup>b</sup><https://github.com/keras-team/keras><sup>c</sup>WaCky corpora<sup>d</sup><https://github.com/ZeeraKW/hatespeech>, [https://github.com/AkshitaJha/NLP\\_CSS\\_2017](https://github.com/AkshitaJha/NLP_CSS_2017)<sup>e</sup><https://github.com/t-davidson/hate-speech-and-offensive-language><sup>f</sup><https://webscope.sandbox.yahoo.com/?guccounter=1><sup>g</sup>Stormfront database, crowdflower, github.com/aitor-garcia-p/hate-speech-dataset, <https://data.world/crowdflower/hate-speech-identification><sup>h</sup><https://data.mendeley.com/datasets/hd3b6v659v/2>, [https://github.com/nuhaalbadii/Arabic\\_hatespeech](https://github.com/nuhaalbadii/Arabic_hatespeech)<sup>i</sup><https://competitions.codalab.org/competitions/19935><sup>j</sup><https://kaggle.com/c/detecting-insults-in-social-commentary><sup>k</sup><http://trac1-dataset.kniagra.org><sup>l</sup>[https://hatebase.org/recent\\_sightings/](https://hatebase.org/recent_sightings/)<sup>m</sup><https://hasocfire.github.io/hasoc/2019/dataset.html><sup>n</sup><https://zenodo.org/record/3520152#XcL00nUzY5k><sup>o</sup>[https://github.com/nuhaalbadii/Arabic\\_hatespeech](https://github.com/nuhaalbadii/Arabic_hatespeech)<sup>p</sup>[https://github.com/HKUST-KnowComp/MLMA\\_hate\\_speech](https://github.com/HKUST-KnowComp/MLMA_hate_speech)<sup>q</sup><https://github.com/Hala-Mulki/L-HSAB-First-Arabic-Levantine-HateSpeech-Dataset><sup>r</sup><http://alt.qcri.org/~hmubarak/offensive/TweetClassification-Summary.xlsx><sup>s</sup><https://github.com/jing-qian/A-Benchmark-Dataset-for-Learning-to-Intervene-in-Online-Hate-Speech>

examples that should be deleted or beneficial examples that should not be deleted.

**Over-sampling** Class imbalance decreases the predictive power of the classification systems. These algorithms frequently attempt to maximize classification accuracy, a parameter that benefits the dominant class. A classifier can nevertheless achieve high classification accuracy even if it cannot accurately anticipate even one instance of a minority class. In this technique, we increase the number of minority class data in the training set. Each point in the minority class tries to increase, to balance with the majority class. It is much more efficient than under-sampling because, in under sampling, we lose some amount of data. However, oversampling is prone to overfitting because we try to duplicate the example of the minority class in the training dataset [58]. In order to address the overfitting problem in oversampling for the binary classification, this research [33] offers combining the k-means clustering algorithm with SMOTE. The proposed over sampler may locate and focus on input space regions where the creation of false data is most efficient by using clustering.

Simple oversampling does not add any new information to the model because it is just duplicating the existing examples, making it vulnerable to overfitting, which can also lead to low bias and high variance results. Therefore, in order to tackle the problem of oversampling, SMOTE was introduced by the author [59] in 2002. SMOTE works on the principle of nearest neighbor and evaluates the average of it by considering the examples that are close in the feature space without duplicating the data points. By using this technique, we can create synthetic examples using skew and rotation in the feature space rather than duplicating them [60].

## 7 Feature extraction IN OHS

Detection of hate speech using machine learning is a prominent approach. The accuracy of traditional machine learning algorithms mainly depends on feature extraction. In this section, we will discuss all the handcrafted features of the machine learning algorithm. In the feature selection process, with the increase in the number of features, the threshold value increases, which in turn may decrease the accuracy of the model. Therefore, whenever we give large feature data, our model gets confused because it is learning too much information. In order to resolve this situation, we do not select all the features from the particular dataset; instead, we use some specific type of features only, which increases the accuracy of the model. In Sect. 6.1, we have discussed the types of features that play an important role in classifying the text as hate or non-hate.

## 8 RQ5: "What are the extracted features in the Traditional machine learning algorithm for OHS?"

### 8.1 Types of features

**Simple surface-level features** In order to classify the text in the different classes, these types of features are basic things to be performed first. The majority of the authors have used BOW, N-gram, char-n-gram, frequency of URL, punctuation, and capitalization in the given sentence. Bow and TF-IDF approach does not store the semantic information because there is a chance of overfitting. The author [61] used a multi-task learning approach, where different features like BOW, N-gram and sub-word embeddings were used. BOW technique [62] is employed to make the dictionary of the misogynistic and non- misogynistic. However, researchers used these features with other high-level features in order to increase the efficiency of the model [3, 56, 58, 60, 63–66]. We conclude that the performances of these features are very predictive.

**Word generalization** Most of the authors yields good classification results using Bow, meaning, in training and testing datasets, these predictive words will appear. If the dataset contains small sentences, then our model can suffer from the data sparsity. Therefore, by using the word generalization technique, this issue can be addressed. To achieve the task [63], the clusters of words are taken as additional features and brown clustering can be used to do so. If new words come up, then, based on some degree of similarity, we assign any one of the clusters to that word. In the paper [67] Word embeddings using gensim's word2vec model had been used which was found to be useful when compared to simple BOW and TF-IDF. The author [27] provides a short survey on OHS using NLP. According to the author, token-level approaches as compared to character-level approaches perform better. Word embedding and paragraph embeddings use the same concept [42, 57].

**Sentiment analysis** Hate speech itself is a negative word. If a sentence is negative in polarity, then it may be a case of hate speech or offensive speech. By taking this assumption in mind, several approaches of sentiment analysis are taken into consideration. The author has two different approaches: a multi-step approach or a single-step approach [68]. In the multistage approach, the author used sentiment analysis in the first step to finding the negative polarity, and then these negative features are further used to find the exact dictionary of the hateful words. On the other hand, in a single-step approach [39], only features are exacted using the sentiment analysis and are classified as hate or non-hate based on the polarity of the word. High variation in the degree of the polarity, such as highly negative words, also plays an important role in the classification. The SentiStrength algorithm can

also be used as a feature extraction algorithm to find the type of polarity of the document [69].

**Lexical-based approach** Generally, the hate speech consists of hate words, so the authors use the general assumption that hate speech contains hate words or negative words (like insulting words, slurs, etc.). In the lexicon approach, hateful words are taken into consideration [70]. If the word is present in the dictionary, then only classifier predict is as hate; otherwise, it will classify the sentence into the non-hate category. Hatebase1 is popularly used to find all the hate or negative words that are present in all the languages. Apart from all the list of hate words, the author focuses on the list of some specific classes of hate like racism, sexism or ethnic hate-related words. Some authors also try to identify the hate words by manual inspection tasks. In paper [71] author used the rule-based approach for subjectivity detection and to develop the hate speech classifier. For the sentiment analysis, subjectivity analysis plays a vital role, and multi-perspective question answering is used for subjective clues. They applied the bootstrapping algorithm to augment the lexicon. The author considers mostly blog and Israel-Palestinian conflict datasets for race, nationality and religion target groups. Most of the authors [8, 55, 58, 72–77] used the lexical approach in addition to other features or as some baseline features.

**Linguistic features** Sometimes, the classifier often confuses between the offensive or hate speech. Identifying the semantics of the sentences plays an essential role in hate speech detection [68] as language often comes both in the form of slurs and insults. Hence, tagging POS (part of speech) information adds some semantic information into the classifier [73]. But POS alone cannot improve the performance; therefore, some authors add more information about the data like type dependency relationship [33]. Example 1: Wipe out the Muslims. Here, the term (wipe out, Muslims) has a typed dependency between both the words. The dictionary-based approach [42] is not very useful for context-specific mapping of the offensive words. Hence, to capture the opinion, the author has used a domain-based corpus approach.

**Knowledge base** To identify the statement as hate or non-hate is not an easy task, not even by using linguistic features. Sometimes, to classify the sentence, we need some background knowledge or domain knowledge [63]. Example: "Put wig and lipstick and behave as who you really are." In the given statement, hate is directed toward a boy and comments about the sexuality (LGBT) or gender of the boy. Therefore, in order to classify, one needs to have world knowledge. The author [78] introduced some world knowledge using automated reasoning, but that requires a lot of manual coding.

**Multimodal information** Modern social media is very popular for publishing multimodal information like audio, video, images and text. The hate does not come in the form of only texts. Lots of other content is circulated every day on social media platforms. To extract the information from the images,

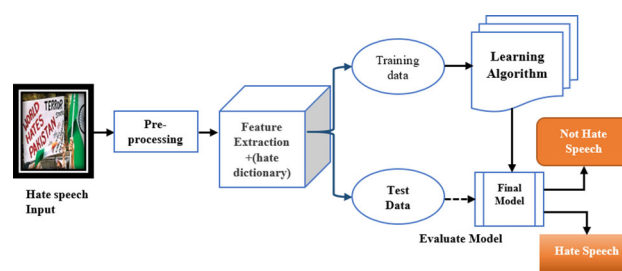


Fig. 10 Traditional framework For OHS

the author uses predictive features like user comments to find the semantics of the image. Also, the author [79] works on text and acoustic speech, but it does not yield very satisfactory results.

We analyzed all the features that are used in the various research on a different algorithm for OHS detection. Finding the best features in traditional machine learning is a very important task. Therefore, we have discussed all the features in table 5 that are used in the previous papers of OHS and we found that the most extracted features are surface-level features, linguistic features and lexicon features which outperformed the other existing features when used with the AI techniques.

## 9 OHS detection using traditional machine learning-based methods

This survey covered the various methods that have been adopted for solving the problem of OHS. The general framework of the OHS detection methodology is shown in Fig. 10. The data are first pre-processed by removing punctuation, tokenization, stopwords and stemming or lemmatization so that they can be made fit for mining and feature extraction. To train the model, features are then extracted using various techniques like Bow, TF-IDF, word embeddings, etc. After pre-processing, the features are extracted from the pre-processed data. The next step is to pass the processed data in our trained classifier which classifies them into positive or negative class.

To answer the RQ6 from Table 1 We explored the various papers of OHS using machine learning to deal with online hate speech.

### 9.1 Support vector machine

The support vector machine (SVM) was invented back in the '90 s by Vladimir Vapnik. SVM makes use of kernel trick to model nonlinear decision boundaries. It draws a decision boundary near the extreme points in the dataset. Therefore, SVM algorithm is essentially a frontier that best segregates the two classes. The author [56] has used SVM to

**Table 5** List of handcrafted features used for the detection of OHS

S. no	Feature class	References
1	<i>Surface-level feature</i> A1: Bag-of-word A2: Negation A3: unigram A4: n-gram A5: Frequency of URL mention A6: Token Length and Capitalization A7: Non- English Words	[77, 80, 81, 82, 83, 84, 85, 79, 56, 65, 78, 86, 87, 3, 63, 88, 89, 55, 90–92]
2	<i>Word-generalization</i> B1: Set of words (Clustering) B2: Word Embeddings B3: TF-IDF	[3, 42, 56, 58–60, 63, 65, 66]
3	<i>Sentiment analysis</i> S1: Positive and Negative polarity S2: Neutral words	[58, 93, 57, 82, 65, 73]
4	<i>Lexical resources</i> L1: General Hate-Related terms L2: Contextual Information	[94, 33, 55, 65, 76, 91, 94, 80, 95, 32, 65, 79, 73, 55, 96, 97, 32, 66]
5	<i>Linguistic feature</i> F1: n-gram + POS information F2: Dependency Relationships F3: Syntactic Feature and Semantic feature	[44, 40, 98, 99, 100, 65, 101, 66]
6	<i>Knowledge-based feature</i> K1: Heteronormative Context	[99, 73, 102, 99]
7	<i>Meta- information</i> M1: Background information about the user of the Post M2: No of Post by User M3: No of reply by user M4: Location M5: Correlation between the number of post and hate speech	[41, 103, 72, 73]
8	<i>Multimodal information</i> C1: Images C2: Audio C3: Video and Audio Content	[104]

find the racist text using different kernel functions on the Bow, bigrams and pos in order to find the best effective technique. The highest accuracy was achieved on Bow using the polynomial function, but Pos performed worse than bow and bigram. It has been observed [73] that the SVM performed best on the surface-level features. On the binary classification [73], the SVM classifier gives the highest results in terms of accuracy. In paper [105] author collected data from yahoo newsgroup posts and the American Jewish Congress. A template-based strategy is used to generate features from the corpus. The author took the problem as word-sense disambiguation and used SVM light classifier as a linear kernel function. The proposed result using this classifier on the dataset was not accurate. Also, the bi-gram and tri-gram degraded the performance of the classifier.

Furthermore, long linguistic pattern was not detected and also resulted in a low recall and precision value. This paper [102] presented the annotation framework for hate speech of tweets that were collected during the Kenyan election. They developed the framework for the extracted text and employed bootstrapping and n-gram technique to obtain the hateful tweets from the 394 k collected data. For the reliability of annotated tweets, the author used Krippendorff's alpha. The same concept described in the duplex theory of hate (i.e., passion, distance and commitment feature for the hate speech framework) was used in the paper [26]. Out of 394 k tweets, 94% of tweets labeled ethnic. The authenticity of the data

are not cared about, i.e., fake news and propaganda. Also, this framework is applicable only for short messages. SVM is one of the major adopted techniques by the researchers [3, 42, 65].

## 9.2 Naive Bayes

It is a supervised learning algorithm that is used for binary and multiclass classification problems. It is based on the Bayes theorem given by Thomas Bayes: the algorithm makes naïve assumption that the features are independent of each other, which makes the algorithm simple and effective.

$$P(A|B) = (P(B|A)P(A))/P(B) \quad (1)$$

$P(A|B)$ : The probability of finding the event A, when event B is true.

$P(A)$ : Prior probability that is the probability of an event before event B.

$P(B)$ : Prior probability that is the probability of an event before event A.

$P(B|A)$ : The probability of finding the event B, when event A is true.

In the detection of hate speech, the author [58] used naïve Bayes by extracting the surface-level features and lexicon features and found that the voting classifier gives the best

results compare to the lexicon-based approach for the classification. The author [3] took at least three annotators to annotate the hate words and compared the results also Standard Pre-processing TF-IDF and n-gram is used after that Naïve Bayes gives the same accuracy as other classifiers. By using the hard ensemble, the author [8] achieved the highest accuracy of 78.3% with naïve compared to other classifiers on the unbalanced dataset.

### 9.3 k-nearest neighbor

It is one of the simplest and most used classification algorithms. This algorithm is used when data points are separated into several classes to predict the classification of a new sample point. KNN captures the idea of similarity. It is used to solve nonlinear classified data points means if the data points are distributed in a nonlinear manner, where we cannot just draw a straight line, there we can use KNN. In order to find the similarity between the data points, Euclidean distance, Manhattan distance is calculated. Then an object is classified based on the number of votes of its neighbors with the object being assigned to the class most common among its nearest neighbors. To find the prominent pages on Facebook, the author [58] used Betweenness Centrality. Very few works have been identified in the field of hate speech detection.

### 9.4 Logistic regression

Logistic regression (LR) is used to solve binary classification and multiclass classification problems, i.e., output  $y \in \{0,1\}$ . Regression estimates the relationship between the dependent and independent variables. Hence, LR is most widely used when the dependent variable or the output is in binary format or categorical format. The author [42] implemented a logistic regression with the surface-level features, which gives comparable results. We did not find much work on word generalization and knowledge-based features in logistic regression. Furthermore, very few works have been seen by considering different features set to classify the sentences shown in Table 6.

### 9.5 Decision tree

Decision tree (DT) is a flow chart-like structure in which each internal node represented a "test" on an attribute, and each branch represents the outcome of the test, and each leaf node represents a class label. DT is used to map nonlinear relationship, means if data are not easily separable, then we draw or split it into different classes. DT is used by the authors [42], and surface-level features were the first choice of the research to use in the classification process.

### 9.6 Random forest

It creates DT on data samples and then gets the predictions from each of them and finally selects the best solution by means of voting. It is an ensemble method that is better than a single DT because it reduces the overfitting by averaging the result. The author [70] used the ensemble of DT to work on the video platform to find the hatred on the multimodal data. The author finds the maximum accuracy of 0.94% with a weighted-vote ensemble. Author [106] detects the hateful content on Twitter and Whisper. As whisper is an anonymous mobile application, they collected nearly one-year data from the whisper app and 1% random sample from Twitter, which is available to all the users. They present the computational method to detect hate speech in which they divide the sentence into four parts, i.e., I, Intensity, user intent and hate target. Also, there is a possibility of biases as the collected data are from the online social network.

### 9.7 Artificial neural networks

It is an interconnection of assembly of nodes to form structures using a directed link. A simple artificial neural network (ANN) consists of only one hidden layer. Perceptron is a simple neural network which can be further classified as a single layer and multilayer. Multilayer perceptron consists of hidden layers and hidden networks. The author [60] fed extracted features into the simple ANN classifier and followed a genetic-based approach to detect the hate speech in the Albanian language.

### 9.8 Explainable artificial intelligence

Explainable artificial intelligence (XAI) is technology which decodes the reason behind the neural networks and presents it in form understandable by humans [107]. With neural networks becoming more and more complex with many more parameters and feature engineering becoming a thing of the past, making deep learning models justifiable is the need of the hour. XAI has already gained significance in the domain of computer vision with visualizations like class activation maps becoming more and more popular. Class Activation maps are made by overlaying the features of a layer in DNN on the image to classified signifying the importance a model places on a particular region or pixel. Class activation maps help data scientist design a model which uses relevant features to make a decision, making the model more reliable. The adoption of XAI has been low though recently sudden interest has been seen. The author [107] released a benchmark dataset in which each tweet has a class-label (hate, offensive, normal), a target community and the rationale behind its class-labels. The author further shows that it is not necessary that the models performing best according to traditional

**Table 6** ML classifier used with general features of OHS

ML classifier	Feature							
	Surface-level feature	Word-generalization	Sentiment analysis	Lexical resources	Linguistic feature	Knowledge-based feature	Meta-information	Multimodal information
SVM	[42, 3, 40 [98, 100, 103, 82, 81, 50, 4 [77, 3, 74, 4]	[32, 65]	[82, 65]	[95, 32, 65]	[44, 40] [98, 99] [100, 65] [101]	[102, 99]	[102, 103]	N/A
NB	[42, 40, 108, 103, 109, 81] [77, 92, 104, 3, 74] [79]	[99, 65] [79, 56]	[65]	[95, 65] [79]	[40, 99] [65, 33]	[102, 99]	[102, 103] [72]	[104]
RF	[42, 40, 100, 82, 77, 3]	[65]	[82, 65]	[65]	[40, 100] [65, 33] [79]	[102]	[102, 13] [72]	N/A
KNN	[40, 74]	[58]	N/A	N/A	[40]	N/A	[72]	N/A
DT	[42, 40, 77, 74, 79]	[79]	N/A	[94]	[40, 94, 33] [79]	[94]	[72]	N/A
Logistic regression	[39, 40, 100, 31, 110, 82, 50, 77, 74]	N/A	[82]	[39]	[100, 111]	N/A	[31, 72]	N/A
Adaptive boosting	[82]	N/A	N/A	N/A	N/A	N/A	N/A	N/A
NLP	[93, 104]	[64]	N/A	N/A	N/A	N/A	N/A	[104]
J48	[3, 73, 55]	[73, 55]	[73]	[73, 55]	[99, 73] [55]	[99, 73]	[72, 73]	[73]

\*\*N/A: Authors did not perform the task



**Table 7** ML algorithms used in the research papers

Algorithms	No of frequencies used in the paper
SVM	26
Naïve Bayes	21
Random forest	13
Decision tree	12
Logistic regression	10
ANN	3
KNN	1
XAI	1

metrics such as accuracy, macro-F1 score and AUROC score will necessarily perform well on explainability metrics such as Plausibility, comprehensiveness and sufficiency.

Based on the total 95 articles in OHS, approximately 40 research papers used the traditional machine learning approach. SVM, Naive Bayes and decision tree are the most common approaches used in the papers of OHS in computer science background as shown in Table 7.

As part of the practical work that has been done, hate speech is being explored in relation to other pertinent concepts, including social media and machine learning. Machine learning techniques are being used to classify hate speech and automatically identify it.

According to the aforementioned literature, 136 research publications provided a variety of strategies for locating online hate speech in social networks. Unsupervised machine learning was discovered to be a relatively recent subject of study. Some researchers combined various techniques, such as sentiment analysis, emotional analysis and text mining, to effectively categorize the hate texts. As a result, each study has a unique perspective and understanding of online hate speech detection. In a nutshell, we have highlighted the following common flaws and limitation with current approaches.

1. From the study, it has been observed that the existing research covers mostly lexicon (simple keywords)-based hate speech analysis. As a result, the outcome of those models would not be able detect semantic of the text.
2. Facebook, Twitter and other social media platforms including the research papers that we have studied do not have a real-time hate speech detecting system and the corrective measures are taken only after the expression is posted online. So, real-time detection system can be made so that the corrective measures can be taken on time.
3. The majority of the methods are quite complex, including deep logical structures, complex equations, derivatives and formulas. Algorithms also required an excessive

amount of computational time to execute. Straightforward and less complex model should be implemented so that the computational cost can be reduced.

4. Most of the researchers worked on highly imbalanced dataset, which would result in an inaccurate result. So, to deal with the class imbalance problem authors should adopt some strategies some of them are already listed in Sect. 5.2.1.
5. We also invested that majority of the study only used supervised learning and none of the author explored the area of unsupervised ML.

In Table 8, we have shown a comparison related to various traditional machine learning approaches and their associated advantages and disadvantages.

Considering the fact that online hate speech can occur in different formats, where the word, sentence, semantic and pragmatic knowledge of the language are significant. So, from the study, it has been observed that ngram and word embeddings can be a suitable approach to achieve better accuracy with machine learning models. Furthermore, LR and SVM often performed well when experimented with different approaches. We can see in Table 6 that surface-level features and linguistic features are most used with different traditional machine learning classifiers. Very little work has been done using other handcrafted features except for surface level and lexical resource. Moreover, some of the areas are not even explored. (Marked as 'NA', Table 6). In the OHS, there is further scope to work on KNN, Adaptive Boosting classifier, "cleaning and stemming" and annotation of the data using automatic machine learning tools.

## 10 OHS detection using traditional deep learning-based methods

Traditional machine learning and deep learning, both offer ways to train models and classify data. In traditional machine learning, we manually extract features, but in deep learning, we skip the manual step of extracting features; instead, we put data directly into the deep learning algorithm like a convolutional neural network (CNN), which then further predicts the object. Therefore, deep learning is a subtype of machine learning which deals directly with data (like images) and is often more complicated. In this section, we have covered the various methods of deep learning that have been adopted for solving the problem of OHS. Figure 11 shows how deep learning model classifies the text as hate speech or not hate speech by taking some inputs. A deep neural network is a type of artificial neural network which has more than one hidden layer that helps to extract higher-level features from the dataset. At each level, the input is slightly transformed, and it gives more details of the data. Deep learning behaves

**Table 8** Traditional frameworks of OHS

References	Approach	Language	Dataset	Methodology	Merits	Limitation
Raufi et al. [60]	ANN	Albanian	3620 words from Albanian forums	The author used Standard Pre-processing, feature extraction using a Bag of Words Extracted Featured Fed into ANN and Classified, and then new words are added to the hate vocabulary	The highest accuracy achieved is 94%, with a 60–30 spilled	In the Long Run, many word features will become irrelevant. Their current system is developed on "per word"-based detection, where deeper language constructs are not in their scope
Martins et al. [65]	RF, NB, and SVM	English	Davidson and Warningsley Total tweets are 24,782 Hate-1430 Offensive-19190 non-hate-4162	The author used lexicon-based and machine learning approaches to predict hate speech contained in a text, using an emotional approach through sentiment analysis	Finds the best accuracy with SVM compared to NB and RF i.e., 80.56%	The author gives emphasis on emotional features only Uses fixed vocabulary found on hatebase, Semantic features are not considered
Sharma et al. [78]	Machine Learning and NLP	English	Dataset available on Kaggle 2235 number of samples from various sites	Standard Pre-processing such as stop word removal stemming, etc., is done followed by labeling and adding the time comment was made	Real-time tweets are extracted from multiple online sites and created a new dataset	Prepares only data but does not build a classifier
Pelzer et al. [64]	NLP and Automated reasoning	Swedish	Collected 17,176 comments from open forums i.e., Avpixlat and Samha llsnytt	The author compared their developed automatic hate speech method with the manual analysis Build a (NLP + Automated reasoning) approach The author took six politicians, three males, and three females	NLP + AR approach is more easily adapted to other languages by modifying the underlying dependency recognition rules	NLP + AR technique finds very small hateful comments compared to manual inspection Hatecategory dictionaries do not capture all hate expressions
Davidson et al. [42]	Logistic regression, SVM and Sentiment Lexicon	English	Collected 24,782 annotated Tweets named as Davidson and Warningsley dataset	Standard Pre-processing, unigram, bigram, trigram features are extracted with POS tagging To classified a text as hate author used NLP, TF-IDF is used to find the most relevant word, and BOW is used to find the most frequent word	Overall F1 score of 90 is achieved with SVM and LR The dataset is highly skewed, not able to classify hate speech with high accuracy, will not be able to handle unseen vocabularies	40% of hate speech is misclassified: the precision and recall scores for the hate class are 0.44 and 0.61, respectively
Diwani et al. [3]	SVM, J48, Naïve Bayes, Random Forest, Random Tree	Turkish	Collected 1288 Tweets from Twitter. Where 159 was classified as hate and, 1129 classified as non-hate	The author took at least three annotators to annotate the hate words and compared the results They used Standard Pre-processing TF-IDF, and n-gram is used after that	Accuracy is in the range of 60 on almost all models	The author has adopted a complete lexical approach. This model will fail if new vocabulary is observed in the data

**Table 8** (continued)

References	Approach	Language	Dataset	Methodology	Merits	Limitation
Rodriguez et al.[58]	Sentiment Analysis (VADER) Emotional Analysis (JAMIN), K-means clustering	English	Collected 1000 comments from each page from the Facebook using FB graph API	The proposed framework intends to identify prominent pages in social media where potential hate speech promoters may exist To find the prominent pages on Facebook, they used Betweenness Centrality	The author proposes a new way of dealing with hate speech, rather than building a classifier that classifies each tweet into hate and non-hate	The method is not completely automated since a man will be required to inspect the words near each centroid
Watanabe et al.[57]	Unigram and pattern classification, J48graft	English	Three different datasets from two from Crowdfunder and one from GitHub. Divide dataset into three categories as hate, offensive, and clean	The author proposed an approach that collects words and expression in a pragmatic way and uses them with patterns, along with other sentiment-based features to detect hate speech	The proposed approach reaches an accuracy equal to 87.4% for the binary classification of tweets into offensive and non-offensive and an accuracy equal to 78.4% for the ternary classification of tweets into, hateful, offensive and clean	Richer dictionary of hate speech patterns can be used for the better classification
Greevy and Smeaton [56]	SVM	English	3 million words collected from Yahoo The phrases and a list of words have collected from the Yahoo dictionary	Applied SVM model on BOW, Bigram and POS feature Experiment conducted using the default linear, polynomial, radial basis function and sigmoid tanh as kernel functions The author used different kernel functions on Bow, bigrams and pos in order to find the best effective technique	Polynomial proved to be the most effective kernel function for both BOW and POS The highest accuracy was achieved on Bow using the polynomial function	The author used the linear kernel function and sigmoid tanh for BOW and POS. But they are computationally expensive Pos performed worse than Bow and bigram Better results can be achieved if experimented with bow + bigram or bow + bigram + pos bow + bigram + pos
Smeelakshmi et al.[67]	Facebook pre-trained word embeddings, SVM-radial bias, Random Forest, SVM-linear	Hindi English code mixed data	10,000 data from different sources	It is found that character-level features give more compared to doc information for code-mixed classification The author used doc2vec and word2vec and FastText library for the feature selection	In the first experiment, the author finds that the RF gives high accuracy of 0.6415% compared to SVM- RBF and simple-linear when incorporating Doc2vec features In the second experiment, the author found that the SVM-RBF gives high accuracy of 0.7511% compared to RF and simple linear by incorporating word2vec embeddings In the 3 <sup>rd</sup> experiment, the author found that SVM-RBF performed better than previous techniques by incorporating FastText embeddings with an accuracy of 0.8581	Classification of the tweets has been not done on multi-classification

\*\*N/A: Authors did not perform the task

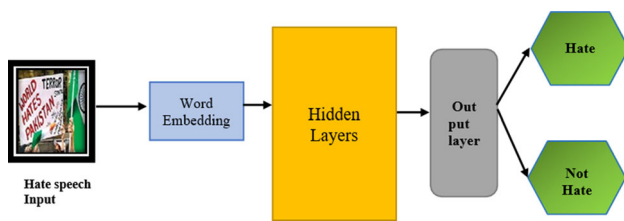


Fig. 11 Deep learning framework For OHS

like a black box for some researchers because it does not require feature engineering. We found that as compared to ML, very little research has been done in the area of deep learning for hate speech detection till 2019. The reasons for the less amount of research in DL can be label data scarcity and unavailability of the high-performance GPU. However, the trend has shifted to deep learning in 2020. According to our findings, we found the majority of research papers from the year 2020 in deep learning as compared to traditional machine learning. In upcoming sections, we will discuss different types of deep learning models that have been used in the previous literature.

### 10.1 Recurrent neural network

ANN cannot capture the sequence of information, which means it does not have an account for the memory. On the other hand, RNN is a type of neural network that captures information about the sequence or time-series data. It can take variable size input and give variable size output and works very well with time-series data. They are a class of artificial neural networks where connections between nodes form a directed graph that allowing information to flow back into the previous parts of the network. Thus, each model in the layers depends on past events, allowing information to persist. RNN works on the given recursive formula in equation 2. In order to detect sentences as hate or not, the author implements tests with RNN, data-partition, epoch, learning rate and batch size. All these parameters affect the system performance. The author [112] used UTFPR models in order to process the text. Then character embeddings fed into the RNN layers. The proposed system is based on the compositional RNN. The proposed model is robust, even when the input data are noisy, but the dataset that is used to feed the RNN is very small, and the performance of the classifier can be affected if a large dataset is taken.

$$S_t = F_w(S_t - 1, X_t) \quad (2)$$

$X_t$  —input at time step  $t$ ;  $S_t$ —state at time step  $t$ ;  $F_w$  —Recursive function

Social media such as Facebook, Twitter and Instagram are becoming a ubiquitous platform for people to share and

express their opinion toward something [113]. Online Social network, especially Twitter, has a prodigious influence on the success or demolition of a person's image [114]. The author [84] used an RNN DL-based approach to detect the hate speech text on Twitter data. Thereafter, 1235 posts were analyzed using case folding, tokenization, cleansing and steaming. The data are collected from the Twitter accounts by the Twitter API. Using RNN (recurrent neural network) and LSTM (long short-term memory), it can process not only single data but also an entire sequence of data at a time. word2vec is used to convert sentences into vector value or to find the semantic meaning. Test the data with epoch, which resultant in high precision of 91% and recall 90% and an accuracy of 91%. The author [115] represents machine learning with a hybrid NLP approach where killer NLP with ensemble deep learning is used to examine the data, which gives 98.71% accuracy of the system. The authors [50] address the problem of identifying speech promoting religious hatred in the Arabic Twitter. They created an Arabic dataset of 6000 tweets annotated for the task of hate speech detection and Arabic lexicon with scores representing their polarity and strength. They also developed the various classification model using a lexicon-based,  $n$ -gram and deep learning-based approach. But the author used GRUs rather than LSTMs because GRUs can be trained faster and may achieve the best performance on datasets that have a limited number of training examples. GRU (gated recurrent unit)-based RNN model produced the best results for the evaluation metrics. The study [134] demonstrates how psychologists have looked into the connection between hate and personality. The author used a text-mining strategy that completely automates the personality inference process. A deep learning algorithm called PERSONA has been developed to identify hate speech online.

### 10.2 Long short-term memory

LSTMs are a modified version of a recurrent neural network capable of learning long-term dependencies, usually used for time series analysis. They can process images, speech and video. It is made up of gates viz. input, output and forget which have the function of, respectively, receiving the data, outputting it and deciding what to pass and what not to. In RNN, we suffer from the vanishing gradient problem, which is as we propagate the error back through all the multiple layers of the RNN. Hence, LSTM solves the problem of vanishing gradient and gives much better accuracy than RNN because RNN fails to establish the long-term dependencies. To classify the OHS, the author [85] used the LSTM classifier and FastText library and found that the binary classifier obtained comparable results as that of sentiment analysis. The author [38] used GloVe embedding-based method and LSTM classifier, in which embeddings learned

from the model, and that leads to high accuracy. The author [79] used two models one Textual model and the second Acoustic model. LSTM model performs better on textual data rather than on acoustic data. To determine the hateful or neural [43], the author used NLP classifiers with paragraph2vec. The performance of the experiment has increased as the number of hidden layers increased, also the author experiment with the five hidden connected units and two hidden layers which gives the 0.99 AUC over 200 iterations. An ensemble of the LSTM classifier improves the classification [115]; also, the author used a combination of various features, which gives the high F score 0.9320. To work in the Hinglish language, author [116] found that the LSTM classifier calculated a maximum recall value of 0.7504 on specific hyperparameter settings.

### 10.3 Convolutional neural network

Convent or CNN, it is a subclass of DNN (deep neural networks). CNN mostly used in the area of analyzing visual imagery. The three layers of an image are converted into a vector of suitable size, and then a DNN is trained on it. Their other applications include video understanding, speech recognition and understanding natural language processing. The author [39] used CNN in order to find racism and sexism speech. The proposed model is tested by ten-fold cross-validation and gives a 78.3% f score. The author [68] employed text features, i.e., surface-level features, linguistic features and sentiment features in deep learning classifiers, and then implemented an ensemble-based novel approach. The author finds the accuracy of 0.918 with the novel approach. Batch size, epoch and learning rate affect the system performance. Also, the studies show that a larger training dataset produces better results [27]. To visualize the online aggression on Twitter and Facebook, the CNN-based web browser plugin had been presented by the authors [117].

### 10.4 Transformer methods

The transformer [118] is the latest innovation that has taken the domain of natural language processing by storm. Transformer like its predecessor has the ability to account long-term dependencies, but unlike LSTMs transformers do not process data sequentially as done in the case of RNN's and LSTMS. Instead, to account for the position of each word is added to its embedding. The transformer was first introduced for machine translations (Sequence to Sequence Model), and thus it has two components, an encoder and a decoder. Though only the encoder is relevant in text classification tasks such as Hate speech detection. It is vital to understand to study transformer in totality. In an encoder the inputs are first fed into a self-attention layer which generates an embedding taking into account all other words in sentence and depicts the

relevance of each word with respect to a particular word. The embeddings obtained from self-attention layer are fed into neural network. This process is repeated many times, i.e., many layers of self-attention and neural networks are stacked to form the encoder. The decoder of a transformer is very similar to the encoder except for an encoder–decoder attention layer, which is added to find the inputs relevant to a particular output [118]. In the context of hate speech detection embedding obtained from the pre-trained model such as BERT (Bidirectional Encoder Representations from Transformers) has been widely used. BERT is a transformer trained using the masked LLM technique. Masked LM technique [119] requires 15% of the words in the sentence to be masked, and the transformer then attempts to predict these words from context during the training process. In the paper [120] the author showed the efficacy of finetuning the Bert in the context of hate speech detection. Comparing pre-trained models for hate speech detection explores and compares various multilingual transformers such as Mbert, Beto.

In this paper [121], the authors argue that, for the multi-class classification problem of online hate speech, transformers must be used over basic traditional machine learning, basic RNN-based deep learning or even attention-based RNN models to achieve the state-of-the-art accuracy. They propose a streamlined version of BERT, called DistilBERT, which has half the number of parameters with no loss in performance. On comparison and experimentation with various LSTM and BERT-based models, DistilBERT outperforms all the models given on various metrics. This paper [122] provides us with a comparative analysis of three different types of models, namely baseline traditional machine learning models, Deep Learning models and Transfer Learning-based models for Hate Speech classification in the Spanish language. This comparison shows how Transfer learning models outperform traditional machine learning models, which are used as the baseline. They evaluate the performance of pre-trained Language models. The authors showcase that the pre-trained monolingual language model (BETO) outperforms pre-trained multilingual models like Bert and XLM, concluding the requirement of hate speech models to be language-specific. In the paper [123], the author uses GPT- 2; it is a language modeling transformer released by open AI. It was trained on a massive dataset of Web text, which required storage space of 40GB and contained parameters ranging from 117 million to 1500 million. Though both BERT and transformers are transformers a stark difference can be observed between these two in their usage; while BERT finds its usage in creating embedding that incorporates the context of whole sentence GPT-2 is widely to generate sentences. The architecture of these transformers presents a stark difference as well, while BERT is entirely made of encoders and GPT-2 is entirely made up of decoders. Further, GPT-2 relies on autoregression that is GPT-2 produces



**Table 9** Classification of type of input to deep learning model

DL classifier	Input type					
	BOW, N-gram, Char n-gram, Skip-gram	Word embeddings, TF-IDF	Positive, negative words	Hate-related terms	POS information	Image, audio, video
ANN	[104, 79]	[43, 104, 79, 84, 80, 79, 78, 77, 75, 74, 73, 72, 64, 65, 46, 45] [44, 43, 42, 41] [40, 39, 35, 38] [37, 36]	N/A	[43, 79]	N/A	[104]
CNN	[37, 110] [8]	[37, 86]	[8]	N/A	[8, 101]	N/A
DNN	[129, 45, 81]	[129]	N/A	[129]	[45]	N/A
RNN	[50]	[84, 130]	N/A	N/A	[101]	N/A
LSTM	[110, 79]	[86, 85]	N/A	[96, 97, 32]	[101]	N/A
Dense NN	N/A	N/A	N/A	N/A	N/A	N/A

\*\*N/A: Authors did not perform the task

tokens sequentially and once one token is produced, it is included as input for the next token. Though the technique of autoregression has its cons since on using auto-regression, the model loses the ability to utilize the context on both sides. It has been proven that GPT-2 achieves excellent results. The authors of this paper [124] propose a novel solution to the hate speech binary classification problem statement by scaling up the small public datasets available using a Deep Generative model, here GPT-2[125] to produce large datasets for the training of Deep Learning-based classifiers and satisfy their extensive data requirements. In the paper, the GPT-2 was finetuned according to the public datasets for the generation of data points. Then they test these models intra-dataset and cross-dataset among the public ones to compare the increase in accuracy and generalization across different probability distributions of datasets. In the paper [126], the author used the transfer learning and Compact Bert variants in a pipeline model. The pre-processed data are loaded into batches of text and true labels and tokenized with a pre-trained BERT tokenizer. The final layer is removed and a dense layer of size 3 is added, because of three different classes then the dense layer SoftMax is used, to get probability scores for each class where maximum probability results in predicted label. Also, Focal loss is used as a cost function. It is beneficial with a class imbalance problem. In order to improve the overall accuracy of the system the author [127] used the ensemble of different features and study the effects of TF-IDF and sentiment bases features. The author also presented the criterion for the selection of computational complexity and classification performance among the existing methods. To detection

of hate speech in Spanish language different pretrained models were analyzed [128], where SVM and logistic regression was used for text categorization and Bert model was finetuned with input of 512 tokens, output vector has dimension of 768. However, the transfer learning models outperformed the traditional machine learning approaches for the Spanish vocabulary.

In Table 9, we have analyzed the types of inputs that can be provided to the deep learning algorithms so that model can perform better by taking low computation resource. However, we did not get satisfactory results as word embedding is the first choice of the researchers for the input parameter and other methods of DL with varied input parameters were not explored. Most of the fields in Table 9 are NA (not applicable), which means that no work has been done using these inputs in the specific type of classifier. In Table 11, we concluded each DL paper's merits and limitations, but it is not very clear in the papers which approach performed better. Also, some recent studies show that deep learning gives better results than a traditional framework, but again these results are not very consistent. Based on the selected 111 papers, we found that most of the authors used SVM, Naive Bayes, Decision Tree in ML and CNN, LSTM in the DL approach also shown in Tables 7 and 9. From the recent trend, we have also all seen that the transformer-based techniques are the most used approaches among the researcher.

From Table 10, we found that most of the authors used SVM, Naive Bayes, Decision Tree in ML and CNN, LSTM in the DL approach. From the recent trend, we have also all seen that the transformer-based techniques are the most used approaches among the researcher (Table 11).



**Table 10** DL algorithms used in the research papers

Algorithm	No of frequencies used in the paper
CNN	11
LSTM	10
RNN	8
DNN	6
Transformer	8

## 10.5 Different organization contribution toward OHS

In this section, we have discussed the various workshops and competitions which contributed to the online hate speech problem.

- o **SemEVAL**
- o It is a research workshop that works to advance the SOTA on semantic analysis and offers different NLP tasks based on semantic analysis to build efficient systems for these problems. Through these challenges, it aims to build datasets that can be publicly used for further research.<sup>2</sup>
- o HASOC (hate speech and offensive content identification in Indo-European Languages)
- o It is a forum that provides datasets in multiple languages for two Hate Speech subtasks for different classification. Participants are expected to use these datasets and create systems as solutions to these subtasks. These datasets comprise ten thousand annotated tweets.<sup>3</sup>
- o **GermEVAL**
- o This is a series of Natural Language Processing tasks in the German language that are released for people to build efficient systems on. The datasets are provided by the forum and are an amalgamation of German tweets.<sup>4</sup>
- o **TRAC**
- o This workshop aims to use NLP and related methods for the detection of online aggression, trolling, cyberbullying and related phenomena in text and speech present on social media platforms to deal with inflammatory content. It has two subtasks, each pertaining to a different set of classes and to solve these problems, it gives 5000 annotated data from social media in Bangla, Hindi and English.<sup>5</sup>
- o Hateful meme challenge

<sup>2</sup> <https://semeval.github.io/SemEval2021/tasks.html>.

<sup>3</sup> <https://hasocfire.github.io/hasoc/2020/index.html>.

<sup>4</sup> <https://swisstext-and-konvens-2020.org/shared-tasks/>.

<sup>5</sup> <https://sites.google.com/view/trac2/live?authuser=0>.

This challenge is organized by Facebook AI, wherein they provide a dataset of memes containing text and images. The goal is to create a system wherein the model is able to accurately identify hate speech in this multimodal dataset and perform classification. The dataset contains 10000+ examples of memes which are annotated.<sup>6</sup>

OSACT4 Shared Task on Offensive Language Detection (Subtask A)

This challenge uses the Arabic SemEVAL dataset for binary classification problem statement of Arabic Hate Speech. The goal is to create a system which is capable of classifying Arabic tweets into offensive or non-offensive.

MEX-A3T

The goal of this community is to improve the further research in misinformation and aggressive speech by improving the research in NLP-related task. This research group provides different tracks to the researchers in the same domain only.

## 10.6 Evaluation metrics

Evaluation metrics are the mathematical functions that provide constructive feedback and are used to measure the quality of a traditional machine learning model. Most of the state-of-the-art online hate speech detection used an F1 score [31, 73, 99], precision [105, 131], recall [43, 131] and accuracy[43] for measuring the effectiveness of the parameters. We have discussed some most used evaluation metrics in the literature. With XAI becoming more and more relevant in artificial Intelligence, it is important to discuss the metrics used to measure the explainability of a model.

1. *Precision* The piece of relevant information from the total information.

$$P = \text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

2. *Recall* The percentage of total relevant information correctly classified by the classifier.

$$R = \text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

3. *F1 score*: An F1 score is defined as the harmonic mean of precision and recall. F1 score has become the preferred choice of measuring the performance of machine learning models. This can be attributed to the fact that F1 score gives equal weightage to both precision and recall and it punishes models that lack even in one of them.

$$F1Score = \frac{(2 * P * R)}{P + R} \quad (3)$$

<sup>6</sup> <https://ai.Facebook.com/blog/hateful-memes-challenge-and-dataset/>.

**Table 11** Deep learning methods for OHS

References	Approach	Language	Dataset	Methodology	Merits	Limitation
Saksesi et al. [84]	RNN	Indonesian	The author collected 1235 Words from Twitter	The author used word2vec embeddings to get the text matrix In order to detect sentences as hate or not, they implement tests with RNN	The author tested the result on a small dataset over which they get PR-91%, RC-90% and, Accuracy-91%	Better results can be achieved if the size of the data increased
Albadi, Kurdi, and Mishra [50]	RNN plus GRU (gated recurrent unit)	Arabic	Collected 600 Arabic tweets from Twitter	The author used MADAMIRA 2.1 to lemmatize the data For a lexicon-based approach features selection, the author used AraHate-PMI (pointwise mutual information), AraHate-Chi, and AraHate-BNS (Bi-Normal Separation)	As compared to lexicon-based and SVM and LR-based model, GRU-Based RNN performs the best with an accuracy of 0.79	The author did not compare the earlier state of the arts with the developed model; also, error analysis is missing
Sazany et al. [85]	LSTM, FastText algorithm	Indonesian	713 Twitter political posts from Twitter	The author used two types of datasets i.e., rpolitics and okkyabusive In this study, two types of word embeddings have been used, i.e., word2vec and FastText library	By using FastText embeddings with rpolitics they achieved the highest results, i.e., 97.39 f1 score	In the proposed research, the model configuration, such as the classifier, number of layers, training batch size, is not analyzed Also, the training and testing dataset size is very small for the desired purpose
Vigna et al. [32]	LSTM SVM	Italian	17567 comments collected from Facebook	To increase the system accuracy, the author also used the sentiment polarity lexicon and word embeddings lexicons	The author found that the binary classifier obtained comparable results as that of sentiment analysis	In the given study, both SVM and LSTM are not able to discriminate between the three classes
Badjatiya et al. [38]	CNN, LSTM, FastText	English	16 K tweets collected on sexism and racism and neither from Twitter publicly available data	In the given study GloVe-pre-trained embedding and 10-Fold Cross-Validation have been used The author applied Adam for CNN and LSTM and RMS-Prop for FastText as an optimizer	In this study, the author found that CNN is performed better than LSTM, which was better than FastText We also learned that Embeddings learned from deep neural network models when combined with gradient boosted decision trees led to the best accuracy Values	Not applicable

**Table 11** (continued)

References	Approach	Language	Dataset	Methodology	Merits	Limitation
Park and Fung [39]	HybridCNN	English	Waseem and Hovy 2016 English dataset (20 k)	The author implements three CNN-based models to classify sexist and racist abusive language i.e., CharCNN, WordCNN and HybridCNN Max pooling is performed after the convolution to capture the feature that is most significant to the output	The proposed model is tested by tenfold cross-validation and gives a 78.3% F score	More precise results can be explored if training the two-step classifiers on separate datasets (larger dataset)
Paetzold et al. [112]	RNN	English and Spanish	HatEval website	The author used UTFPR (minimalistic Recurrent Neural Networks) models in order to process the text The character embeddings are used in the RNN layers The proposed system is based on the compositional RNN	The proposed model is robust, even when the input data are noisy	More reliable ways of re-using pre-trained compositional models can be tested
Sutejo and Lestari [79]	LSTM	Indonesian	Two types of data, text-2273 and audio-2469. Collected from different social media websites	The author used two models one Textual model and the second Acoustic model Word n-gram features are employed for the classification For the acoustic model author used low-level descriptor features and, Uni-bi-bow features which give the high F1 score for the textual model and MFCC_E_D_A features (of an acoustic model)	The author found the textual model gives the best result as that of the acoustic model	CBOW (87.98%) performed better than word n-gram and their combination (the highest achieved 83.91%) In the study, there are several incorrect results due to the bias of some people
Andreou et al. [68]	Ensemble-based classification using CNN DNN RNN	English	Davidson et al	The author employed text features, i.e., surface-level features, linguistic features and sentiment features The author implemented novel ensemble-based classification	Mandola processes the information in real The author finds the accuracy of 0.918 with the novel approach Mandola is the first system that provides a systematic and integrated approach for detecting hate speech	The proposed system is not compatible with any cross-lingual

In multiclass classification there are mainly two methods of calculating F1 score, namely microaveraged F1 score and macroaverage F1 score.

A) *F1 microaveraged* This metric is simply calculated by taking the harmonic mean of microprecision and microrecall. An important feature of this metric is that it assigns equal value to each label, the repercussion of which is the not enough attention is given to minority classes in case of imbalanced datasets. Since Imbalanced datasets are seen in abundance in the domain of hate speech detection, the use of microaveraged F1 score should be minimized.

$$\text{Micro Averaged Precision} = \frac{\sum TP}{\sum TP + \sum FP} \quad (4)$$

$$\text{Micro averaged recall} = \frac{\sum TP}{\sum TP + \sum FN} \quad (5)$$

B) *F1 macroaveraged* This is calculated by simply taking the mean of F1 scores obtained on each class individually. This metric assigns equal value to each class and thus should be the preferred metric in the context hate speech detection where datasets are generally imbalanced and models are expected to be proficient in detecting all classes.

4. *Confusion matrix* It is a performance measurement matrix comparing the actual and predicted observations through the values of False Positives (FP), True Negatives (TN), True Positive (TP) and False Negative (FN) labels (Matrix 1).

$$\text{Confusion Matrix} = \begin{bmatrix} TP & FN \\ FN & TN \end{bmatrix} \quad (6).$$

5. *Accuracy* Is the measure which tells how efficiently the classification models produce the results correctly.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (7)$$

6. *Comprehensiveness*: In XAI, we essentially try to predict the factors which led to a model's decision. To calculate the comprehensiveness, the factors predicted by the XAI model are first removed from the datapoint. In the context of hate speech detection, the equivalent of this is removing the words predicted by the XAI model. Now, this new modified datapoint is then fed into the model. The change in the model's confidence in prediction is noted. A change implies that the factors predicted by the model indeed contributed to the model's decision[132].

7. *Sufficiency*: This metric measures how important the extracted rationales(words or phrases in the context of Hate speech detection) for the model to make a prediction[132].

8. *Matthews correlation coefficient (MCC)*: It tries to find the relation between the true and predicted values. Higher value of the coefficient shows the better results. Whenever the given dataset is highly imbalance in that case it is found that

MCC has given best results compared to the accuracy [133]. Its value always lies between -1 and 1. The given formula is shown in Eq. 7.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FN)(TN + FP)(TN + FN)}} \quad (8).$$

Both precision and recall are very important and the most used evaluation metrics in traditional machine learning and deep learning classification. We can calculate the accuracy by providing the given values to TN, TP, FP, and FN. By getting the values of precision and recall from equations 1 and 2, we can calculate the F1 score that is used to test the accuracy of the parameter. Some authors also used AUC (area under the curve) to compute the performance of the model. The aforementioned metric evaluation formulas were used by mostly all other authors mentioned in related works to evaluate the performance of their machine learning model.

## 11 Findings, conclusion and research gaps

The growth of social media has been exponential and people are sharing information, expressing opinions like never before. However, research on hate speech has not been able to keep pace with the multiplicity of social media platforms and their associated problems. Our goal was to cover all the aspects that play an essential role in the field of OHS detection. But our study is limited to computer science background, and we have not considered the culture-specific ways of communication in a different language for detecting OHS. In this survey, we presented a systematic approach that investigates the types of features and classifiers that are most used in OHS detection. From the survey, we found that SVM, Naïve Bayes, Decision Tree, CNN and LSTM are the most used algorithm, and surface-level features are the first choice of the researcher. We learned the concept of hate speech and laws to limit hate speech. Additionally, we presented an application of hate speech. We concluded that very limited studies and papers had been published in the OHS detection from the computer science perspective. We also found that most of the authors used self-generated datasets which are not available online so to find the credibility of these dataset and results achieved with these datasets is also a problem in itself. Finally, we identified some challenges in the field of OHS, the availability of open-source code and the self-generated dataset link, which leads to the lack of comparative studies that can evaluate the existing approaches.

Based on our study, we found several research gaps which can be considered in future work.

- From the study, it has been observed that the existing research covers mostly lexicon (simple keywords)-based features for the hate speech analysis, which restricted the results because the models will not be suitable if whole meaning of the sentence is needed. So, knowledge-based

feature, semantic features can be taken into consideration with lexicon-based features. By this, the accuracy of the model can be increased.

- Facebook, Twitter and social media platforms do not have a real-time hate speech detecting system, and the corrective measures are taken only after the expression is posted online. So, the hate speech detecting plugin can be made, which can analyze hate speech in real time.
- We also invested that hate speech does not only come in the form of text but can take the form of audio, video, picture, etc. But in the area of hate speech detection multimodal OHS detection is very less unexplored.
- The research work has been limited to spotting hate in the English language and few pieces of research in Arabic, Indonesian, Italian, Turkish, Swedish, Albanian Language and hate content in the rest of the languages like Hindi goes unfiltered.
- Another limitation that we found is to get the balanced dataset for the OHS. A very limited and less skewed dataset is available online.
- To lubricate the online hate speech detection and analysis, the unlabeled data should be examined for the unsupervised machine learning model as the labeling of data is a very time-consuming task. Therefore, to address hate speech problems, further study of the deep learning model is essential and advantageous.
- In order to furnish research in the field, a multimodal and multilingual dataset should be developed.
- Some cultures may represent anger and hate in linguistically distinct ways, which can be taken into consideration while building the online hate speech model.

#### Implication of study

This study is highlighting the need for interdisciplinary collaboration between computer science and other fields, such as linguistics, sociology and psychology, to develop more comprehensive approaches to OHS detection that take into account language and cultural differences.

Academics can benefit from this study by understanding the current state of the art in OHS detection, the most commonly used algorithms and surface-level features. This study's limitations can help researchers identify gaps in the field and focus on exploring culture-specific ways of communication for detecting OHS. Practitioners in the field of social media moderation can use this study to inform their strategies for identifying and removing hate speech from social media platforms. This research's findings can help them determine which algorithms and features are most effective in OHS detection. Policymakers and politicians can use this study to inform legislation and regulations around hate speech and social media. The study's presentation of hate speech and the laws that limit it can help policymakers better understand

the issue and take informed actions to address it. The challenges identified in the study, such as the lack of open-source code and self-generated datasets, can inform future research and development efforts in OHS detection. Addressing these challenges can lead to the development of better approaches to OHS detection and more reliable datasets, enabling more comparative studies to evaluate existing approaches. In summary, this study on OHS detection in the context of social media can provide valuable insights for various stakeholders and inform future research, policymaking and social media moderation strategies.

**Data availability statements** Data generated or analyzed during this study are included in this published article.

#### Declarations

**Conflict of interest** All the authors of this paper declare that he/she has no conflict of interest.

**Ethical approval** This article does not contain any studies with human participants or animals performed by any of the authors.

#### References

1. Newman, N., Fletcher, R., Kalogeropoulos, A. et al.: Reuters Institute Digital News Report 2018 (2018)
2. Global social media ranking (2019). <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>
3. Diwhu, G., Ghdwk, W.K.H., Ihpdo, R.I.D., Vwxghqw, X.: Automated detection of hate speech towards woman on Twitter. In: International Conference On Computer Science And Engineering. pp 7–10 (2018)
4. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. *ACM Comput Surv* (2018). <https://doi.org/10.1145/3232676>
5. bbc Facebook launches initiative to fight online hate speech. In: bbc. ps://[www.bbc.com/news/technology-40371869](http://www.bbc.com/news/technology-40371869)
6. Organisation International Alert (2016) A plugin to counter hate speech online. <https://europeanjournalists.org/mediaagainsthate/hate-checker-plugin-to-counter-hate-speech-online/>
7. Salminen, J., Guan, K., Jung, S.G. et al.: A literature review of quantitative persona creation. In: Conf Hum Factors Comput Syst - Proc 1–15 (2020). <https://doi.org/10.1145/3313831.3376502>
8. Biere, S., Analytics, M.B.: Hate speech detection using natural language processing techniques. *VRIJE Univ AMSTERDAM* 30 (2018)
9. DePaula, N., Fietkiewicz, K.J., Froehlich, T.J. et al.: Challenges for social media: misinformation, free speech, civic engagement, and data regulations. In: Proceedings of the Association for Information Science and Technology. pp. 665–668 (2018)
10. Varade, R.S., Pathak, V.: Detection of hate speech in hinglish language. In: ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers) (2020)
11. Djuric, N., Zhou, J., Morris, R. et al.: Hate speech detection with comment embeddings. In: Proceedings of the 24th International Conference on World Wide Web. Association for Computing Machinery, New York, NY, USA, pp. 29–30 (2015)



12. Davidson, T., Warmley, D., Macy, M., Weber, I.: Automated Hate Speech Detection and the Problem of Offensive Language. (2017). arXiv170304009v1 [cs.CL] 11 Mar 2017 Autom
13. Miró-Llinares, F., Moneva, A., Esteve, M.: Hate is in the air! But where? Introducing an algorithm to detect hate speech in digital microenvironments. *Crime Sci.* **7**, 1–12 (2018). <https://doi.org/10.1186/s40163-018-0089-1>
14. Daniel Burke The four reasons people commit hate crimes. In: CNN. <https://edition.cnn.com/2017/06/02/us/who-commits-hate-crimes/index.html>
15. Equality and Diversity Forum (2018) Hate Crime: Cause and effect | A research synthesis. Equal Divers Forum
16. ONTARIO PO, GENERAL MOA: CROWN POLICY MANUAL (2005). [https://files.ontario.ca/books/crown\\_prosecution\\_manual\\_english\\_1.pdf](https://files.ontario.ca/books/crown_prosecution_manual_english_1.pdf)
17. Räsänen, P., Hawdon, J., Holkeri, E., et al.: Targets of online hate: examining determinants of victimization among young finnish Facebook users. *Violence Vict.* **31**, 708–725 (2016)
18. Contributors, W.: Hate crime. In: Wikipedia (2020). [https://en.wikipedia.org/wiki/Hate\\_crime](https://en.wikipedia.org/wiki/Hate_crime)
19. twitter Twitter policy against Hate speech. <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>
20. facebook Hate speech. [https://www.facebook.com/communitystandards/hate\\_speech](https://www.facebook.com/communitystandards/hate_speech)
21. Instagram Instagram policy for hate speech. <https://help.instagram.com/477434105621119>
22. Youtube YouTube hate policy. <https://support.google.com/youtube/answer/2801939?hl=en>
23. Dr. Amarendra Bhushan Dhiraj: Countries Where Cyber-bullying Was Reported The Most In 2018 (2018)
24. United nations: Universal Declaration of Human Rights (1948)
25. Nations S-G of the U: European Convention on Human Rights, the International Covenant on Civil and Political Rights (1966)
26. Gagliardone, I., Patel, A., Pohjonen, M.: Mapping and analysing hate speech online. In: SSRN Electronic Journal. p 41 (2015)
27. Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Pp. 1–10 (2017)
28. Nastiti, F.E., Prastyanti, R.A., Taruno, R.B., Hariyadi, D.: Social media warfare in Indonesia political campaign: a survey. In: Proceedings - 2018 3rd International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2018. IEEE, pp 49–53 (2019)
29. Kumar, A., Sachdeva, N.: Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis. *Multimed. Tools Appl.* (2019). <https://doi.org/10.1007/s11042-019-7234-z>
30. Waqas, A., Salminen, J., Jung, S., et al.: Mapping online hate: a scientometric analysis on research trends and hotspots in research on online hate. *PLoS ONE* **14**, 1–21 (2019). <https://doi.org/10.1371/journal.pone.022194>
31. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In: Association for Computational Linguistics Proceedings of NAACL-HLT. pp 88–93 (2016)
32. Vigna, F. Del, C. A., Orletta, F.D. et al.: Hate me , hate me not : Hate speech detection on Facebook. In: In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy. pp 86–95 (2017)
33. Agarwal S, Sureka A (2017) But I did not mean it! - Intent classification of racist posts on tumblr. In: Proceedings - 2016 European Intelligence and Security Informatics Conference, EISIC 2016. IEEE, pp 124–127
34. CodaLab Competition. <https://competitions.codalab.org/competitions/19935>.
35. Wang, G., Wang, B., Wang, T. et al: Whispers in the dark: Analysis of an anonymous social network. In: Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC. pp 137–149 (2014)
36. Ziai, A.: cohen kappa. In: Medium (2017). <https://towardsdatascience.com/inter-rater-agreement-kappas-69cd8b91ff75>
37. Gambäck, B., Sikdar, U.K.: Using Convolutional Neural Networks to Classify Hate-Speech. In: Proceedings of the First Workshop on Abusive Language Online. pp 85–90 (2017)
38. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep Learning for Hate Speech Detection in Tweets. In: arXiv:1706.00188v1 [cs.CL]. p 2 (2017)
39. Park, J.H., Fung, P.: One-step and two-step classification for abusive language detection on Twitter. In: Association for Computational Linguistics Proceedings of the First Workshop on Abusive Language Online, pages 41–45, Vancouver, Canada, July 30. pp 41–45 (2017)
40. Waseem, Z.: Are you a racist or am i seeing things ? Annotator influence on hate speech detection on Twitter. In: Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science. pp 138–142 (2016)
41. Jha, A: When does a Compliment become Sexist ? Analysis and Classification of Ambivalent Sexism using Twitter Data. In: Proceedings of the Second Workshop on Natural Language Processing. pp 7–16 (2017)
42. Davidson, T., Warmley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language \*. In: arXiv (2017)
43. Alorainy, W., Burnap, P., Liu, H.A.N., Williams, M.L.: “ The Enemy Among Us ”: detecting cyber hate speech with threats-based othering language embeddings. *ACM Trans. Web* **13** (2019)
44. Nobata, C., Tetreault, J.: Abusive language detection in online user content. In: International World Wide Web Conference. Pp. 145–153 (2016)
45. Al, Z., Amr, M.: Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach. *Springer Comput.* (2019) <https://doi.org/10.1007/s00607-019-00745-0>
46. Detecting Insults in Social Commentary. <https://www.kaggle.com/c/detecting-insults-in-social-commentary>
47. MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N.F.O. (2019) Hate speech detection: challenges and solutions. *PLoS ONE* **14**(8): e0221152. <https://doi.org/10.1371/journal.pone.0221152>. <https://sites.google.com/view/trac1/shared-task>
48. Timothy Quinn: Hatebase database. (2017). <https://www.hatebase.org/>
49. Charitidis, P., Doropoulos, S., Vologiannidis, S., et al.: Towards countering hate speech against journalists on social media. *Online Soc. Netw. Media* **17**, 10 (2020). <https://doi.org/10.1016/j.osnem.2020.100071>
50. Albadi, N., Kurdi, M., Mishra, S.: Are they our brothers? Analysis and detection of religious hate speech in the Arabic Twittersphere. In: Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018. IEEE, (pp. 69–76) (2018)
51. Al-Hassan, A., Al-Dossari, H.: Detection of hate speech in Arabic tweets using deep learning. *Multimed. Syst.* (2021). <https://doi.org/10.1007/s00530-020-00742-w>
52. Ousidhoum, N., Lin, Z., Zhang, H. et al.: Multilingual and multi-aspect hate speech analysis. *EMNLP-IJCNLP 2019 - 2019 Conf Empir Methods Nat Lang Process 9th Int Jt Conf Nat Lang Process Proc Conf* 4675–4684 (2020). <https://doi.org/10.18653/v1/d19-1474>
53. Mulki, H., Haddad, H., Bechikh Ali, C., Alshabani, H.: L-HSAB: A Levantine Twitter dataset for hate speech and abusive language, pp. 111–118 (2019). <https://doi.org/10.18653/v1/w19-3512>



54. Ljubešić, N., Erjavec, T., Fišer, D.: Datasets of Slovene and Croatian moderated news comments, pp. 124–131 (2019). <https://doi.org/10.18653/v1/w18-5116>
55. Dinakar, K.: Modeling the detection of textual cyberbullying. In: 2011, Association for the Advancement of Artificial Intelligence, pp 11–17 (2011)
56. Greevy, E., Smeaton, A.F.: Classifying racist texts using a support vector machine. In: ACM Proceeding, pp 468–469 (2004)
57. Watanabe, H., Bouazizi, M., Ohtsuki, T.: Hate speech on Twitter: a pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access* **6**, 13825–13835 (2018). <https://doi.org/10.1109/ACCESS.2018.2806394>
58. Rodriguez, A., Argueta, C., Chen, Y.L.: Automatic detection of hate speech on facebook using sentiment and emotion analysis. In: 1st International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2019. Pp. 169–174 (2019)
59. Hall, L.O., WPKNVCKWB.: snopes.com: Two-striped Telamonia Spider. *J Artif Intell Res* **2009**, 321–357 (2006). <https://doi.org/10.1613/jair.953>
60. Raufi, B., Xhaferri, I.: Application of machine learning techniques for hate speech detection in mobile applications. In: 2018 International Conference on Information Technologies, InfoTech 2018 - Proceedings. IEEE, pp 1–4 (2018)
61. Waseem, Z., Thorne, J., Bingel, J.: Bridging the gaps: multi task learning for domain transfer of hate speech detection. In: Online Harassment, Human–Computer Interaction Series, pp 29–55 (2018)
62. Lynn, T., Endo, P.T., Rosati, P., et al.: Data set for automatic detection of online misogynistic speech. *Data Br.* **26**, 104223 (2019). <https://doi.org/10.1016/j.dib.2019.104223>
63. Plaza-Del-Arco, F.-M., Molina-González, M.D., Ureña-López, L.A., Martín-Valdivia, M.T.: Detecting Misogyny and Xenophobia in Spanish Tweets using language technologies. *ACM Trans. Internet Technol.* **20**, 1–19 (2020). <https://doi.org/10.1145/3369869>
64. Pelzer, B., Kaati, L., Akrami, N.: Directed digital hate. In: 2018 IEEE International Conference on Intelligence and Security Informatics, ISI 2018, pp. 205–210 (2018)
65. Martins, R., Gomes, M., Almeida, J.J. et al.: Hate speech classification in social media using emotional analysis. In: Proceedings - 2018 Brazilian Conference on Intelligent Systems, BRACIS 2018, pp. 61–66 (2018)
66. Basak, R., Sural, S., Ganguly, N., Ghosh, S.K.: Online public shaming on Twitter: detection, analysis, and mitigation. *IEEE Trans. Comput. Soc. Syst.* **6**, 208–220 (2019). <https://doi.org/10.1109/TCSS.2019.2895734>
67. Sreelakshmi, K., Premjith, B., Soman, K.P.: Detection of hate speech text in Hindi-English Code-mixed Data. *Procedia Comput. Sci.* **171**, 737–744 (2020). <https://doi.org/10.1016/j.procs.2020.04.080>
68. Andreou, A., Orphanou, K., Pallis, G.: MANDOLA : A Big-Data Processing and Visualization. *ACM Trans. Internet Technol.* **20** (2020)
69. Zimbra, D., Abbasi, A., Zeng, D., Chen, H.: The state-of-the-art in Twitter sentiment analysis. *ACM Trans. Manag. Inf. Syst.* **9**, 1–29 (2018). <https://doi.org/10.1145/3185045>
70. Mariconti, E., Suarez-Tangil, G., Blackburn, J., et al.: “You know what to do”: proactive detection of YouTube videos targeted by coordinated hate attacks. *Proc ACM Hum.-Comput. Interact* (2019). <https://doi.org/10.1145/3359309>
71. Gitari ND, Zuping Z, Damien H, Long J (2015) A Lexicon-based approach for hate speech detection a Lexicon-based approach for hate speech detection. *Int. J. Multimed. Ubiquitous Eng.* <https://doi.org/10.14257/ijmue.2015.10.4.21>
72. Lima, L., Reis, J.C.S., Melo, P. et al.: Inside the right-leaning echo chambers: characterizing gab, an unmoderated social system. In: Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018. pp 515–522 (2018)
73. Watanabe, H., Bouazizi, M., Ohtsuki, T.: Hate speech on Twitter : a pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, pp. 13825–13835 (2018)
74. Ruwandika, N.D.T., Weerasinghe, A.R.: Identification of hate speech in social media. In: 2018 International Conference on Advances in ICT for Emerging Regions (ICTer) : Identification. IEEE, pp. 273–278 (2018)
75. Alorainy W, Burnap P, Liu H, et al.: Suspended accounts : a source of tweets with disgust and anger emotions for augmenting hate speech data sample. In: Proceeding of the 2018 International Conference on Machine Learning and Cybernetics. IEEE (2018)
76. Setyadi, N.A., Nasrun, M., Setianingsih, C.: Text analysis for hate speech detection using backpropagation neural network. In: The 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC). IEEE, pp 159–165 (2018)
77. Alfina, I., Mulia, R., Fanany, M.I., Ekanata, Y.: Hate speech detection in the Indonesian language: A dataset and preliminary study. In: 2017 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2017. pp 233–237 (2018)
78. Sharma, H.K., Singh, T.P., Kshitiz, K., et al.: Detecting hate speech and insults on social commentary using NLP and machine learning. *Int. J. Eng. Technol. Sci. Res.* **4**, 279–285 (2017)
79. Sutejo, T.L., Lestari, D.P.: Indonesia hate speech detection using deep learning. In: International Conference on Asian Language Processing. IEEE, pp 39–43 (2018)
80. Lekea, I.K.: Detecting hate speech within the terrorist argument : a greek case. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, pp 1084–1091 (2018)
81. Liu, H., Burnap, P., Alorainy, W., Williams, M.L.: A fuzzy approach to text classification with two-stage training for ambiguous instances. *IEEE Trans. Comput. Soc. Syst.* **6**, 227–240 (2019). <https://doi.org/10.1109/TCSS.2019.2892037>
82. Wang, J., Zhou, W., Li, J., et al.: An online sockpuppet detection method based on subgraph similarity matching. In: Proceedings - 16th IEEE International Symposium on Parallel and Distributed Processing with Applications, 17th IEEE International Conference on Ubiquitous Computing and Communications, 8th IEEE International Conference on Big Data and Cloud Computing, 11th IEEE, pp. 391–398 (2019)
83. Wu, K., Yang, S., Zhu, K.Q.: False rumors detection on Sina Weibo by propagation structures. In: Proc - Int Conf Data Eng 2015-May:651–662 (2015). <https://doi.org/10.1109/ICDE.2015.7113322>
84. Saksesi, A.S., Nasrun, M., Setianingsih, C.: Analysis text of hate speech detection using recurrent neural network. In: The 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC) Analysis. IEEE, pp. 242–248 (2018)
85. Sazany, E.: Deep learning-based implementation of hate speech identification on texts in Indonesian : Preliminary Study. In: 2018 International Conference on Applied Information Technology and Innovation (ICAITI) Deep. IEEE, pp 114–117 (2018)
86. Son, L.H., Kumar, A., Sangwan, S.R., et al.: Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network. *IEEE Access* **7**, 23319–23328 (2019). <https://doi.org/10.1109/ACCESS.2019.2899260>

87. Salminen, J., Hopf, M., Chowdhury, S.A., et al.: Developing an online hate classifier for multiple social media platforms. *Human-centric Comput. Inf. Sci.* **10**, 1–34 (2020). <https://doi.org/10.1186/s13673-019-0205-6>
88. Coste, R.L. (2000) Fighting speech with speech: David Duke, the anti-defamation league, online bookstores, and hate filters. In: *Proceedings of the Hawaii International Conference on System Sciences*. p 72
89. Gelber, K.: Terrorist-extremist speech and hate speech: understanding the similarities and differences. *Ethical Theory Moral Pract.* **22**, 607–622 (2019). <https://doi.org/10.1007/s10677-019-10013-x>
90. Zhang, Z.: Hate speech detection: a solved problem ? The challenging case of long tail on Twitter. *Semant WEB IOS Press* **1**, 1–5 (2018)
91. Hara, F.: Adding emotional factors to synthesized voices. In: *Robot and Human Communication - Proceedings of the IEEE International Workshop*, Pp. 344–351 (1997)
92. Fatahillah, N.R., Suryati, P., Haryawan, C.: Implementation of Naive Bayes classifier algorithm on social media (Twitter) to the teaching of Indonesian hate speech. In: *Proceedings—2017 International Conference on Sustainable Information Engineering and Technology, SIET 2017*, pp. 128–131 (2018)
93. Ahmad Niam, I.M., Irawan, B., Setianingsih, C., Putra, B.P.: Hate speech detection using latent semantic analysis (LSA) method based on image. In: *Proceedings - 2018 International Conference on Control, Electronics, Renewable Energy and Communications, ICCEREC 2018*. IEEE, pp. 166–171 (2019)
94. Gitari, N.D., Zuping, Z., Damien, H., Long, J.: A lexicon-based approach for hate speech detection. *Int. J. Multimed. Ubiquitous Eng.* **10**, 215–230 (2015)
95. Chen, Y., Zhou, Y., Zhu, S., Xu, H.: Detecting offensive language in social media to protect adolescent online safety. In: *Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012*. IEEE, pp. 71–80 (2012)
96. Pitsilis, G.K., Ramampiaro, H., Langseth, H.: Effective hate-speech detection in Twitter data using recurrent neural networks. *Appl. Intell.*, Pp. 4730–4742 (2018)
97. Pitsilis, G.K., Ramampiaro, H., Langseth, H.: Detecting offensive language in Tweets using deep learning (2018). *arXiv:180104433v1* 1–17. <https://doi.org/10.1007/s10489-018-1242-y>
98. Warner, W., Hirschberg, J.: Detecting hate speech on the World Wide Web. In: *Association for Computational Linguistics Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012)*, pp. 19–26 (2012)
99. Dinakar, K., Jones, B., Havasi, C., Lieberman, H.: Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Trans. Interact. Intell. Syst.* **2**, 30 (2012). <https://doi.org/10.1145/2362394.2362400>
100. Burnap, P., Williams, M.L.: Cyber hate speech on twitter: an application of machine classification and statistical modeling for policy and decision making. *Policy Internet* **7**, 223–242 (2015). <https://doi.org/10.1002/poi3.85>
101. Garc, A: Hate speech dataset from a white supremacy forum. In: *Proceedings of the Second Workshop on Abusive Language Online*, pp. 11–20 (2018)
102. Ombui, E., Karani, M., Muchemi, L.: Annotation framework for hate speech identification in Tweets : Case Study of Tweets During Kenyan Elections. In: *2019 IST-Africa Week Conference (IST-Africa)*. IST-Africa Institute and Authors, pp. 1–9 (2019)
103. Hosseinmardi, H., Mattson, S.A., Rafiq, R.I. et al.: Detection of cyberbullying incidents on the Instagram Social Network. In: *arXiv:1503.03909v1 [cs.SI]* 12 Mar 2015 Abstract (2015)
104. Raufi, B., Xhaferri, I.: Application of machine learning techniques for hate speech detection in mobile applications. In: *2018 International Conference on Information Technologies (InfoTech-2018)*, IEEE Conference Rec. No. 46116 20–21 September 2018, St. St. Constantine and Elena, Bulgaria. IEEE (2018)
105. Warner, W., Hirschberg, J.: Detecting hate speech on the World Wide Web. In: *19 Proceedings of the 2012 Workshop on Language in Social Media (LSM)*. pp 19–26 (2012)
106. Wang, G., Wang, B., Wang, T. et al.: Whispers in the dark : analysis of an anonymous social network categories and subject descriptors. *ACM* **13** (2014)
107. Mathew, B., Saha, P., Yimam, S.M. et al.: HateXplain: a benchmark dataset for explainable hate speech detection. In: *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*. p 12 (2020)
108. Kiilu, K.K., Okeyo, G., Rimiru, R., Ogada, K.: Using Naïve Bayes Algorithm in detection of Hate Tweets. *Int. J. Sci. Res. Publ.* **8**:99–107. <https://doi.org/10.29322/ijsrp.8.3.2018.p7517> (2018)
109. Sanchez, H.: Twitter Bullying Detection, pp. 1–7 (2016). In: <https://www.researchgate.net/publication/267823748>
110. Gröndahl, T., Pajola, L., Juuti, M. et al.: All you need is “love”: Evading hate speech detection. In: *Proceedings of the ACM Conference on Computer and Communications Security*. pp 2–12 (2018)s
111. Correa, D., Silva, L.A., Mondal, M., et al.: The many shades of anonymity : characterizing anonymous social media content. *Assoc Adv. Artif. Intell.* **10** (2015)
112. Paetzold, G.H., Malmasi, S., Zampieri, M.: UTFPR at SemEval-2019 Task 5: Hate Speech Identification with Recurrent Neural Networks. In: *arXiv:1904.07839v1*. p 5 (2019)
113. Miro-Llinares, F., Rodriguez-Sala, J.J.: Cyber hate speech on twitter: analyzing disruptive events from social media to build a violent communication and hate speech taxonomy. *Int. J. Design Nat. Ecodyn.* pp 406–415 (2016)
114. Rizoiu, M.-A., Wang, T., Ferraro, G., Suominen, H.: Transfer learning for hate speech detection in social media. *arXiv:190603829v1* (2019)
115. Pitsilis, G.K., Ramampiaro, H., Langseth, H.: Effective hate-speech detection in Twitter data using recurrent neural networks. *Appl. Intell.* **48**, 4730–4742 (2018). <https://doi.org/10.1007/s10489-018-1242-y>
116. Varade, R.S., Pathak, V.B.: Detection of hate speech in hinglish language. *Adv. Intell. Syst. Comput.* **1101**, 265–276 (2020). [https://doi.org/10.1007/978-981-15-1884-3\\_25](https://doi.org/10.1007/978-981-15-1884-3_25)
117. Modha, S., Majumder, P., Mandl, T., Mandalia, C.: For surveillance detecting and visualizing hate speech in social media: a cyber watchdog for surveillance. *Expert Syst. Appl.* (2020). <https://doi.org/10.1016/j.eswa.2020.113725>
118. Maxime: What is a Transformer?No Title. In: *Medium* (2019). <https://medium.com/inside-machine-learning/what-is-a-transformer-d07dd1fbec04>
119. Horev R BERT Explained: State of the art language model for NLP Title. <https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270>
120. Mozafari, M., Farahbakhsh, R., Crespi, N.: A BERT-based transfer learning approach for hate speech detection in online social media. *Stud. Comput. Intell.* **881** SCI:928–940 (2020). [https://doi.org/10.1007/978-3-030-36687-2\\_77](https://doi.org/10.1007/978-3-030-36687-2_77)
121. Mutanga, R.T., Naicker, N., Olugbara, O.O. (2020) Hate speech detection in twitter using transformer methods. *Int. J. Adv. Comput. Sci. Appl.*; **11**, 614–620 . <https://doi.org/10.14569/IJACSA.2020.0110972>
122. Plaza-del-Arco, F.M., Molina-González, M.D., Ureña-López, L.A., Martín-Valdivia, M.T.: Comparing pre-trained language

- models for Spanish hate speech detection. *Expert Syst. Appl.* **166** (2021)
123. Pandey, P.: Deep generative models. In: medium. <https://towardsdatascience.com/deep-generative-models-25ab2821afd3>
  124. Wullach, T., Adler, A., Minkov, E.M.: Towards hate speech detection at large via deep generative modeling. *IEEE Internet Comput.* (2020). <https://doi.org/10.1109/MIC.2020.3033161>
  125. Dugas, D., Nieto, J., Siegwart, R., Chung, J.J.: NavRep : Unsupervised representations for reinforcement learning of robot navigation in dynamic human environments (2021)
  126. Behzadi, M., Harris, I.G., Derakhshan, A.: Rapid cyber-bullying detection method using compact BERT models. In: *Proc - 2021 IEEE 15th Int Conf Semant Comput ICSC 2021* 199–202. (2021) <https://doi.org/10.1109/ICSC50631.2021.00042>
  127. Araque, O., Iglesias, C.A.: An ensemble method for radicalization and hate speech detection online empowered by sentic computing. *Cognit. Comput.* (2021). <https://doi.org/10.1007/s12559-021-09845-6>
  128. Plaza-del-Arco, F.M., Molina-González, M.D., Ureña-López, L.A., Martín-Valdivia, M.T.: Comparing pre-trained language models for Spanish hate speech detection. *Expert Syst. Appl.* **166**, 114120 (2021). <https://doi.org/10.1016/j.eswa.2020.114120>
  129. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: *26th International World Wide Web Conference 2017, WWW 2017 Companion* (2019)
  130. Mossie, Z., Wang, J.H.: Vulnerable community identification using hate speech detection on social media. *Inf. Process Manag.* **57**, 102087 (2020). <https://doi.org/10.1016/j.ipm.2019.102087>
  131. Magu, R., Joshi, K., Luo, J.: Detecting the hate code on social media. In: *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*. pp 608–611 (2017)
  132. Qian, J., Bethke, A., Liu, Y., et al.: A benchmark dataset for learning to intervene in online hate speech. In: *EMNLP-IJCNLP 2019 - 2019 Conf Empir Methods Nat Lang Process 9th Int Jt Conf Nat Lang Process Proc Conf* 4755–4764 (2020). <https://doi.org/10.18653/v1/d19-1482>
  133. Chicco, D., Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **21**, 1–13 (2020). <https://doi.org/10.1186/s12864-019-6413-7>
  134. Lee, K., Ram, S.: PERSONA: Personality-based deep learning for detecting hate speech. In: *International Conference on Information Systems, ICIS 2020 - Making Digital Inclusive: Blending the Local and the Global. Association for Information Systems* (2021)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

# Impact of Feature Selection Algorithms on Network Intrusion Detection

Samyak Jain

Dept. of Applied Mathematics  
Delhi Technological University  
New Delhi, India

samyakjain\_2k19mc114@dtu.ac.in

Siddharth Bihani

Dept. of Applied Mathematics  
Delhi Technological University  
New Delhi, India

siddharthbihani\_2k19mc125@dtu.ac.in

Satyam Jaiswal

Dept. of Applied Mathematics  
Delhi Technological University  
New Delhi, India

satyamjaiswal\_2k19mc116@dtu.ac.in

Anshul Arora

Dept. of Applied Mathematics  
Delhi Technological University  
New Delhi, India

anshularora@dtu.ac.in

**Abstract**—This research paper discusses the importance of efficient network intrusion detection as a key element in the field of cybersecurity. Network Intrusion Detection Systems (NIDS) are crucial for spotting and stopping unwanted activity on computer networks as cyber-attacks become more sophisticated. Various feature selection algorithms have been used in the literature for network intrusion detection systems. In this work, we select three feature ranking techniques namely, Anova F-test, Mutual Information, and Chi-square test to rank the traffic features. The UNSW-NB15 dataset was used in the research to assess the effectiveness of five well-known machine learning models, namely Logistic Regression, Decision Trees, Random Forest, K-Nearest Neighbours, and Naive Bayes. In order to assess how feature selection affects algorithm accuracy, we compare the detection accuracy obtained from each of the feature ranking techniques. The study emphasizes how crucial accurate feature selection is in improving NIDS' accuracy and lowering false positives. The results also show that each algorithm gives the highest accuracy at varying numbers of features. Overall, this research offers insightful information about the efficacy of machine learning algorithms and feature selection techniques for network intrusion detection. The findings have significant ramifications for enhancing computer network security and defending against cyber threats.

**Keywords**—Network Intrusion Detection, UNSW-NB15, Machine Learning, Feature Selection, Cyber Security

## I. INTRODUCTION

The rising concerns of cybersecurity in today's digital environment have made robust network intrusion detection systems (NIDS) crucial. Machine learning algorithms have proven effective in detecting network breaches by analyzing traffic patterns. However, the selection of relevant features plays a vital role in the algorithm's effectiveness. Feature selection involves identifying and removing irrelevant and redundant features to enhance the algorithm's accuracy.

This study evaluates the performance of five popular machine learning techniques, namely Logistic Regression, Decision Trees, Random Forest, K-Nearest Neighbors, and Naive

Bayes, for network intrusion detection. The evaluation is conducted using the UNSW-NB15 [1] dataset, which resembles real network traffic and includes various features like packet sizes, durations, and protocols.

The study compares the effects of three feature selection algorithms (mutual information, chi-square, and ANOVA) on the accuracy of different machine learning algorithms. The number of selected features is varied from 10 to 41 to understand its impact on accuracy.

The goal of this research is to provide insights into the performance of network intrusion detection systems using different feature selection techniques and machine learning algorithms. The findings have the potential to improve computer network security, reduce false positives, and enhance the accuracy of network intrusion detection systems.

The paper is structured as follows: Section II reviews existing related work in network intrusion detection. Section III explains the proposed methodology in detail. Section IV presents the results obtained from the approach. Finally, Section V concludes the paper and discusses future directions for research.

## II. RELATED WORK

Intrusion detection systems (IDS) are used to identify unauthorized access, misuse, or malicious activities on computer networks. To evaluate the effectiveness of IDS, researchers rely on datasets to test their algorithms. Two well-known IDS datasets are KDD99 and UNSW-NB15 [1]. By lowering the dimensionality of the input data and choosing only the most pertinent features, feature selection is essential to the effectiveness of an IDS. Khammassi et al. (2017) [2], who chose features from the UNSW-NB15 and KDDCup99 datasets using the Genetic Algorithm (GA) and Logistic Regression (LR). The findings show that employing the GA-LR in combination with the DT on a smaller set of 20 features of the original list of 42 features present in the UNSW-NB15 [1] feature

space, they were able to achieve a detection score of 81.42% with a FAR of 6.39%. Whereas in the case of KDDCup99 dataset, employing the same algorithm on 18 features helped them achieve a detection score of 99.90% with a FAR rate of 0.105%.

Osanaiye et al. [3], worked on the detection of Distributed Denial of Service (DDoS) using a filter-based methods, where Information Gain, Chi-Square, and Gain Ratio, ReliefF were utilised as filtering techniques. According to the experimental findings, an accuracy of 99.67% and a FAR of 0.42% was achieved in this study using just 13 features out of the total 42 features when the DT classifier was used. .

Alazzam et al. [4], utilising the Pigeon Inspired Optimizer (PIO), put into practise a feature reduction technique for intrusion detection systems. This algorithm is inspired from white pigeon flying, and this is falls in the category of bio-inspired algorithm. In the KDDCup99 dataset 10 and 7 features were chosen by the Sigmoid and Cosine PIOs, respectively. While, in the NSL-KDD dataset 18 and 5 features were chosen by the Sigmoid and Cosine PIOs, respectively. In the case of UNSW-NB15 [1] dataset 14 and 5 features were chosen by the Sigmoid and Cosine PIOs, respectively. Sigmoid PIO was able to achieve an accuracy of 94.7%, 86.9% and 91.3% on the above three mentioned datasets, respectively. In comparison to this, the Cosine PIO was able to achieve an accuracy of 96.0%, 88.3% and 91.7% on the above three mentioned datasets, respectively.

Janarthanan et al. [5], developed several algorithms for feature selection using the UNSW-NB15 [1]. Two subsets were taken into consideration after many simulations. With only 8 important features, the first subset had an accuracy of 75.6617% and a Kappa value of 0.6891. Whereas comprising of only 5 important features, the second subset had an accuracy of 81.6175% and a Kappa value of 0.7639.

Kumar et al. [6], utilised the UNSW-NB15 [1] dataset validate an IDS that was put into action. An integrated rule-based model IDS was suggested by the authors in this work that used several Tree-based classifiers to carry out the classification. The Attack Accuracy (AAc) reached of 0.75 was used to assess the system's performance, while the SVM model earned training accuracy of 93.77%, FPR of 11.18%, and FM of 0.74. Owing of the higher accuracy, lower false positive and false negative rates of PSO and GA ove FO and GO, the authors of this paper concluded that PSO and GA outperformed FO and GO.

Khan et al. [7], made use of the RF algorithm to develop a feature reduction algorithm, and thus used it to calculate the Feature Importance score for all the features present in the UNSW-NB15 [1] dataset. The performance of multiple ML techniques was assessed using a feature subset of 11 qualities, with RF showing the best results.

Tama et al. [8], suggested a ensemble model developed over two levels for IDS using the Rotation Forest and Bagging algorithms. Also, Particle Swarm Optimization, Ant Colony Algorithm, and Genetic Algorithm were used to build a feature selection approach on the UNSW-NB15 [1] dataset.

They achieved high performance metrics of AC, precision, and sensitivity for binary classification using 10-fold cross-validation and hold-out method.

Hamid et al. [9] analysed a number of IDS datasets, comprising of the KDD99 dataset, the NSL-KDD dataset, the UNSW-NB15 [1] dataset, the center for applied internet data analysis dataset (CAIDA), the Australian defence force academy Linux dataset (ADFA-LD), and the University of New Mexico dataset (UNM). The research included a general overview of each dataset, with UNSW-NB15 [1] receiving more attention. To determine the accuracy, precision, and recall of all the datasets included in the study, the k-NN classifier was used. The NSL-KDD dataset, which has fewer redundant records distributed equally, showed the best results with the classifier, according to the results. Deep neural networks (DNN) did best on UNSW-NB15 [1] and attained accuracy above 90% for all datasets, according to a comparable assessment metric and F1 measure.

Binbusayyis et al. [10], used several feature-selection techniques, namely, the correlation measure (CFS), the consistency measure (CBF), the information gain (IG), and the distance measure (ReliefF) to evaluate the list of attributes present in the NSL-KDD and UNSW-NB15 [1] datasets. Furthermore to determine the training and testing performance, the features chosen using the techniques mentioned above, were then assessed using four classifiers, namely k-NN, random forests (RF), support vector machine (SVM), and deep belief network (DBN). In order to aid researchers in creating successful IDS, the work revealed the features selected using all the techniques for feature selection mentioned above and also the classification outcomes.

Rajagopal et al. [11], utilised a neural network to evaluate the importance of the features present in the UNSW-NB15 [1] dataset. According to their type, the authors divided the features into five groups: flow-based, content-based, time-based, essential, and supplementary features. 31 potential feature combinations were assessed from these groups. 39 features from the categorised categories were used to get the maximum accuracy (93%) possible. Additionally, a combination of 23 characteristics from the study were chosen using the SelectFromModel meta estimator, which chooses features depending on their ratings. The 23 chosen features had a greater accuracy (97%) than the previously listed 39 features.

Almomani et al. [12] compared the UNSW-NB15 [1] dataset's features with a few set of features that have previously been proposed earlier. They were assessed using supervised machine learning to determine classification performance. According to the study's findings, the existing feature vectors could be strengthened by making them smaller and enhancing them to handle encrypted traffic.

It is crucial to keep in mind that the UNSW-NB15 [1] and KDD99 datasets do not include vulnerabilities like SQL injection that are specific to cloud computing. A countermeasure was put forth to identify such attacks, specifically in a cloud-based setting, without the necessity for an application's source code, by Wu et al. [13].

Li et al. [14], performed experiments on the UNSW-NB15 [1] dataset to evaluate the performance of the proposed approach and used the sparse autoencoder in order to improve the rate of detection and lower the false positive rate in the classification discrepancies in intrusion detection.

Disha et al. [15], assessed UNSW-NB15 [1] dataset, using binary classification-based models. Gradient Boosting Tree, Decision Tree, Random Forest, and Multi-Layer Perceptron were among the models that were put to the test. They discovered that DT was the most effective classifier with the best accuracy and least False Positive Rate after using the Chi-Square test to eliminate characteristics which were independent of each other. response. Except for RF, all models' overall performance was enhanced through feature selection. Based on their research, it is determined that their proposed method is more accurate than other current ML approaches.

Yin et al. [16], presented a hybrid feature selection method known as IGRF-RFE, in order to increase the accuracy of detecting anomalies in multi-class network datasets utilising a multilayer perceptron, or MLP, network. The strategy successfully manages less significant features with high-frequency values and selects more pertinent features by combining a filtering technique which uses information gain and random forest and a wrapper method which utilises recursive feature elimination. The suggested strategy can help reduce the number of features from 42 to 23 and increase the multi-classification accuracy of multilayer perceptron from 82.25 to 84.24 percent, according to experimental findings using the UNSW-NB15 [1] dataset.

Zou et al. [17], introduced HC-DTTWSVM, which combines hierarchical clustering and decision tree twin support vector machines to efficiently identify various types of network intrusion. The technique uses a bottom-up merging strategy to maximise the separation of the decision tree's upper nodes, hence preventing the accumulation of errors. Twin SVMs are then integrated into the decision tree in order to determine the network intrusion classification top-down. The results of the proposed method's evaluation on the intrusion detection data sets NSL-KDD and UNSW-NB15 [1] show that it can efficiently detect numerous categories of network intrusion having performance that is comparable to that of other recently proposed methods.

In conclusion, the field of intrusion detection systems has seen significant advancements in recent years, thanks to the availability of various datasets and the development of sophisticated machine learning techniques. The KDD99 and UNSW-NB15 [1] datasets have played a crucial role in evaluating the effectiveness of IDSs, with the latter being preferred due to its more efficient features in terms of detection accuracy and false alarm rates. Additionally, researchers have proposed different feature selection methods and classifiers to improve the performance of IDSs.

Despite the significant progress made in this field, there is still much work to be done. The ever-evolving threat landscape requires continued research and development to ensure the effectiveness of IDSs. Furthermore, no single approach or

algorithm is universally optimal for all datasets and classifiers, making it necessary to explore the effectiveness and limitations of these techniques further.

Overall, the advancements in machine learning and feature selection techniques, combined with the availability of various datasets, provide researchers with the necessary tools to design effective IDSs. However, it is crucial to keep in mind that this is an ongoing process, and further research is necessary to address emerging threats and improve the effectiveness of IDSs.

### III. PROPOSED METHODOLOGY

In this study, we analyze the detection accuracy of various machine learning techniques for network intrusion detection both before and after feature selection. To accomplish this, we'll loop over choosing the top k features from a range of 10 to 41 using the feature selection algorithms of Mutual Information, Chi-Square, and Anova. The UNSW-NB15 [1] dataset will then be used to assess the effectiveness of Logistic Regression, Decision Trees, Random Forest, K-Nearest Neighbours, and Naive Bayes on the chosen features. We explain the various phases of the system design in the following subsections.

#### A. Dataset Explanation

The UNSW-NB15 [1] dataset is a benchmark dataset for network intrusion detection research that contains both legitimate and malicious traffic, with 2.5 million instances and 42 key features per instance. The dataset has only two labels, 0 for malware presence and 1 for malware absence, with a severe class imbalance that makes it difficult to accurately classify malware instances. The dataset has been used by the research community to evaluate the effectiveness of intrusion detection systems, revealing that many attacks are difficult to detect using conventional techniques, and serves as a useful baseline for assessing the performance of machine learning systems for network intrusion detection.

#### B. Data Preprocessing

Data preparation is a crucial step in getting the UNSW-NB15 [1] dataset ready for machine learning research. Machine learning methods often require numerical input, and the UNSW-NB15 [1] dataset contains both symbolic and numerical properties. So, in this study, we carried out two key preprocessing steps: data encoding and data normalization.

1) *Data Encoding for Symbolic Features:* The symbolic aspects of the dataset, such as the protocol (proto), service, and state, were initially encoded. Many machine-learning techniques call for encoding, which is the act of turning category data into numerical data. In this study, the symbolic features were encoded using the LabelEncoder function from the scikit-learn toolkit. To help the machine learning algorithms process the data, the LabelEncoder function gives each category a distinct integer value.



2) *Data Normalization*: Data normalization is an important step in preprocessing, which involves scaling numerical data to a common range to minimize the effects of variations in feature magnitudes. The scikit-learn library's StandardScaler function was used to normalize the UNSW-NB15 dataset's numerical characteristics, ensuring that all features had a comparable scale with a mean of 0 and a standard deviation of 1.

Data encoding and normalization were critical steps in preparing the UNSW-NB15 dataset for machine learning analysis. By transforming categorical features into numerical values and normalizing the numerical features, the dataset was prepared for feature selection and algorithm training. These steps are essential for developing an accurate and efficient network intrusion detection system.

### C. Detection Algorithm

The algorithm involved comparing the performance of five machine learning algorithms (Logistic Regression, Decision Trees, Random Forest, K-Nearest Neighbors, and Naive Bayes) on the UNSW-NB15 dataset before and after feature selection. Three feature selection algorithms (Mutual Information, Chi-Square, and Anova) were applied, and the dataset was preprocessed by handling missing values, removing duplicates, encoding symbolic features, and normalizing numerical features. The dataset was divided into training and testing sets using k-fold cross-validation, with k set to 10. The algorithms were run on the entire dataset without feature selection, and then with feature selection using all three algorithms. The accuracy of each algorithm was recorded and compared to determine the impact of feature selection on performance. A loop was run from k=10 to k=41 to determine the relationship between the number of selected features and algorithm accuracy, with the results presented in graphs. This algorithm provides a comprehensive approach to evaluating feature selection techniques for network intrusion detection. This above-mentioned procedure is summarized in Algorithm 1.

## IV. RESULTS AND DISCUSSION

In this study, three feature selection algorithms - ANOVA, Mutual Information, and Chi-Square - were used to evaluate their impact on the effectiveness of different machine learning models. The performance of these models was analyzed by varying the number of features selected using each algorithm. The details of the effect of each feature selection method on the models are discussed below.

### A. Results With ANOVA

Figure 1 shows the variation in the detection accuracy of each model by varying the number of selected features after using ANOVA. It was observed that the accuracies of Logistic Regression, Decision Tree, K-Nearest Neighbors, and Random Forest were either more than or equal to that of the initial accuracies when all 42 features were used. This means that the selected features did not significantly affect the performance of these models.

### Algorithm 1 Experimental Design for Feature Selection and Classification

---

```

procedure RUN_EXPERIMENT()
    Load preprocessed dataset into memory.
    for each feature selection algorithm do
        for k in range(10, 41) do
            Select the k best features from the training data.
            Transform the training and testing data to include
            only the selected features.
            for each classification model do
                Train the model on transformed training data.
                Predict the class labels of the testing data using
                the trained model.
                Compute the accuracy of the model
                Store the accuracy in a list.
            end for
            Compute the average accuracy across k-fold cross-
            validation for each model.
            Store the average accuracy for the current k and
            feature selection algorithm.
        end for
        Compute the mean and standard deviation of accuracies
        for each model and feature selection algorithm.
        Compare the accuracies before and after feature selec-
        tion for each algorithm.
    end for
end procedure

```

---

However, the accuracy of the Naive Bayes model showed a significant improvement after using ANOVA as the feature selection algorithm. The accuracy of the Naive Bayes model increased from 79% to 85% when the number of selected features was 23. Results with other machine learning models applied on ANOVA are summarized in Table I. This indicates that the selected features had a positive impact on the performance of the Naive Bayes model.

Overall, the experiment showed that feature selection can have a significant impact on the performance of classification models. In this case, ANOVA as a feature selection algorithm improved the performance of the Naive Bayes model but did not have a significant impact on the other models.

TABLE I  
ACCURACY OF DIFFERENT MODELS BEFORE AND AFTER FEATURE  
SELECTION USING ANOVA

Models	Before ANOVA	After ANOVA
Logistic Regression	89.95	89.90 (k = 36)
Decision Tree	93.74	93.82 (k = 34)
Naive Bayes	79.91	85.10 (k = 23)
K-Nearest Neighbors	91.68	92.26 (k = 23)
Random Forest	95.16	95.16 (k = 37)

### B. Results With Mutual Information

Figure 2 presents the analysis of the impact of mutual information as a feature selection algorithm on the accuracy

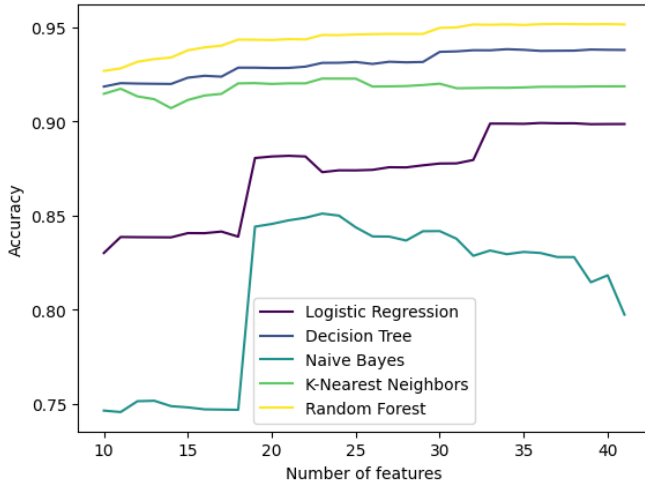


Fig. 1. Variation of accuracy of different models with number of features using ANOVA

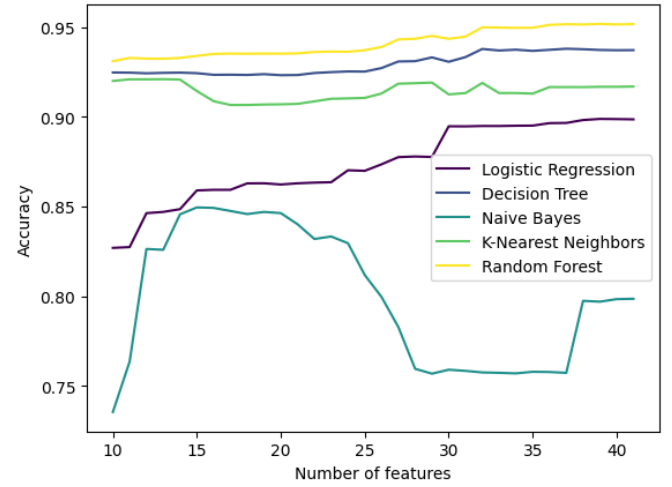


Fig. 2. Variation of accuracy of different models with number of features using Mutual Information

of different models. The results indicate that the accuracies of Logistic Regression, Decision Tree, K-Nearest Neighbors, and Random Forest were not significantly affected by the selected features when compared to the initial accuracy of using all 42 features. This suggests that the performance of these models was robust to the chosen features. However, it was observed that the accuracy of Random Forest, Decision Tree, and Logistic Regression kept on improving continuously as the number of selected features increased, whereas, for K-nearest neighbors, the accuracy dipped at 15 features and then increased steadily.

When it comes to Naive Bayes, the maximum accuracy was 84% when 15 features were selected using mutual information. However, there was a sudden dip in accuracy between 25 to 30 features, which remained constant until 37 features when the accuracy increased again. Interestingly, applying mutual information as a feature selection algorithm also led to a significant increase of 5% in the accuracy of Naive Bayes when compared to using all 42 features. Results with other machine learning models applied to Mutual Information are summarized in Table II. These findings indicate that mutual information is a useful feature selection algorithm that can enhance the performance of certain models, including Naive Bayes, by selecting relevant features.

TABLE II  
ACCURACY OF DIFFERENT MODELS BEFORE AND AFTER FEATURE SELECTION USING MUTUAL INFORMATION

Models	Before Mutual Information	After Mutual Information
Logistic Regression	89.95	89.88 (k = 39)
Decision Tree	93.74	93.79 (k = 37)
Naive Bayes	79.91	84.95 (k = 15)
K-Nearest Neighbors	91.68	92.09 (k = 14)
Random Forest	95.16	95.17 (k = 39)

### C. Results With Chi Square

In Figure 3, it can be observed that Logistic Regression, Decision Tree, and Random Forest models exhibit an increase in accuracy with an increase in the number of features selected using chi-square. However, the difference in accuracy before and after applying the chi-square is insignificant, and the maximum accuracy is reported before applying feature selection.

On the other hand, in K-Nearest Neighbors, there is a continuous increase in accuracy as the number of features varies from 10 to 22, and it remains almost constant at a maximum accuracy of 92% when  $k=27$ . After that, the accuracy decreases and remains almost constant. It is noteworthy that there is no significant difference in accuracy before and after applying the chi-square feature selection on K-Nearest Neighbors, despite variations in the accuracy.

In the case of Naive Bayes, there is a spike in accuracy at  $k=14$ , followed by a decrease until  $k=41$ . It is also evident that there is a significant increase in accuracy after applying the chi-square feature selection. Before applying the chi-square test, the accuracy was 79%, but after selecting 14 features using the feature selection, the accuracy increased to 85%. Results with other machine learning models applied on the Chi-Square test are shown in Table III.

To summarize, the results suggest that while the chi-square feature selection method can improve accuracy in some models, it may not be significant in others. Additionally, the number of features selected can also have an impact on the accuracy of the models.

## V. CONCLUSION AND FUTURE WORK

In conclusion, this study explored the impact of three different feature selection algorithms, ANOVA, mutual information, and linear regression, on the accuracy of five classification models: Logistic Regression, Decision Trees, Random Forests, K-Nearest Neighbors, and Naive Bayes. The results of the

TABLE III  
VARIATION OF ACCURACY OF DIFFERENT MODELS BEFORE AND AFTER  
FEATURE SELECTION USING CHI SQUARE

Models	Before Square	Chi After Square
Logistic Regression	89.95	89.58 (k = 40)
Decision Tree	93.74	93.91 (k = 32)
Naive Bayes	79.91	85.49 (k = 14)
K-Nearest Neighbors	91.68	92.03 (k = 27)
Random Forest	95.16	95.21 (k = 40)

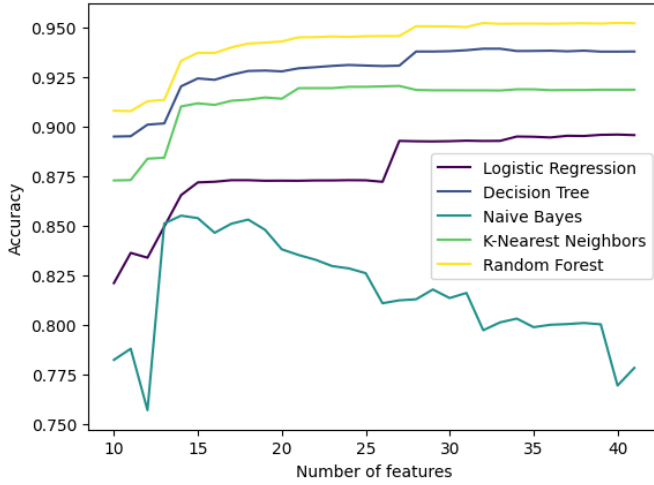


Fig. 3. Variation of accuracy of different models with number of features using Chi Square

experiments indicate that feature selection can have a significant impact on the performance of classification models, and the choice of feature selection algorithm can affect the performance of different models differently.

ANOVA was found to improve the performance of Naive Bayes while having little impact on the other models. Mutual information improved the performance of Naive Bayes and led to continuous improvement in the accuracy of logistic regression, decision trees, and random forest. However, mutual information had a dip in the accuracy of K-nearest neighbors at 15 features, after which accuracy improved steadily. Chi-square had a mixed effect on the models, with significant improvement in Naive Bayes and little impact on the other models.

Overall, the results suggest that feature selection can be used to improve the performance of classification models, and the choice of feature selection algorithm should be tailored to the specific problem and the models being used. Future work can explore the impact of other feature selection algorithms and combinations of algorithms on a wider range of classification models. Additionally, the study only considered a limited set of feature selection algorithms and models, and further investigation may be needed to determine the most effective approach for a particular problem. Finally, more research can be done to explore the impact of feature selection on

other types of machine-learning tasks, such as regression or clustering.

## REFERENCES

- [1] <https://research.unsw.edu.au/projects/unsw-nb15-dataset>
- [2] Khammassi, Chaouki and Saoussen Krichen. "A GA-LR wrapper approach for feature selection in network intrusion detection." *Comput. Secur.* 70 (2017): 255-277.
- [3] Osanaiye, O., Cai, H., Choo, K.K.R., Dehghantanha, A., Xu, Z., Dlodlo, M.. Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing. *J Wireless Com Network* 2016, 130 (2016). <https://doi.org/10.1186/s13638-016-0623-3>
- [4] Alazzam, Hadeel & Sharieh, Ahmad & Sabri, Khair Eddin. (2020). A Feature Selection Algorithm for Intrusion Detection System Based on Pigeon Inspired Optimizer. *Expert Systems with Applications*. 148. 113249. [10.1016/j.eswa.2020.113249](https://doi.org/10.1016/j.eswa.2020.113249).
- [5] T. Janarthanan and S. Zargari, "Feature selection in UNSW-NB15 and KDDCUP'99 datasets," 2017 IEEE 26th International Symposium on Industrial Electronics (ISIE), Edinburgh, UK, 2017, pp. 1881-1886, doi: 10.1109/ISIE.2017.8001537.
- [6] Kumar, V., Sinha, D., Das, A.K., Pandey, S.C., Goswami, R.T.. An integrated rule based intrusion detection system: analysis on UNSW-NB15 data set and the real time online dataset. *Cluster Comput* 23, 1397-1418 (2020). <https://doi.org/10.1007/s10586-019-03008-x>
- [7] Khan, N. & C, Nalina & Negi, Anjali & Thaseen, Sumaiya. (2020). Analysis on Improving the Performance of Machine Learning Models Using Feature Selection Technique. [10.1007/978-3-030-16660-1\\_7](https://doi.org/10.1007/978-3-030-16660-1_7).
- [8] B. A. Tama, M. Comuzzi and K. -H. Rhee, "TSE-IDS: A Two-Stage Classifier Ensemble for Intelligent Anomaly-Based Intrusion Detection System," in *IEEE Access*, vol. 7, pp. 94497-94507, 2019, doi: 10.1109/ACCESS.2019.2928048.
- [9] Hamid, Y.; Ranganathan, B.; Journaux, L.; Sugumaran, M. Benchmark Datasets for Network Intrusion Detection: A Review. *Int. J. Netw. Secur.* 2018, 20, 645-654.
- [10] Binbusayyis, Adel & Vaiyapuri, Thavavel. (2020). Comprehensive analysis and recommendation of feature evaluation measures for intrusion detection. *Heliyon*. 6. e04262. [10.1016/j.heliyon.2020.e04262](https://doi.org/10.1016/j.heliyon.2020.e04262).
- [11] Rajagopal, S.; Hareesha, K.S.; Kundapur, P.P. Feature Relevance Analysis and Feature Reduction of UNSW NB-15 Using Neural Networks on MAMLS. In *Advanced Computing and Intelligent Engineering-Proceedings of ICACIE 2018*; Pati, B., Panigrahi, C.R., Buyya, R., Li, K.-C., Eds.; *Advances in Intelligent Systems and Computing*; Springer: Paris, France, 2020; pp. 321-332.
- [12] Almomani, O. A Feature Selection Model for Network Intrusion Detection System Based on PSO, GWO, FFA and GA Algorithms. *Symmetry* 2020, 12, 1046.
- [13] Wu, T.; Chen, C.; Sun, X.; Liu, S.; Lin, J. A Countermeasure to SQL Injection Attack for Cloud Environment. *Wirel. Pers. Commun.* 2017, 96, 5279-5293.
- [14] Y. Li, P. Gao and Z. Wu, "Intrusion Detection Method Based on Sparse Autoencoder," 2021 3rd International Conference on Computer Communication and the Internet (ICCCI), Nagoya, Japan, 2021, pp. 63-68, doi: 10.1109/ICCCI51764.2021.9486776.
- [15] R. A. Disha and S. Waheed, "A Comparative study of machine learning models for Network Intrusion Detection System using UNSW-NB 15 dataset," 2021 International Conference on Electronics, Communications and Information Technology (ICECIT), Khulna, Bangladesh.
- [16] Yin, Y., Jang-Jaccard, J., Xu, W. et al. IGRF-RFE: a hybrid feature selection method for MLP-based network intrusion detection on UNSW-NB15 dataset. *J Big Data* 10, 15 (2023). <https://doi.org/10.1186/s40537-023-00694-8>
- [17] L. Zou, X. Luo, Y. Zhang, X. Yang and X. Wang, "HC-DTTSVM: A Network Intrusion Detection Method Based on Decision Tree Twin Support Vector Machine and Hierarchical Clustering," in *IEEE Access*, vol. 11, pp. 21404-21416, 2023, doi: 10.1109/ACCESS.2023.3251354.
- [18] Brownlee, J. (2020). Mutual Information for Machine Learning. *Machine Learning Mastery*. <https://machinelearningmastery.com/information-gain-and-mutual-information/>
- [19] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. John Wiley & Sons.
- [20] Agresti, A. (2018). *An Introduction to Categorical Data Analysis* (3rd ed.). John Wiley & Sons.

# Investigation on the impact of elevated temperature on sustainable geopolymer composite

Manvendra Verma<sup>1</sup>, Rahul Kumar Meena<sup>2</sup>, Indrajeet Singh<sup>3</sup>,  
Nakul Gupta<sup>1</sup> , Kuldeep K Saxena<sup>4</sup> , M Madhusudhan Reddy<sup>5</sup>,  
Karrar Hazim Salem<sup>6</sup> and Ummal Salmaan<sup>7</sup>

## Abstract

Geopolymer concrete (GPC) is an eco-friendly, sustainable, cementless and green concrete. It could be an alternative to the conventional concrete. In alkaline circumstances, the alumina and silica concentration in geopolymer concrete creates the geopolymer bond, while regular concrete creates C-S-H (calcium silicate hydrate bond). The final result of the geopolymer bond does not include any water. At elevated temperatures, geopolymer concrete would thus be more stable. Due to its greater strength and durability quality, geopolymer concrete may be the ideal replacement for ordinary portland cement (OPC) concrete. This research intends to examine how specimens of geopolymer concrete and regular concrete respond to exposure to increased temperatures between 100°C and 800°C. Mass loss, ultrasonic pulse velocity, compressive strength, X-ray diffraction, thermogravimetric analysis and derivative thermogravimetric analysis were all examined throughout the experimental examination. Both concrete specimens lose mass or weight as the exposure temperature rises; OPC concrete samples spalls at 600°C, while GPC sample fail at 800°C. GPC specimens lose around 12% of their original mass after being exposed to temperatures of 800°C, while OPC specimens lose about 7%. The GPC specimens maintained 60% of their initial compressive strength after being exposed to a temperature of 700°C, but the OPC concrete specimens only kept 52%. With each increase in exposure to extreme temperatures, the peaks of quartz and cristobalite are lowered. Only the form or structure of the mineral oxide would change; the chemical linkages would remain. The GPC samples subjected to temperatures of 100°C exhibit effective thermal stability than all other specimens exposed to extreme temperatures. As the exposure temperature rises, the GPC specimens become more thermally stable. According to the experimental findings, the GPC specimens' bonding structure makes them more resistant to high temperatures than regular concrete specimens. Micropores are present in the voids of the geopolymer matrix, while mesopores and micropores are present in the voids of the OPC matrix. While OPC bonding is C-S-H formed by the hydration of lime and silica contained in the cement, the geopolymer bonding did not include the water content in the final or end result of geopolymerisation for strengthening.

<sup>1</sup>Department of Civil Engineering, GLA University, Mathura, Uttar Pradesh, India

<sup>2</sup>Department of Civil Engineering, Punjab Engineering College, Chandigarh, Punjab, India

<sup>3</sup>Department of Civil Engineering, Delhi Technological University, Delhi, India

<sup>4</sup>Division of Research and Development, Lovely Professional University, Phagwara, Jalandhar, Punjab, India

<sup>5</sup>Department of Civil Engineering, Institute of Aeronautical Engineering, Hyderabad, Telangana, India

<sup>6</sup>Al-Mustaqbal University College, Hillah, Babil, Iraq

<sup>7</sup>Department of Automotive Engineering, Aksum University, Aksum, Tigray, Ethiopia

## Corresponding author:

Ummal Salmaan, Department of Automotive Engineering, Aksum University, Aksum, Tigray 1010, Ethiopia.

Email: ummalsalmaan90@gmail.com



Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (<https://creativecommons.org/licenses/by/4.0/>) which permits any use, reproduction and distribution of the work

without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

## Keywords

Higher temperature, mass loss, geopolymer concrete, compressive strength, durability

Date received: 9 May 2023; accepted: 31 July 2023

Handling Editor: Chenhui Liang

## Introduction

Geopolymer concrete would be an example of sustainable construction. Concrete is a primary building material around the globe for many decades, in which OPC works as a binder material in the mix. High energy utilises production of Portland cement and causes significant carbon dioxide emissions due to burning large quantities of fuel and calcareous breakdown. The carbon dioxide emission during the production of cement contributes around 8% of total carbon dioxide emissions.<sup>1</sup> In future years, the carbon emissions increase due to the increment in the demand for concrete. So, GPC could be an alternative to conventional concrete because it utilises pozzolans like GGBFS, flyash, calcined clay, rice husk ash, etc. as a binding material for the replacement of cement.<sup>2,3</sup> Geopolymer is a name of the chemical bond which were initially developed by Prof. Davidovits in 1991. This novel bond is a three-dimensional (3D) inorganic polymeric matter produced using an alkaline solution including sodium silicate and sodium hydroxide that reacts with any substance rich in silica and alumina.<sup>4–8</sup> GPC is an environment-friendly inorganic polymers that are produced by the synthesis action of aluminosilicate sources in alkaline and hydrothermal conditions.<sup>9</sup> In contrast to Portland cement, geopolymer concrete's manufacturing cuts energy and produces less greenhouse gases because it utilises industrial waste and by-products as binding.<sup>10</sup> As an engineering binder material, geopolymers may also be used to create geopolymer mortars and concrete.<sup>11–14</sup>

Due to its extensive spectrum of chemical components and reactions, geopolymer concrete exhibits several characteristics that are common to concrete.<sup>15–17</sup> Geopolymer concrete has a growing concern in many areas as an alternative construction material to ordinary concrete in the context of environmental preservation and sustainable development.<sup>18,19</sup> Fly ash and slag are the key raw materials used in the production of geopolymers. In recent years, there has been a lot of interest in the properties of geopolymer paste formed from blast furnace slag or fly ash.<sup>20–31</sup> Currently, there is less research publication on the high-temperature performance of geopolymer concrete. Concrete buildings should be resistant to high-temperature impacts due to natural or human-related incidents over their service lifetime. The increase in temperature leads to a sequence

of physicochemical changes that are not reversible, resulting in many internal defects like cracks and pores.<sup>32</sup> These defects might lead to concrete degradation and cause local damage or collapse in the structure. The characteristics and kind of binder vary between the GPC and OPC concrete. Calcium silicate hydrates are formed via the hydration processes of calcium oxide and silicon dioxide in the OPC binder. Furthermore, the GPC binder is made by alkali activating aluminosilicate raw materials, then polymerised in a high pH environment and hydrothermal circumstances at fairly low temperatures to produce the reaction result. This chemical structure gives GPC materials advantages over their OPC concrete. It also improved the strength performance against exposure to high temperatures.<sup>33</sup> Previous research on the thermal characteristics of flyash geopolymers found that these materials had little thermal shrinkage and maintained their strength after being exposed to high temperatures.<sup>34–37</sup> The flexibility of geopolymer mortars after 800°C exposure has a significant influence on their residual strength.<sup>38</sup> Alkali-activated metakaolin with  $\text{Na}_2\text{SiO}_3$  and NaOH solutions creates a lightweight geopolymer mortar that can unite broken shale, haydite sand and withstand fire.<sup>39</sup> After being exposed to a high temperature of 950°C for 30 min, the lightweight geopolymer mortars that had been cured at 20°C for 28 days indicated a strength loss of up to 63% of their original strength.<sup>40</sup>

It offers a scientific basis for the destruction or remediation of fireproof structures made with geopolymer concrete and helps to generalise the use of geopolymer concrete.<sup>41–45</sup> The non-destructive identification and deterioration of fire-damaged buildings are always utilised for ultrasonic testing.<sup>46</sup> Geopolymers are highly fire-resistant due to their inorganic structure, and they have exceptional thermal resistance with little gel structure deterioration identification up to 700°C–800°C. Therefore, the GPC may be a replacement for OPC concrete in the building of high-risk infrastructures. It is important to realise that geopolymeric materials change their compressive strength after heat exposure.<sup>47,48</sup> Gel structures analyse the GPC's pore structure network, which manages moisture at high temperatures with lesser porosity, improving fire resistance by releasing trapped water vapour.<sup>49,50</sup> Particle size distribution, fly ash content, the kind of alkaline solution used and the exposure temperature are all

factors that may affect the reaction rate, chemical composition and microstructure of a GPC that is based on fly ash.<sup>51–59</sup> The GPC's main reaction is an alkaline silicon tetrahedron alloy arranged into three dimensions.<sup>60–64</sup> Recent study has concentrated on the following topics: (1) A study of the mechanical properties of geopolymers at high temperatures, (2) investigating the effectiveness of different FA-based geopolymer and geopolymer composite types, as well as their performance at elevated temperatures and (3) researching the details of the unique behaviours of GPC and OPC concrete in fires.<sup>38,55,65–67</sup>

The geopolymer mortar gains or loses strength after being exposed to high temperatures.<sup>68</sup> After conducting the experiments, the authors concluded that sintering or additional geopolymerisation at a higher temperature has the impact of increasing strength, and the second is thermal incompatibility damage.<sup>9,69</sup> The smaller degree of damage was shown, the greater the frequency of retention components. The small frequency components would otherwise prevail over the observed frequency range.<sup>70,71</sup> Therefore, by modifying the recorded ultrasonic signals in the frequency area, you may indirectly demonstrate high-temperature damage to concrete.<sup>72,73</sup>

## Research significance

This study is to determine the thermal stability of GPC and ordinary concrete to elevated temperatures and validate them with predict models by other authors and code. Firstly, cast the GPC and ordinary concrete cube samples and check the compressive strength UPV of that. Both concrete samples were kept in the muffle furnace for 2 h from 100°C to 800°C and found that the ordinary concrete samples fail at 600°C, whereas the GPC samples fail at 800°C. Continuously, check the mass loss, UPV and compressive strength after the exposed temperature. After the destructive test, matrix powers were extracted from the samples and checked the microstructure analysis by the XRD and TGA tests. Validate the results of the test with the model equation givens by the various authors. After the study, it could be concluded that the GPC is highly stable than ordinary concrete.

## Experimental programme

The experimental programme includes the properties of the materials, mixing, casting, curing and test setups.

### Materials

Cement, fine aggregates, coarse aggregates, binding materials (Sodium silicate and sodium hydroxide), water and superplasticiser are all included in the materials section. The materials used to create the OPC

concrete and GPC specimens for the experimental analysis are introduced in the next paragraph.

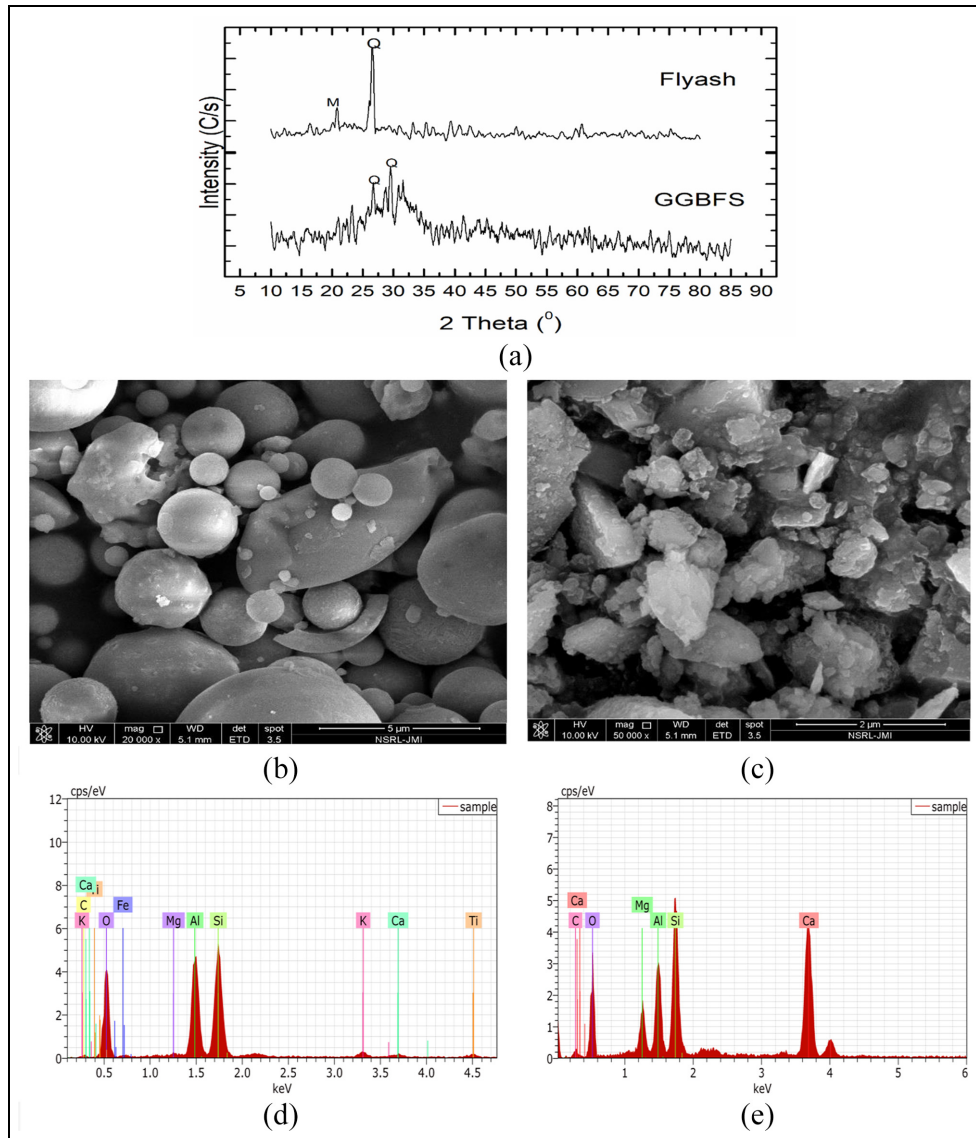
To determine the properties preliminary tests were performed on material samples at the DTU Concrete Laboratory, Civil Engineering Department. The materials were subjected to SEM and EDS testing at Jamia Millia Islamia's nanotechnology lab, while the XRD test was performed at DTU, Delhi. JK Cement OPC 43 grade was purchased from the market and its qualities were tested in accordance with Indian standards.<sup>74</sup> All basic properties of cement are checked in the laboratory for quality check as per the Indian standard. Flyash were collected from the thermal power plant because it is a solid waste of the thermal power industry.

During the experimental study, the GPC mix design includes class c flyash. In Figure 1(b), the SEM image of flyash reveals the spherical and porous character of the particles. Figure 1(a) shows fly ash amorphous character. Table 1 lists the chemical components identified in the flyash by XRF analysis. The elemental composition of flyash was determined by the Jamia Millia Islamia Nanotechnology Laboratory in New Delhi, India, as seen in Figure 1(d). Slag is the waste material found in iron ore and steel ore after the steel industry's rubbish has been removed. In a steel manufacturing facility, molten iron steel is used to remove slag. The GGBFS composition also comprises silica and alumina in substantial amounts. GGBFS was brought in from the Bhilai steel factory in order to test and create GPC mix specimens. On exhibit in Table 2 is the mineral oxides of GGBFS, and on display in Figure 1(a) is the GGBFS XRD graph, which illustrates the amorphous nature of samples. SEM image of the GGBFS sample is shown in Figure 1(c), which illustrates the irregular shape of the particles at 2 microns, and Figure 1(e) depicts the EDS graph, which illustrates the components in the samples. Figure 2 depicts the flyash and GGBFS gradation curve.

In order to get the pozzolanic materials ready for the geopolymerisation process, an alkaline solution is applied to them, which produces a geopolymer link. Prior to mixing and casting, an alkaline solution must be produced (20–24 h). The ratio of sodium hydroxide to sodium silicate solution was exact. Fisher Scientific chemicals Pvt. Ltd. supplied samples of sodium hydroxide flakes, while CDH Pvt. Ltd. supplied sodium silicate. The mix design sample of sodium hydroxide flakes is shown in Figure 3(a), whereas the sodium silicate sample is illustrated in Figure 3(b).

The aggregate acts as the skeleton of the concrete, comprising up to 85% of its overall volume. Coarse and fine aggregates are the most common kinds of aggregates used in the concrete mix. The design mix contained two sizes of coarse aggregates: 10 and 20 mm. As fine aggregates in the mix design, crushed stone or stone dust were used. Prior to incorporating the aggregate into the





**Figure 1.** (a) Graph of XRD, (b) flyash SEM image, (c) GGBFS SEM image, (d) flyash EDS graph and (e) GGBFS EDS graph.

**Table 1.** GGBFS and flyash mineral oxides.

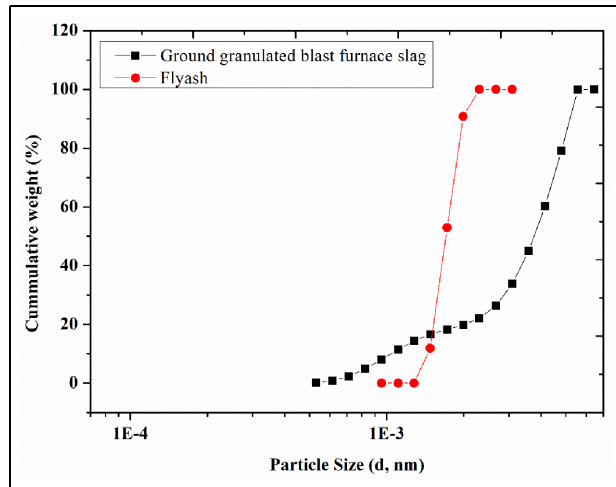
Mineral oxides	GGBFS (%)	Flyash (%)
SiO <sub>2</sub>	34.52	45.8
Al <sub>2</sub> O <sub>3</sub>	20.66	21.4
CaO	32.43	13.7
Fe <sub>2</sub> O <sub>3</sub>	0.57	12.6
MgO	10.09	1.3
SO <sub>3</sub>	0.77	1.9
LOI	0.3	0.1

mix design, its quality was evaluated in accordance with Indian standards, which included a check of its grading, fineness modulus, size, water absorption, specific gravity, silt content, soundness, crushing, impact, abrasion value, flakiness and elongation index. Figure 3(d) provides an

**Table 2.** Mix designs.

Constituents	OPC concrete mix content (kg/m <sup>3</sup> )	GPC mix content (kg/m <sup>3</sup> )
OPC	370	00
Flyash	00	303.75
GGBFS	00	101.25
NaOH	00	40.5
Na <sub>2</sub> SiO <sub>3</sub>	00	101.25
Fine aggregate	683	683
Coarse aggregate	1289	1269
Water	148	40.5
Superplasticiser	3.7	4.05
Total	2493.7	2543.7

illustration of the use of stone dust in the blended design. The sand quality was identified in the laboratory test for



**Figure 2.** GGBFS and flyash cumulative particle size.

concrete. The coarse and fine aggregate gradation curves are shown in Figure 4, which demonstrates that both aggregates are properly graded.

All concrete compositions included local raw materials and resources. The fraction of aggregates in the sample was used to determine the particle sizes, and the fineness modulus of the coarse aggregate was computed. Figure 3(e) illustrates samples of coarse aggregate. Sieve examination reveals that the fineness modulus of the coarse aggregate is 7.29, as indicated by the sample. In

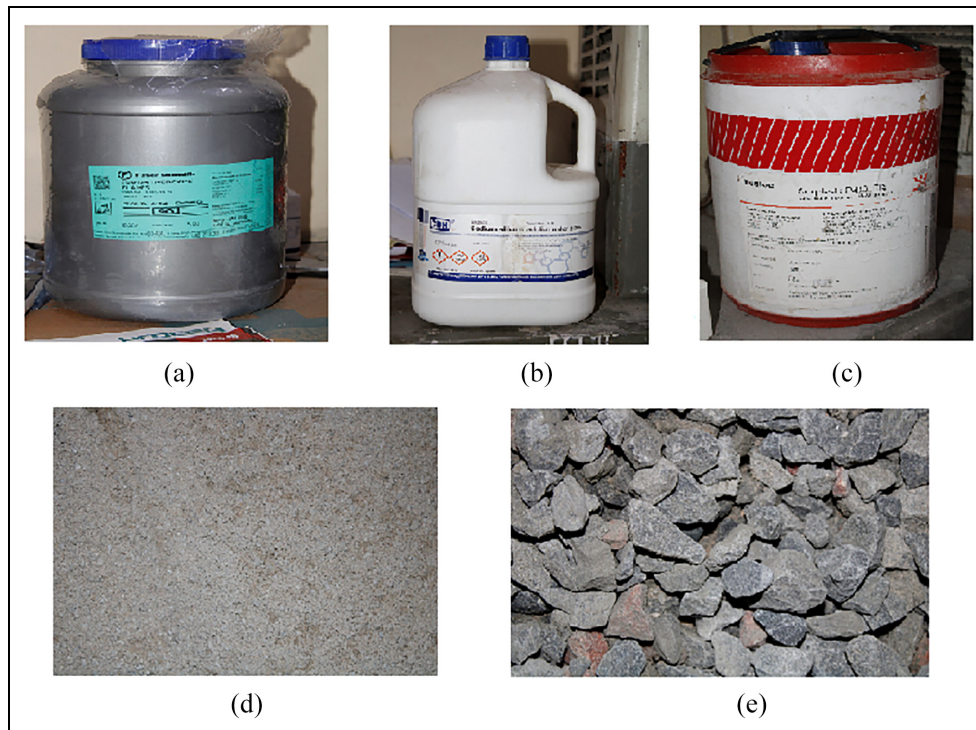
the hybrid design, the SNF-based superplasticizer SP Conplast 430 was used. The superplasticizer enhances the workability of concrete mix by reducing the amount of water while simultaneously improving its strength. Figure 3(c) depicts the superplasticizer used during the testing.

### Mixing, casting and curing

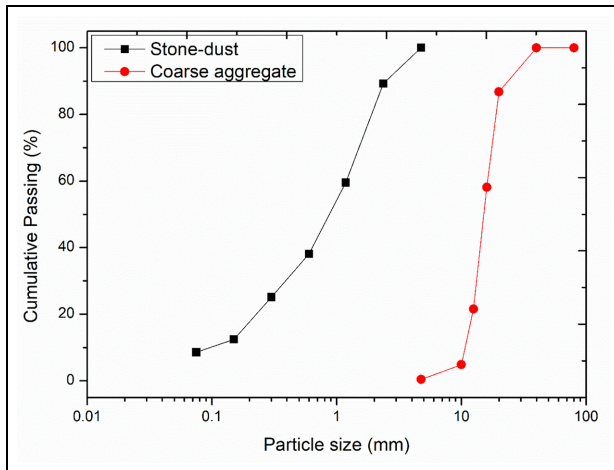
Both GPC and OPC concrete were mixed for 2–5 min in the pan mixture before being poured into 150 mm × 150 mm × 150 mm cube moulds. The GPC alkaline solution was made before 20 h of mixing, while OPC concrete components are used directly without any preparation. Both samples were cast in similar moulds, however, the conventional concrete specimens were cured in a water tank, while the GPC specimens were cured in an oven for 24 h at 600°C. Table 2 displays the total amount of GPC and OPC concrete used to create specimens for testing.

### Test setups

Initial samples consisted of cube-shaped concrete mix design specimens and tests for compressive strength, density or mass, as well as UPVT for future reference. The specimens of reinforced concrete were placed in the muffle furnace to expose them to increased temperatures, and the impact of elevated temperatures on the



**Figure 3.** Picture of raw GPC materials: (a) Sodium Hydroxide, (b) Sodium Silicate, (c) Superplasticiser, (d) Stone Dust and (e) Coarse Aggregates.



**Figure 4.** Aggregates gradation curve.

specimens' compressive strength, mass loss and UPV was evaluated. The GPC matrix was also subjected to microstructural investigation, which included XRD and TGA testing.

Both cubes of concrete were subjected to severe temperatures between 100°C and 800°C. The cube samples were heated for 2 h in the muffle furnace at a predetermined temperature, with the temperature rising from room temperature to a certain raised temperature at a rate of 10°C/min and cooling randomly to the ambient temperature of the surrounding environment. Figure 5(a) shows the variation of temperature exposure on concrete samples, whereas Figure 5(b) depicts the muffle furnace used in this examination at raised temperatures. After being exposed to high temperatures and allowed to cool, the specimens were determined for strength weight loss.

The XRD and thermogravimetric examinations provided the microstructural analysis of the GPC samples.

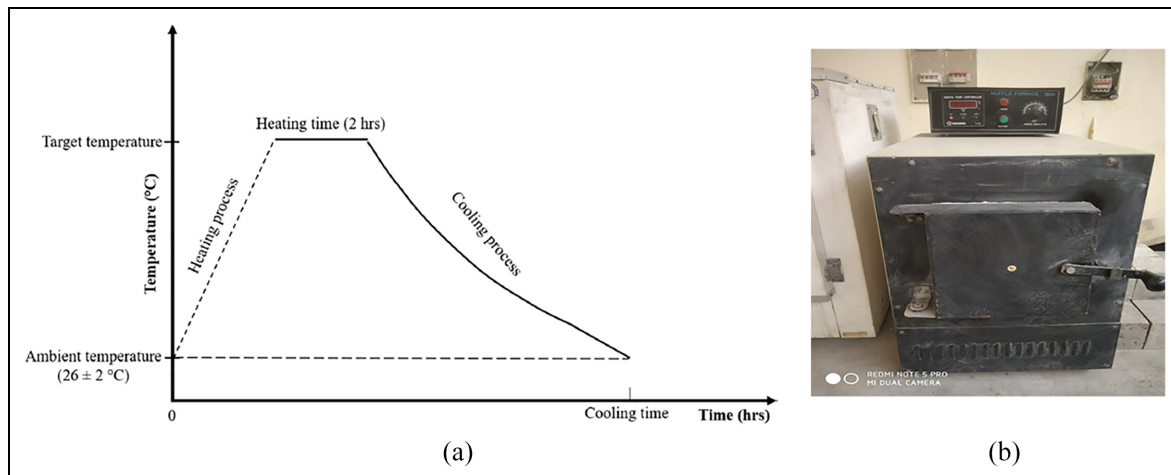
The XRD test revealed the intensity of mineral oxides in the GPC sample.<sup>75</sup> The TGA study revealed that the samples had their thermal stability up to 850°C. The sample could be weighed sequentially at different temperature stages while the temperature increases by 10°C/min.<sup>76</sup>

## Results and discussion

In the cube moulds, samples of both GPC and normal concrete were cast. Before the compressive strength test, density and UPVT were measured 28 days after the casting of the specimen and before its compression. Figure 6(a) illustrates the relationship between compressive strength and concrete age for both concrete samples. The compressive strength of the GPC and standard concrete specimens at 28 days is 35 and 36.3 MPa, respectively. In 28-day tests, the density and UPVs of the GPC specimen are 2476 kg/m<sup>3</sup> and 4492 m/s, while the corresponding values for conventional concrete are 2490 kg/m<sup>3</sup>, and 4520 m/s. The fluctuations of the UPV of both concrete specimens are illustrated in Figure 6(b), which demonstrates that the UPV of OPC concrete is larger than that of GPC concrete at all ages except the 3-day test.

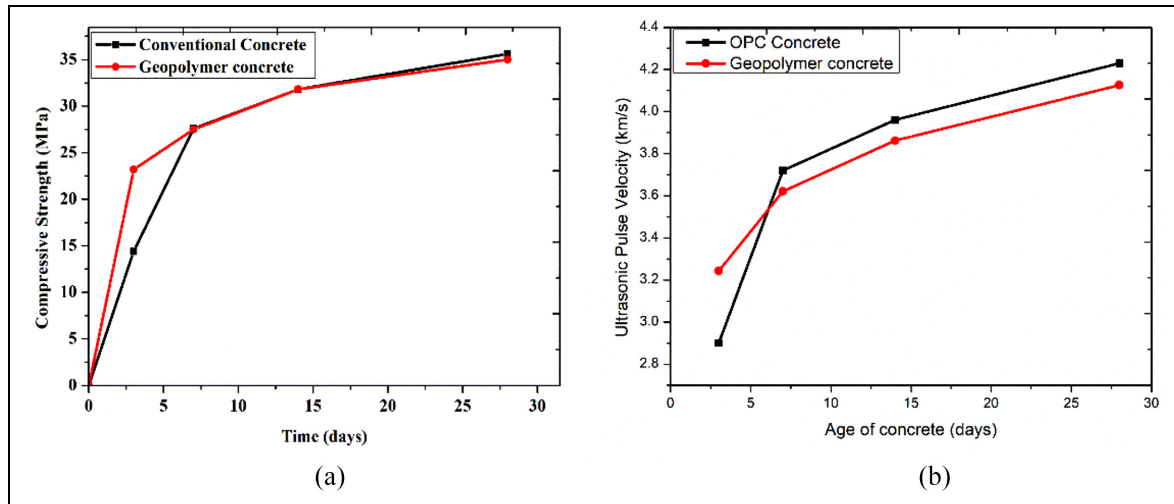
### Effect of elevated temperature

Muffle furnace was used for concrete sample temperature exposure, which allowed for total control over the temperature's stability as well as the pace at which the temperature increased during time. GPC and OPC concrete samples were put into the muffle furnace for a period of 2 h at a precise controlled elevated temperature that ranged from 100°C to 800°C, with a rate of temperature rise of 10°C/min. After being subjected to high temperatures, the specimens were investigated in



**Figure 5.** (a) Picture of the heating process and (b) picture of muffle furnace.





**Figure 6.** (a) Graph of compressive strength and (b) graph of UPV variation of both concrete.

terms of the amount of mass lost, the UPV and the compressive strength.

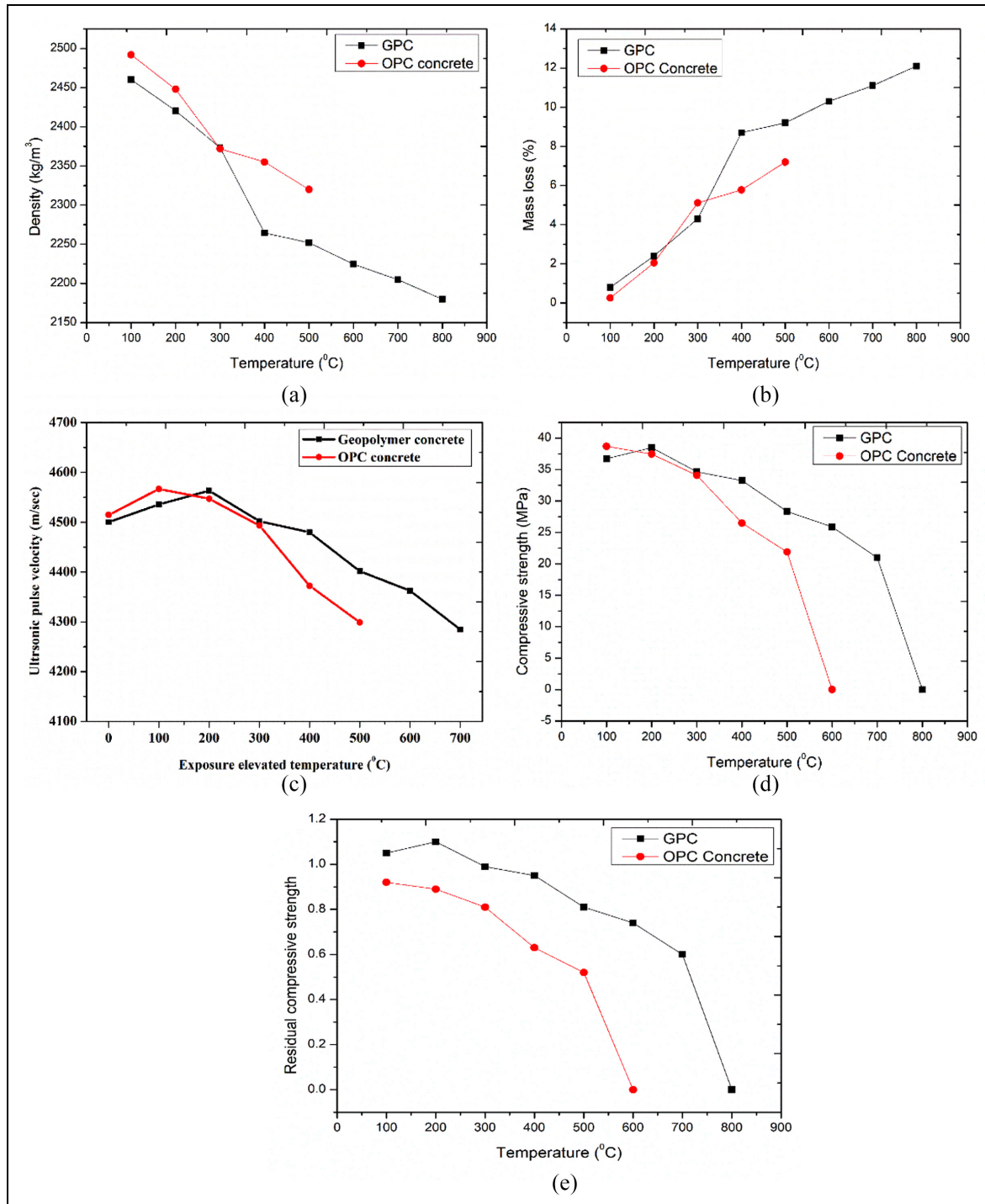
**Mass loss.** The OPC concrete specimens fail at a temperature of 600°C, while the GPC specimens fail at a temperature of roughly 800°C. Both concrete mix specimens weight reduces as the exposure temperature rises. After being subjected to increased temperatures from 100°C to 800°C by interval of 100°C, the density variation of both concrete concrete mix samples is depicted in Figure 7(a). Both concrete samples mass loss varies as per exposed temperature ranging from 100°C to 800°C is depicted in Figure 7(b). In contrast to OPC concrete samples, which fail at 600°C, this figure shows that both concrete mix mass loss which is rise with the exposed temperature. The OPC concrete specimens can withstand elevated temperatures up to 800°C. After being heated to 800°C, GPC specimens lose around 12% of their initial weight, while an OPC concrete sample will fail at 600°C and will demonstrate a mass loss of 7% after exposure to 500°C. It has been determined that GPC specimens are less likely to degrade when exposed to high temperatures.

The high-temperature hits cause a large amount of damage to the geopolymer's solid matrix, and the increasing temperatures fasten the development of cracks and loss their strength, both of which eventually result in the formation of voids in the matrix. When the temperature increases, a process known as dehydration starts. During this process, moisture flows on the surface of the specimens and then exits. This results in interior destruction due to the microstructure of samples as well as a loss of weight in the geopolymer matrix. During the early stages of heating, the geopolymer sample experiences a rapid loss in weight owing to the presence of free water and structured water.

**Ultrasonic pulse velocity test (UPVT).** The UPVT is a method for determining the quality of materials that does not involve any damaging processes. It does this by using ultrasonic pulse waves. Figure 7(c) is a graph that shows the variation of UPV of both concrete mix samples with exposed temperature. This graph demonstrates that the GPC specimens UPV first rises up to a temperature of 200°C, whereas the UPV of OPC concrete samples initially increases up to a temperature of 100°C when subjected to the same temperature. Following this point, the UPV of both concrete specimens will continue to decrease with each succeeding increase in the exposure temperature. The UPVs of the GPC specimens are noticeably lower than those of the normal concrete specimens during the first exposure to 100°C, although they attain high values. The examples of OPC concrete failed at 600°C, but the GPC samples were failed at a temperature of 800°C.

Both the sample's porous structure and the amount of water that evaporates from the matrix rise as the exposed temperature increased. As a consequence of the loss of mass, there are now more voids. As a direct consequence of their being more vacancies, the ultrasonic pulse velocity would be decreased. In addition, the production of microfractures was hastened by an increase in temperature, which resulted in a decrease in the density of the composites. The time it takes for ultrasonic velocity waves to propagate has been lengthened, which has led to decreased UPV levels. The melting point of GPC fibre matrix and the formation of tiny channels both occurred above a temperature of 300°C, which contributed to the discovery of lower UPV values.

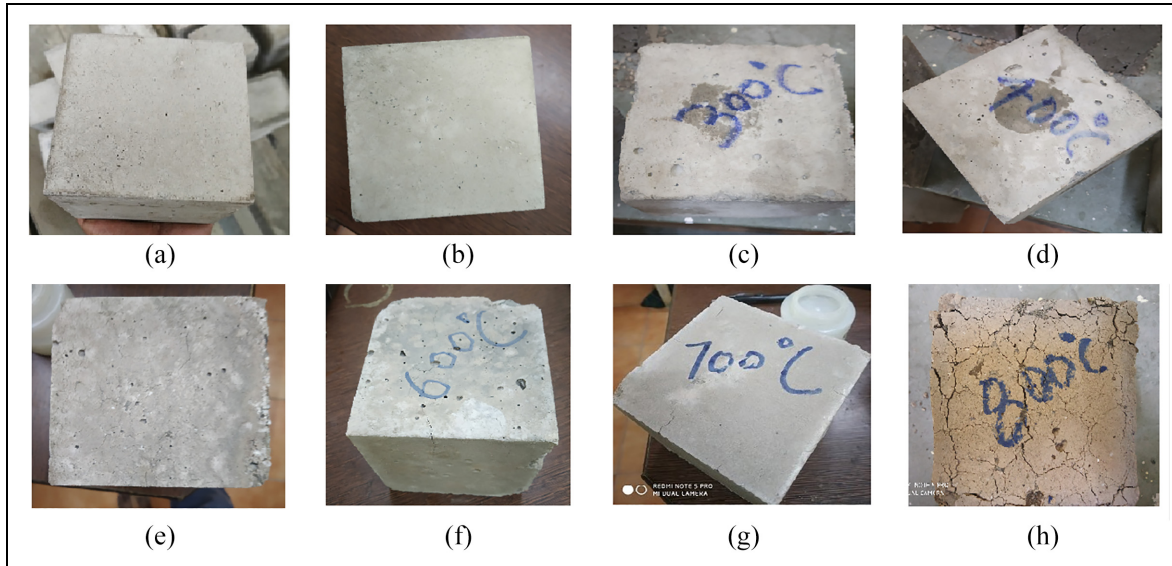
**Compressive strength.** The specimens were exposed to 100°C, 200°C, 300°C, 400°C, 500°C, 600°C, 700°C and



**Figure 7.** (a) Graph of density variation, (b) graph of mass loss variation, (c) graph of UPV variation, (d) graph of compressive strength variation and (e) graph of residual compressive strength variation with the temperature.

800°C, and the compressive strength of OPC concrete and GPC were determined after exposure. Figure 7(d) illustrates the variability in compressive strength that occurred in each of the concrete specimens after being subjected to higher temperatures. At a temperature of 100°C, the compressive strength of OPC concrete samples are higher than the GPC samples, but after

reaching that temperature, it begins to drop. The GPC samples were strengthened up to 200°C temperature exposure, but beyond that point, they begin to disintegrate continuously. The fluctuation in the residual strength of both concrete mix samples that were exposed to higher temperatures is shown graphically in Figure 7(e)'s graph. The GPC specimens and the



**Figure 8.** Picture of high temperature exposed cubes: (a) 100°C, (b) 200°C, (c) 300°C, (d) 400°C, (e) 500°C, (f) 600°C, (g) 700°C and (h) 800°C.

normal concrete specimens both failed when exposed to temperatures of 600°C and 800°C, respectively. After being subjected to a temperature of 700°C, the GPC specimens maintained their initial compressive strength of 60%, while the OPC concrete samples were maintained their compressive strength of 52% after being subjected to a temperature of 500°C. In light of this, GPC samples are superior than regular concrete specimens in terms of their resistance to high temperatures.

Due to water evaporation and dehydration of the geopolymer matrix, melting of the bonding between the matrices, and thermal exposure in response to free water evaporation, the strength of the geopolymer matrix was drastically reduced between 600°C and 900°C. The strength of mortar and matrix are higher than the GPC made with lightweight particles (LWAGC). The porous structure of LWA concrete mix matrix would initiate to weaken the strength of the samples, it was expected that the strength of LWA concrete would be substantially lower than geopolymer paste and mortar. This was the case because LWA has a lower strength. Due to high temperature exposure, the geopolymer mortar sample saw a greater decrease in its flexural strength than in its compressive strength. At elevated temperatures, the flexural strength was more vulnerable to the formation of microstructural defects, such as the propagation of fractures and the formation of porous structures.<sup>77–82</sup> Losses in compressive strength are varied from 31% to 85% in fibreless geopolymer specimens, but losses in polypropylene fibrous geopolymer specimens from 32% to 86%. This rate has risen considerably in comparison to samples lacking fibres. Losses in flexural strength in fibreless

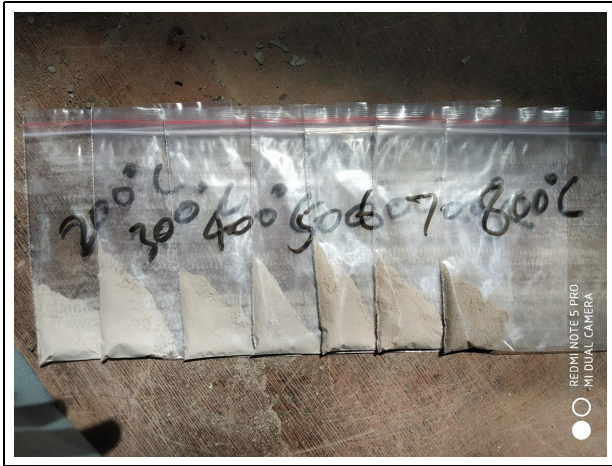
geopolymer samples were between 49% and 84%, but the rate of flexural strength losses in polypropylene fibrous samples was lower, ranging from 43.91% to 84.60%. Both flexural strength and toughness may be significantly improved by including fibres into cement or geopolymer mortar. It has been established that the addition of polymer fibres to a geopolymer composite may significantly increase both the strength and toughness of the composite.<sup>83–86</sup> Various polymer fibres are common kinds of materials that are used to reinforce the geopolymer matrix.

**Visual inspection.** Concrete cubes that were heated in a muffle furnace for 2 h at a higher temperature were subjected to a visual evaluation to determine the outcome of the experiment. The cube GPC specimens are shown in Figure 8 after being subjected to elevated temperatures ranging from 100°C to 800°C. This demonstrates that the GPC specimens fail at temperatures higher than 800°C.<sup>87–91</sup> Figure 9 shows the geopolymer matrix powder after being subjected to exposed elevated temperatures. This illustration demonstrates that the geopolymer matrix becomes darker after being subjected to higher exposure temperatures. As a result of the evaporation of hygroscopic water contained in the GPC matrix, cracks appear when the exposure temperature is raised over a certain point.

### Microstructural analysis

The XRD and TGA-DTG analyses were tested on the GPC sample after elevated temperatures for microstructural analysis.





**Figure 9.** Powder geopolymer matrix picture after elevated temperature exposure from 100°C to 800°C.

**XRD analysis.** The XRD analysis checks the intensity of mineral oxides in the sample as per ASTM C1365–18. It shows the crystalline behaviour of the mineral particles present in the sample. The XRD tests were conducted after elevated exposure of the GPC samples from 100°C to 800°C. Figure 10(a) shows the XRD graphs of the GPC specimens, which show that the quartz and cristobalite peaks are reduced with the increment of elevated temperature exposure. The quartz and cristobalite are silica content, but changes in their shapes, where quartz is hexagonal and cristobalite is tetragonal in shape. The other components show a negligible presence in the crystalline peaks. Most silica contents are present in the GPC paste matrix, which shows the higher thermal stability of the GPC samples to elevated temperatures. The XRD graph shows the mostly amorphous nature of the geopolymer paste. After the XRD analysis, it could be concluded that the GPC matrix is highly stable to the elevated. It would change the mineral oxide's shape or structure only but not the chemical bonds. The geopolymer bonds are highly stable up to 800°C.<sup>92–97</sup>

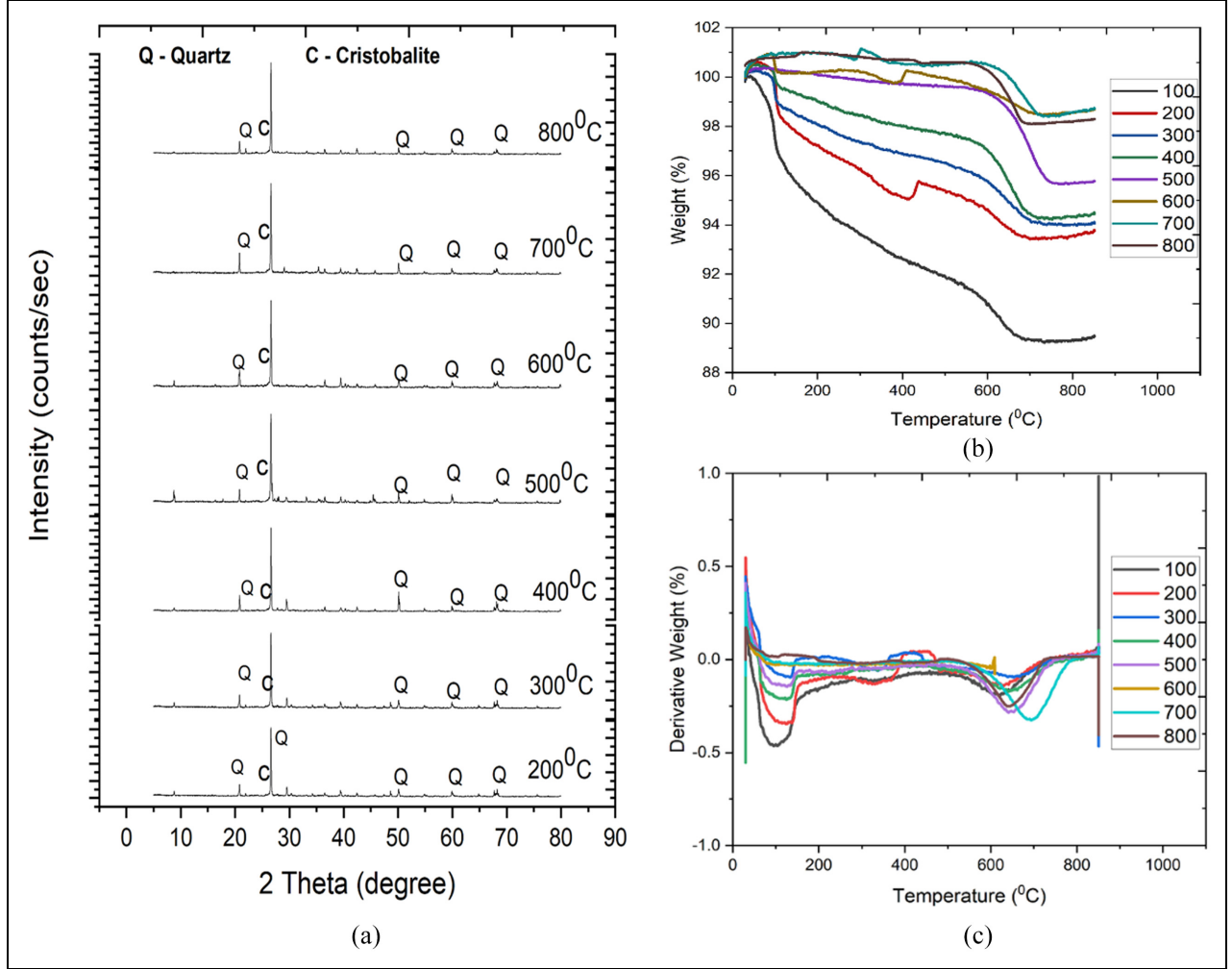
Because of all of these reasons, the number of pores in synthetic geopolymers was significantly reduced. The acceleration of pore volume reduction and water loss in the geopolymer as a result of evaporation and dihydroxylation that occurs during a thermal assault can lead to the formation of structural faults or the expansion of existing faults. These processes can also contribute to the loss of strength that occurs after a thermal assault. Additionally, the production of new crystal phases as a consequence of uncontrolled diffusion is the reason of the observed loss in sample strength. These new crystal phases produced the observed loss in sample strength. In spite of this, as shown by XRD, the

fundamental structure of traditional metakaolin-based geopolymers was preserved, and the geopolymer materials revealed reduced microstructural degeneration at increasing temperatures, which resulted in less strength loss.

**Thermogravimetric analysis.** The TGA test is used to find the stability of samples to elevated temperatures. It works based on the weight of the sample continuously with an increment of the temperature at the rate of 10°C/min. Figure 10(b) shows the TGA (thermogravimetric analysis) graph of the GPC samples exposed to elevated temperatures. The GPC samples were exposed to elevated temperatures from 100°C to 800°C at intervals of 100°C for 1 h. The 100°C exposed GPC samples show less thermal stability than all other elevated temperature exposed specimens. The thermal stability of the GPC specimens increases with the elevated temperature. The GPC specimens exposed to the elevated didn't show linear behaviour; they varied randomly. The 100°C exposed specimens show a mass loss of around 10%–11%, whereas, after 600°C, exposed specimens show a negligible mass loss of around 1%–2%. After the experimental analysis, it could be concluded that the GPC matrix is highly stable up to 800°C because it retains more than 90% mass up to 800°C.

The weight loss of GPC samples in the TGA test due to the evaporation of water absorbed water in the matrix from 25°C to 225°C.<sup>98</sup> This could be due to the nano clay filling the space and resulting in denser matrices. The physical free water would evaporate between 225°C and 525°C, the rate of evaporation or water loss would be slow for all specimens. It is possible that the de-hydroxylation of the silicon hydroxyl group is to blame for the constant weight loss of the silicon oxygen group as well as the water that has evaporated. The inadequate weight reduction that occurred between 500°C and 700°C was caused by the excessive amount of coal and fly ash that was burned.<sup>99–101</sup> This may be seen in DTG curves over 600°C, when a little bump indicates a slight change in weight loss.<sup>102,103</sup>

The DTG analysis shows the regularity or linearity of the stability of the material against the temperature elevation. It clearly shows the negative or positive loss of the sample's mass. Figure 10(c) shows the DTG graph of the GPC samples that were exposed to various elevated temperatures. All GPC samples after the elevated temperature exposure show a similar pattern. The graph shows that the initiative goes positive and after going negative side randomly, describes its gain initially and loss mass randomly after gaining. The mass loss rate increases up to around 160°C–180°C, and the mass-loss rate decreases continuously in all types of specimens.



**Figure 10.** (a) XRD graph after elevated temperature, (b) thermogravimetric analysis graph of the GPC and (c) DTG graph of the GPC.

#### Validation of experimental values with the proposed model

A model was developed to calculate the compressive strength variation with exposure to fire or elevated temperatures for the OPC concrete specimens.<sup>104</sup> Equation (1) shows the effect of fire temperature exposure on concrete specimens in terms of strength.

$$f_{CT} = f_C \times (1 - 0.001 \times T), T \leq 500^\circ\text{C} \quad (1a)$$

$$f_{CT} = f_C \times (1.375 - 0.00175 \times T), 500^\circ\text{C} \leq T \leq 700^\circ\text{C} \quad (1b)$$

$$f_{CT} = 0, T > 700^\circ\text{C} \quad (1c)$$

Eurocode recommended equation (2) for demonstrating the effect of fire on reinforced concrete members.<sup>105</sup> Equations (2a)–(2c) describes the correlation between compressive strength and elevated temperature variation.

$$f_{CT} = f_C, T \leq 100^\circ\text{C} \quad (2a)$$

$$f_{CT} = f_C \times (1.067 - 0.00067 \times T) \geq 0, 100^\circ\text{C} \leq T \leq 400^\circ\text{C} \quad (2b)$$

$$f_{CT} = f_C \times (1.44 - 0.0016 \times T) \geq 0, T \geq 400^\circ\text{C} \quad (2c)$$

Equation (3) was developed for reinforced concrete elements exposed to elevated temperatures,<sup>106</sup> and it was also considered by Zha (2003)<sup>107</sup> to create a nonlinear finite element programme to calculate the resistivity of the concrete element against elevated temperatures.

$$f_{CT} = f_C \times \left( 2.011 - 2.353 \times \frac{T - 20}{1000} \right) \leq f_C \quad (3)$$

**Table 3.** Comparisons of experimental values with the proposed model of compressive strength.

Temp. °C	$f_c$ (MPa)	Predicted values (MPa)				Experimental values (MPa)	
		Lin et al. <sup>104</sup>	Eurocode	Han et al. <sup>106</sup>	Li and Purkiss <sup>108</sup>	GPC	OPC concrete
Ambient	35	35	35	35	35	35	38.3
100	35	31.5	35	35	34.95	36.75	38.7
200	35	28	32.655	35	33.08	38.5	37.45
300	35	24.5	30.31	35	29.80	34.65	34.1
400	35	21	27.965	35	25.47	33.25	26.5
500	35	17.5	22.4	30.85	20.41	28.35	21.9
600	35	11.375	16.8	22.62	14.99	25.9	Fail
700	35	5.25	11.2	14.38	9.55	21	–
800	35	0	5.6	6.15	4.44	Fail	–

A correlation was developed in terms of equation (4) between stress-strain constitutive equations at higher temperatures.<sup>108</sup>

$$f_{CT} = f_c \times \left( 0.00165 \times \left( \frac{T}{100} \right)^3 - 0.03 \times \left( \frac{T}{100} \right)^2 + 0.025 \times \left( \frac{T}{100} \right) + 1.002 \right) \quad (4)$$

where  $f_{CT}$  is the residual compressive strength at  $T$  temperature in MPa,  $f_c$  are the characteristics of the compressive strength of concrete in MPa and  $T$  is the temperature in °C.

Table 3 shows the comparison between equations (1)–(4) and the experimental results. It describes the effects of elevated temperatures on both types of concrete tested in the laboratory and discusses the equations given by various codes or authors. This table describes the strength values as per the author's and euro code equations and validates them with the value of the experimental results. Conventional concrete sample degradation is somehow similar to the author's model equation, but the GPC results didn't match the given models.

## Conclusion

In the experimental investigation in the laboratory, the GPC and OPC concrete samples were exhibited at elevated temperatures to analyse the mass loss and strength loss of the specimens. In microstructural analysis, check the XRD, TGA and DTG of the exposed GPC specimens. Based on the experimental investigation, the following conclusion is reached:

- The fact that the weight loss of both concrete samples increased after being subjected to a variety of elevated temperatures ranging from 100°C to 800°C reveals that the mass loss of both concrete samples increases as the temperatures

climbs. The OPC concrete specimens collapse at a temperature of 600°C, whereas the GPC specimens collapse after being exposed to a temperature of 800°C in a muffle furnace for a period of 2 h. In the case of GPC specimens, the mass loss before failure was around 12% of the total, but in the case of standard concrete samples, the mass loss before failure was approximately 7%. It may be deduced from this that the GPC samples are less prone to deformation at higher temperatures. After being subjected to a temperature of 700°C, the GPC samples were maintained 60% of their strength, whereas the OPC concrete specimens kept 52% of their compressive strength after being subjected to a temperature of 500°C. As a direct consequence of this, GPC specimens demonstrate a higher level of tolerance to high temperatures.

- The peaks of quartz and cristobalite are reduced with the increment of elevated temperature exposure. The quartz and cristobalite are silica content, but changes in their shapes, where quartz is hexagonal and cristobalite is tetragonal in shape. The other components show a negligible presence in the crystalline peaks. Most silica contents are present in the GPC paste matrix, which shows the higher thermal stability of the GPC samples to elevated temperatures. The 100°C exposed GPC samples show less thermal stability than all other elevated temperature exposed specimens. The thermal stability of the GPC specimens increases with the elevated temperature. Geopolymer bonds did not need water in the end product of the reaction or geopolymerisation to develop hardening. The GPC specimens are higher stable to thermal attack than the ordinary concrete due to the presence of maximum micropores in the geopolymer matrix, whereas the OPC matrix has mesopores at a large amount.

## Author contribution

All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for valuable intellectual content; and (c) approval of the final version.

## Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is supported by the Civil Engineering Department, GLA University, Mathura, Uttar Pradesh, India.

## Compliance with ethical standard

This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.


## Consent to participate


As a corresponding author or on behalf of all authors of the research paper, I consent to participate.

## Consent to publication

All author of the research paper consent to the publication.

## ORCID iDs

Nakul Gupta  <https://orcid.org/0000-0001-9525-8482>

Kuldeep K Saxena  <https://orcid.org/0000-0003-4064-5113>

## Data availability

Data would be available on request.

## References

- Ishak SA and Hashim H. Low carbon measures for cement plant – a review. *J Clean Prod* 2015; 103: 260–274.
- Kong DLY and Sanjayan JG. Damage behavior of geopolymer composites exposed to elevated temperatures. *Cem Concr Compos* 2008; 30: 986–991.
- Verma M and Dev N. Geopolymer concrete: a way of sustainable construction. *Int J Recent Res Asp* 2018; 5: 201–205.
- Davidovits J. Geopolymers and geopolymeric materials. *J Therm Anal* 1989; 35: 429–441.
- Davidovits J. 30 years of successes and failures in geopolymer applications. Market trends and potential breakthroughs. In: *Geopolymer 2002 conference*, Melbourne, VIC, Australia, 28–29 October 2002, pp.1–16.
- Davidovits J. Geopolymers: inorganic polymeric new materials. *J Therm Anal* 1991; 37: 1633–1656.
- Davidovits J. *Geopolymer chemistry & applications*. Saint-Quentin, France: Geopolymer Institute, 2015.
- Komnitsas K and Zaharaki D. Geopolymerisation: a review and prospects for the minerals industry. *Miner Eng* 2007; 20: 1261–1277.
- Pan Z, Sanjayan JG and Rangan BV. An investigation of the mechanisms for strength gain or loss of geopolymer mortar after exposure to elevated temperature. *J Mater Sci* 2009; 44: 1873–1880.
- Kong DLY and Sanjayan JG. Effect of elevated temperatures on geopolymer paste, mortar and concrete. *Cem Concr Res* 2010; 40: 334–339.
- Turk J, Cotič Z, Mladenović A, et al. Environmental evaluation of green concretes versus conventional concrete by means of LCA. *Waste Manag* 2015; 45: 194–205.
- El-Gamal SMA and Selim FA. Utilization of some industrial wastes for eco-friendly cement production. *Sustain Mater Technol* 2017; 12: 9–17.
- Shang J, Dai JG, Zhao TJ, et al. Alternation of traditional cement mortars using fly ash-based geopolymer mortars modified by slag. *J Clean Prod* 2018; 203: 746–756.
- Pavithra P, Srinivasula Reddy M, Dinakar P, et al. A mix design procedure for geopolymer concrete with fly ash. *J Clean Prod* 2016; 133: 117–125.
- Aly T and Sanjayan JG. Effect of pore-size distribution on shrinkage of concretes. *J Mater Civ Eng* 2010; 22: 525–532.
- Lloyd NA and Rangan BV. Geopolymer concrete with fly ash. In: Zachar J, Claisse P, Naik TR, et al. (eds) *Second international conference on sustainable construction materials and technologies*. Ancona, Italy: UWM Center for By-Products Utilization, 2010, pp.1493–1504.
- Sagoe-crentsil K. Role of oxide ratios on engineering performance of fly-ash geopolymer binder systems. *Am Ceram Soc* 2009; 20: 175–183.
- Chandrasekhar Reddy K. Investigation of mechanical and microstructural properties of fiber-reinforced geopolymer concrete with GGBFS and metakaolin: novel raw material for geopolymerisation. *Silicon* 2021; 13: 4565–4573.
- Nuaklong P, Jongvivatsakul P, Pothisiri T, et al. Influence of rice husk ash on mechanical properties and fire resistance of recycled aggregate high-calcium fly ash geopolymer concrete. *J Clean Prod* 2020; 252: 119797.
- Ismail I, Bernal SA, Provis JL, et al. Modification of phase evolution in alkali-activated blast furnace slag by the incorporation of fly ash. *Cem Concr Compos* 2014; 45: 125–135.
- Ismail I, Bernal SA, Provis JL, et al. Microstructural changes in alkali activated fly ash/slag geopolymers with sulfate exposure. *Mater Struct* 2013; 46: 361–373.
- Kim HK, Jeon JH and Lee HK. Workability, and mechanical, acoustic and thermal properties of lightweight aggregate concrete with a high volume of entrained air. *Constr Build Mater* 2012; 29: 193–200.
- Yang K-H, Song J-K, Lee K-S, et al. Flow and compressive strength of alkali-activated mortars. *ACI Mater J* 2009; 106: 50–58.

24. Deb PS and Sarker PK. Effects of ultrafine fly ash on setting, strength, and porosity of geopolymers cured at room temperature. *J Mater Civ Eng* 2017; 29: 06016021.
25. Mustakim SM, Das SK, Mishra J, et al. Improvement in fresh, mechanical and microstructural properties of fly ash-blast furnace slag based geopolymer concrete by addition of nano and micro silica. *Silicon* 2021; 13: 2415–2428.
26. Kumar S, Kumar R and Mehrotra SP. Influence of granulated blast furnace slag on the reaction, structure and properties of fly ash based geopolymer. *J Mater Sci* 2010; 45: 607–615.
27. Nath P and Sarker PK. Effect of GGBFS on setting, workability and early strength properties of fly ash geopolymer concrete cured in ambient condition. *Constr Build Mater* 2014; 66: 163–171.
28. Deb PS, Nath P and Sarker PK. The effects of ground granulated blast-furnace slag blending with fly ash and activator content on the workability and strength properties of geopolymer concrete cured at ambient temperature. *Mater Des* 2014; 62: 32–39.
29. Praveen Kumar VV and Dey S. Influence of metakaolin on strength and durability characteristics of ground granulated blast furnace slag based geopolymer concrete. *Struct Concr* 2020; 21: 1040–1050.
30. Lee NK and Lee HK. Setting and mechanical properties of alkali-activated fly ash/slag concrete manufactured at room temperature. *Constr Build Mater* 2013; 47: 1201–1209.
31. Ryu GS, Lee YB, Koh KT, et al. The mechanical properties of fly ash-based geopolymer concrete with alkaline activators. *Constr Build Mater* 2013; 47: 409–418.
32. Arioiz O. Effects of elevated temperatures on properties of concrete. *Fire Saf J* 2007; 42: 516–522.
33. Zhao R and Sanjayan JG. Geopolymer and Portland cement concretes in simulated fire. *Mag Concr Res* 2011; 63: 163–173.
34. Bakharev T. Thermal behaviour of geopolymers prepared using class F fly ash and elevated temperature curing. *Cem Concr Res* 2006; 36: 1134–1147.
35. Dombrowski K, Buchwald A and Weil M. The influence of calcium content on the structure and thermal performance of fly ash based geopolymers. *J Mater Sci* 2007; 42: 3033–3043.
36. Kong DLY, Sanjayan JG and Sagoe-Crentsil K. Comparative performance of geopolymers made with metakaolin and fly ash after exposure to elevated temperatures. *Cem Concr Res* 2007; 37: 1583–1589.
37. Kong DLY, Sanjayan JG and Sagoe-Crentsil K. Factors affecting the performance of metakaolin geopolymers exposed to elevated temperatures. *J Mater Sci* 2008; 43: 824–831.
38. Pan Z and Sanjayan JG. Stress–strain behaviour and abrupt loss of stiffness of geopolymer at elevated temperatures. *Cem Concr Compos* 2010; 32: 657–664.
39. Xu H and Van Deventer JSJ. Geopolymerisation of multiple minerals. *Miner Eng* 2002; 15: 1131–1139.
40. Hu SG, Wu J, Yang W, et al. Preparation and properties of geopolymer-lightweight aggregate refractory concrete. *J Central South Univ Technol* 2009; 16: 914–918.
41. Sharma U, Gupta N and Verma M. Prediction of the compressive strength of flyash and GGBS incorporated geopolymer concrete using artificial neural network. *Asian J Civil Eng*. Epub ahead of print 1 May 2023. DOI: 10.1007/s42107-023-00678-2.
42. Kumar R, Verma M and Dev N. Investigation on the effect of seawater condition, sulphate attack, acid attack, freeze–thaw condition, and wetting–drying on the geopolymer concrete. *Iran J Sci Technol Trans Civil Eng* 2022; 46: 2823–2853.
43. Upreti K, Verma M, Agrawal M, et al. Prediction of mechanical strength by using an artificial neural network and random forest algorithm. *J Nanomat* 2022; 2022: 1–12.
44. Chouksey A, Verma M, Dev N, et al. An investigation on the effect of curing conditions on the mechanical and microstructural properties of the geopolymer concrete. *Mater Res Express* 2022; 9: 055003.
45. Upreti K and Verma M. Prediction of compressive strength of geopolymer concrete using artificial neural network. *J Eng Res Appl* 2022; 1: 24–32.
46. Uysal M. Self-compacting concrete incorporating filler additives: performance at high temperatures. *Constr Build Mater* 2012; 26: 701–706.
47. Duxson P, Lukey GC and van Deventer JSJ. Thermal evolution of metakaolin geopolymers: part 1 – physical evolution. *J Non-Cryst Solids* 2006; 352: 5541–5555.
48. Aslani F. Thermal performance modeling of geopolymer concrete. *J Mater Civ Eng* 2016; 28: 040150621.
49. Cao VD, Pilehvar S, Salas-Bringas C, et al. Microencapsulated phase change materials for enhancing the thermal performance of Portland cement concrete and geopolymer concrete for passive building applications. *Energy Convers Manag* 2017; 133: 56–66.
50. Barbosa VFF and MacKenzie KJD. Synthesis and thermal behaviour of potassium sialate geopolymers. *Mater Lett* 2003; 57: 1477–1482.
51. Verma M and Dev N. Effect of ground granulated blast furnace slag and fly ash ratio and the curing conditions on the mechanical properties of geopolymer concrete. *Struct Concr* 2022; 23: 2015–2029.
52. Verma M and Dev N. Effect of liquid to binder ratio and curing temperature on the engineering properties of the geopolymer concrete. *Silicon* 2022; 14: 1743–1757.
53. Verma M and Dev N. Effect of SNF-based superplasticizer on physical, mechanical and thermal properties of the geopolymer concrete. *Silicon* 2022; 14: 965–975.
54. Verma M and Dev N. Sodium hydroxide effect on the mechanical properties of flyash-slag based geopolymer concrete. *Struct Concr* 2021; 22: E368–E379.
55. Verma M and Dev N. Review on the effect of different parameters on behavior of geopolymer concrete. *Int J Innov Res Sci Eng Technol* 2017; 6: 11276–11281.
56. Krivenko PV and Kovalchuk GY. Directed synthesis of alkaline aluminosilicate minerals in a geocement matrix. *J Mater Sci* 2007; 42: 2944–2952.
57. Assi L, Carter K, Deaver E, et al. Sustainable concrete: building a greener future. *J Clean Prod* 2018; 198: 1641–1651.
58. Yunsheng Z, Wei S, Zongjin L, et al. Impact properties of geopolymer based extrudates incorporated with fly ash and PVA short fiber. *Constr Build Mater* 2008; 22: 370–383.



59. Puertas F, Santos H, Palacios M, et al. Polycarboxylate superplasticiser admixtures: effect on hydration, micro-structure and rheological behaviour in cement pastes. *Adv Cem Res* 2005; 17: 77–89.
60. Fernández-Jiménez A. Engineering properties of alkali-activated fly ash concrete. *ACI Mater J* 2006; 103: 106–112.
61. Criado M, Fernández-Jiménez A, Palomo A, et al. Effect of the  $\text{SiO}_2/\text{Na}_2\text{O}$  ratio on the alkali activation of fly ash. Part II:  $^{29}\text{Si}$  MAS-NMR survey. *Microporous Mesoporous Mater* 2008; 109: 525–534.
62. Duxson P, Fernández-Jiménez A, Provis JL, et al. Geopolymer technology: the current state of the art. *J Mater Sci Technol* 2007; 42: 2917–2933.
63. Criado M, Fernández-Jiménez A, de la Torre AG, et al. An XRD study of the effect of the  $\text{SiO}_2/\text{Na}_2\text{O}$  ratio on the alkali activation of fly ash. *Cem Concr Res* 2007; 37: 671–679.
64. Bernal SA, Mejía de, Gutiérrez R and Provis JL. Engineering and durability properties of concretes based on alkali-activated granulated blast furnace slag/metakaolin blends. *Constr Build Mater* 2012; 33: 99–108.
65. Pan Z, Sanjayan JG and Kong DLY. Effect of aggregate size on spalling of geopolymer and Portland cement concretes subjected to elevated temperatures. *Constr Build Mater* 2012; 36: 365–372.
66. Rickard WDA, Temuujin J and van Riessen A. Thermal analysis of geopolymer pastes synthesised from five fly ashes of variable composition. *J Non-Cryst Solids* 2012; 358: 1830–1839.
67. Abdulkareem OA, Mustafa AI, Bakri AM, Kamarudin H, et al. Effects of elevated temperatures on the thermal behavior and mechanical performance of fly ash geopolymer paste, mortar and lightweight concrete. *Constr Build Mater* 2014; 50: 377–387.
68. Yaraswini K and Venkateshwara Rao A. Behaviour of geopolymer concrete at elevated temperature. *Mater Today Proc* 2020; 33: 239–244.
69. Yim HJ, Kim JH, Park SJ, et al. Characterization of thermally damaged concrete using a nonlinear ultrasonic method. *Cem Concr Res* 2012; 42: 1438–1446.
70. Zhao J, Trindade ACC, Liebscher M, et al. A review of the role of elevated temperatures on the mechanical properties of fiber-reinforced geopolymer (FRG) composites. *Cem Concr Compos* 2023; 137: 104885.
71. Bidgoli MR and Saeidifar M. Time-dependent buckling analysis of  $\text{SiO}_2$  nanoparticles reinforced concrete columns exposed to fire. *Comput Concr* 2017; 20: 119–127.
72. Ren W, Xu J and Bai E. Strength and ultrasonic characteristics of alkali-activated fly ash-slag geopolymer concrete after exposure to elevated temperatures. *J Mater Civ Eng* 2016; 28: 1–8.
73. Khater HM. Studying the effect of thermal and acid exposure on alkali-activated slag geopolymer. *Adv Cem Res* 2014; 26: 1–9.
74. IS 8112 1989. 43 Grade ordinary Portland cement specification. 1997.
75. ASTM D934 13. Identification of crystalline compounds in water-formed deposits by X-ray diffraction. 2014.
76. ASTM C1872 18. Standard test method for TGA of hydraulic cement. 2009.
77. Verma M. Prediction of compressive strength of geopolymer concrete using random forest machine and deep learning. *Asian J Civ Eng*. Epub ahead of print 26 April 2023. DOI: 10.1007/s42107-023-00670-w.
78. Sharma U, Gupta N and Verma M. Prediction of compressive strength of GGBFS and flyash-based geopolymer composite by linear regression, lasso regression, and ridge regression. *Asian J Civ Eng*. Epub ahead of print 29 May 2023. DOI: 10.1007/s42107-023-00721-2.
79. Verma M, Upreti K, Khan MR, et al. Prediction of compressive strength of geopolymer concrete by using random forest algorithm. In: Shaw RN, Paprzycki M and Ghosh A (eds) *Advanced communication and intelligent systems. ICACIS 2022. Communications in computer and information science*. Vol. 1749. Cham: Springer, 2023, pp.170–179.
80. Verma M. Prediction of compressive strength of geopolymer concrete by using ANN and GPR. *Asian J Civ Eng*. Epub ahead of print 2 May 2023. DOI: 10.1007/s42107-023-00676-4.
81. Verma M, Upreti K, Dadhich P, et al. Prediction of compressive strength of green concrete by Artificial Neural Network. In: Shaw RN, Paprzycki M and Ghosh A (eds) *Advanced communication and intelligent systems. ICACIS 2022. Communications in computer and information science*. Vol. 1749. Cham: Springer, 2023, pp.622–632.
82. Verma M, Upreti K, Vats P, et al. Experimental analysis of geopolymer concrete: a sustainable and economic concrete using the cost estimation model. *Adv Mater Sci Eng* 2022; 2022: 1–16.
83. Gupta P, Gupta N, Saxena KK, et al. A novel hybrid soft computing model using stacking with ensemble method for estimation of compressive strength of geopolymer composite. *Adv Mater Process Technol* 2022; 8: 1494–1509.
84. Gupta A, Gupta N and Saxena KK. Experimental study of the mechanical and durability properties of slag and calcined clay based geopolymer composite. *Adv Mater Process Technol* 2022; 8: 655–669.
85. Shukla A, Gupta N, Dixit S, et al. Effects of various *Pseudomonas* bacteria concentrations on the strength and durability characteristics of concrete. *Buildings* 2022; 12: 993.
86. Gupta P, Gupta N, Saxena KK, et al. Multilayer perceptron modelling of geopolymer composite incorporating fly ash and GGBS for prediction of compressive strength. *Adv Mater Process Technol* 2022; 8: 1441–1455.
87. Kalyani G, Janakiramaiah B, Karuna A, et al. Diabetic retinopathy detection and classification using capsule networks. *Complex Intell Syst* 2023; 9: 2651–2664.
88. Godavarthi B, Nalajala P and Ganapuram V. Design and implementation of vehicle navigation system in urban environments using Internet of Things (Iot). *IOP Conf Ser Mater Sci Eng* 2017; 225: 012262.
89. Peddakrishna S and Khan T. Design of UWB monopole antenna with dual notched band characteristics by using  $\pi$ -shaped slot and EBG resonator. *AEU - Int J Electron Commun* 2018; 96: 107–112.
90. Atchudan R, Edison TNJI, Mani S, et al. Facile synthesis of a novel nitrogen-doped carbon dot adorned zinc oxide composite for photodegradation of methylene blue. *Dalton Trans* 2020; 49: 17725–17736.



91. Ray R, Choudhary SS, Roy LB, et al. Reliability analysis of reinforced soil slope stability using GA-ANFIS, RFC, and GMDH soft computing techniques. *Case Stud Constr Mater* 2023; 18: e01898.
92. Poul Raj IL, Valanarasu S, Hariprasad K, et al. Enhancement of optoelectronic parameters of Nd-doped ZnO nanowires for photodetector applications. *Opt Mater* 2020; 109: 110396.
93. Kalpana G, Kumar PV, Aljawarneh S, et al. Shifted adaption homomorphism encryption for mobile and cloud learning. *Comput Electr Eng* 2018; 65: 178–195.
94. Jayanthi N, Babu BV and Rao NS. Survey on clinical prediction models for diabetes prediction. *J Big Data* 2017; 4: 1.
95. Dhanalaxmi B, Naidu GA and Anuradha K. Adaptive PSO based association rule mining technique for software defect classification using ANN. *Procedia Comput Sci* 2015; 46: 432–442.
96. Sai Shravan Kumar P and Viswanath Allamraju K. A review of natural fiber composites [Jute, Sisal, Kenaf]. *Mater Today Proc* 2019; 18: 2556–2562.
97. Kota VR and Bhukya MN. A novel global MPP tracking scheme based on shading pattern identification using artificial neural networks for photovoltaic power generation during partial shaded condition. *IET Renew Power Gener* 2019; 13: 1647–1659.
98. Yeole K, Kadam P and Mhaske S. Synthesis and characterization of fly ash-zinc oxide nanocomposite. *J Mater Res Technol* 2014; 3: 186–190.
99. Kumar R, Dev N, Ram S, et al. Investigation of dry-wet cycles effect on the durability of modified rubberised concrete. *Forces Mech* 2023; 10: 100168.
100. Nigam M and Verma M. Effect of nano-silica on the fresh and mechanical properties of conventional concrete. *Forces Mech* 2023; 10: 100165.
101. Kumar N, Raut RD, Upreti K, et al. Environmental concern in TPB model for sustainable IT adoption. In: Al-Emran M, Al-Sharafi MA and Shaalan K (eds) *International conference on information systems and intelligent applications. Lecture notes in networks and systems*. Vol. 550. Cham: Springer, 2023, pp.59–70.
102. Selin C. Expectations and the emergence of nanotechnology. *Sci Technol Human Values* 2007; 32: 196–220.
103. Ul Haq E, Kunjalukkal Padmanabhan S and Licciulli A. Synthesis and characteristics of fly ash and bottom ash based geopolymers—a comparative study. *Ceram Int* 2014; 40: 2965–2971.
104. Lin C, Chen S and Hwang T. Residual strength of reinforced concrete columns exposed to fire. *J Chin Inst Eng* 1989; 12: 557–566.
105. EN 1992-2. Eurocode 2 - design of concrete structures - concrete bridges - design and detailing rules Eurocode. 2011.
106. Han L-H, Tan Q-H and Song T-Y. Fire performance of steel reinforced concrete columns. *J Struct Eng* 2015; 141: 1–10.
107. Zha X. Three-dimensional non-linear analysis of reinforced concrete members in fire. *Build and Environ* 2003; 38: 297–307.
108. Li LY and Purkiss J. Stress-strain constitutive equations of concrete material at elevated temperatures. *Fire Saf J* 2005; 40: 669–686.

# Kinetic treatment of lower hybrid waves excitation in a magnetized dusty plasma by electron beam

Anshu<sup>1</sup>, S C Sharma<sup>1\*</sup> and J Sharma<sup>2</sup>

<sup>1</sup>Department of Applied Physics, Delhi Technological University (DTU), Bawana Road, Delhi 110042, India

<sup>2</sup>Department of Physics, Amity School of Applied Sciences, Amity University Haryana, Manesar, Gurugram 122051, India

Received: 21 February 2023 / Accepted: 18 July 2023

**Abstract:** The theoretical modelling of electrostatic lower hybrid waves (LHWs) is investigated using a kinetic treatment involving an electron beam. This electron beam propagates through a magnetized dusty plasma cylinder that consists of dust grains, electrons, and positively charged potassium ions ( $K^+$ ). The excitation of LHWs via Cerenkov interaction using an analytical model is driven to instability. In order to explain how a population of charged dust particles affects the LHWs growth rate in a plasma that has been stimulated by an electron beam, a dispersion relation has been developed. The dust grain particles impact has been discussed on the growth rate of LHWs, and it was discovered that with the rise in relative density of dust grains, the growth rate of LHWs augments. Also, the growth rate of the unstable mode decreases with the frequency of the lower hybrid wave. Furthermore, the critical drift velocity for excitation of the mode is derived, and it was observed that it decreases as the relative density of negatively charged dust grains augments. The current study's findings align with the existing experimental observations.

**Keywords:** Dusty plasma; Lower hybrid wave; Electron beam; Kinetic treatment

## 1. Introduction

In the captivating realm of beam plasma systems [1, 2], a multitude of waves and instabilities exist such as upper hybrid waves [3], ion-cyclotron waves [4], and two-stream instabilities [5] but amongst them, the lower hybrid waves (LHWs) [6] are one of the most prevalent waves. For the last several years, there has been a significant surge of interest in studying quasi-static LHWs. The linear theory of excitation of electrostatic LHWs without charged dust grains by electron beams in plasma has been studied by Papadopoulos and Palmadesso [7]. Lower hybrid waves are electrostatic low-frequency plasma waves due to the longitudinal oscillation of ions in a magnetized plasma. LHWs have been observed in a few experimental and theoretical studies [8–15] by utilizing an ion and a beam of electrons. Thus, these waves are valuable in plasma warming devices to a high degree. Chang [9] had experimentally observed that the maximum growth rate of LHWs was achieved when the phase velocity, together with the magnetic field,

was practically identical to the thermal velocity of the electrons. Prakash et al. [10] studied LHWs driven by an electron beam in dusty plasma using the fluid treatment and examined the reliance of the growth rate on the beam velocity. Sharma et al. [11] demonstrated the LHWs excitation in a negative ion plasma via gyrating ion beam. They observed that the growth rate of negative and positive ion modes, i.e. unstable modes, augments with the relative density ratio of negative ions.

The dust grain charge fluctuations have a significant impact on the growth and damping of the wave [16–18]. Lately, extensive work has been done regarding waves and gained a lot of enthusiasm for different waves and instabilities in dusty plasma [19]. A few wave modes have been considered tentatively and analytically, starting with the Bliokh and Yarashenko [20] work by handling waves in Saturn's rings in dusty plasma. Seiler and Yamada [21] have studied the lower hybrid wave instability in a linear Princeton Q-1 device by a spiralling ion beam and reported the results. Barkan et al. [22] have proclaimed exploratory outcomes in a dusty plasma on the current-driven electrostatic ion cyclotron (EIC) instability. They reported that the growth rate of the instability was improved in the presence

\*Corresponding author, E-mail: suresh321sharma@gmail.com

of negatively charged dust grains. Song et al. [23] studied the current-driven EIC instability in negative ion plasmas and found that for the excitation of the wave in a Q-machine, with the increase in the ratio of negative-to-positive ion density, the critical drift velocity of the electron decreased. The instability growth rate augments with the beam density, as observed by Gupta et al. [24].

Dust grains can acquire charge through various methods, such as energetic ion sputtering, plasma currents, photoelectric effects, auxiliary electron outflow, and other techniques commonly employed in laboratory settings [25]. They are often negatively charged; however, the smaller grains might be positively charged. The presence of charged dust grains in close proximity significantly impacts the properties of the surrounding plasma. D' Angelo [26] investigated the dispersion relation for inhomogeneous dusty plasma immersed in a uniform and steady magnetic field for low-frequency electrostatic waves. Within sight of charged dust grains that are negative, D' Angelo found that the frequency modes increase in a definite proportion to the density ratio of negatively and positively charged ions. The principal distinction of multi-component plasmas from dusty plasmas is that the charges in the dust are not settled but rather preferably controlled by the parameters of plasma in their environment. Here, this impact is considered to define the general active approach [27] for all types of dusty plasma. Chow and Rosenberg [28] observed that the critical drift velocity is reduced for the wave excitation as the relative dust concentration increases, indicating that the wave mode is further destabilized in a negatively charged dusty plasma. Sharma and Sugawa [29] have examined that with the increase in the relative density of negatively charged dust grains, the growth rate and frequency of wave instability increases. Merlino et al. [30] have examined analytically and experimentally the negatively charged dust grains' effect on electrostatic waves of low frequency in a dusty plasma.

In recent years, a kinetic treatment [31] of dusty plasmas has emerged, encompassing the crucial aspect of dust particles absorbing plasma constituents, namely ions and electrons. More recently, Sharma et al. [32] developed a model based on the impact of dust charge fluctuations by relativistic runaway electrons in a tokamak on the parametric decay of lower hybrid wave instability using kinetic theory. They studied the impact of dust charge fluctuations on LHWs instabilities growth rate. Anshu et al. [33] studied the correlation between the growth rate and relative density ratio using the kinetic theory model by the relativistic electron beam on lower hybrid waves. They showed that when the relative density ratio increases, so does the rate of growth increase.

In this present work, the excitation of LHWs is studied via the kinetic theory model by an electron beam in a

magnetized dusty plasma. The distribution function of dust is taken to be Maxwellian, as it appears to be a steady arrangement of the present kinetic theory model. Using the Vlasov theory approach [34], the response of the beam electrons has been obtained. Electrostatic LHWs are driven to instability via Cerenkov interaction by a magnetized dusty plasma interacting with the electron beam. In Sect. 2, the instability of LHWs has been examined, and the outcome of the current work is explained in Sect. 3. Section 4 summarizes the conclusions.

## 2. Mathematical model

We consider a system filled with plasma that contains electrons and ions with initial densities  $n_e^0$  and  $n_i^0$ , respectively, in the existence of an externally applied static magnetic field  $B_0$  in the z-direction. Negatively charged dust grains of density  $n_d^0$  are introduced into the system. We examine a lower hybrid low-frequency wave travelling in the x-z plane, which lies firmly perpendicular to the applied magnetic field. An electron beam propagates in the z-direction with an initial velocity as  $v_b \hat{z}$  and an initial beam density as  $n_b^0$ . A small amplitude perturbed electrostatic potential  $\phi$  is given by

$$\phi = \phi_0 \exp[-i(\omega t - k_x x)], \quad (1)$$

where  $(\omega, k_x)$  are the frequency and wave vector, and  $\phi_0$  is the amplitude of the unperturbed electrostatic potential of the lower hybrid wave.

The perturbed density can be evaluated using the Vlasov equation as

$$\frac{\partial F}{\partial t} + \sum_j \left( r_j \nabla f + \frac{\partial}{\partial r_j} \cdot r_j F \right) = 0, \quad (2)$$

where  $F$  is the distribution function,  $f$  is the interacting force,  $r$  is phase space and  $\dot{r}_j$  is the time derivative of phase space of a single particle and  $f = ma$ , where  $m$  and  $a$  are the mass and acceleration of the particle. The expansion of the distribution function of the electron about the equilibrium is as follows  $F = f_0 + f_{md}$ , where  $f_0$  is the equilibrium distribution function and

$$f_{md} = \frac{\omega_c^e / m_e^2}{\pi^{3/2} v_t^e} \exp\left(\frac{-\mu_m \omega_c^e + p^2 / 2m_e}{T_e}\right) \quad (3)$$

is the perturbed Maxwellian distribution function,  $\omega_c^e \left( = \frac{eB_0}{m_e c} \right)$  denotes the cyclotron frequency of electron, the mass of the electron is  $m_e$ , the speed of light is  $c$ ,  $e$  is an electronic charge,  $p$  refers to the axial momentum,  $v_t^e \left( = \sqrt{\frac{2T_e}{m_e}} \right)$  denotes the thermal velocity of the electron,

$T_e$  defines electron temperature,  $\mu_m \left( = \frac{m_e v_e^2}{2\omega_e^2} \right)$  is the electron magnetic moment.

The density perturbation term is evaluated as.

$$n = \frac{m_e^2}{\omega_e^2} \iiint f_{md} d^3 \vec{v} \quad (4)$$

Equation (4) can be simplified by substituting the expression of  $f_{md}$  from Eq. (3), and the density perturbation term for electrons and ions can be evaluated as

$$n_{1e} = \frac{n_e^0 e \phi}{T_e} \left[ 1 + \frac{\omega_l}{k_l v_t^e} \sum_n Z \left( \frac{\omega_l + n \omega_e^e}{k_l v_t^e} \right) I_n(u_e) \exp(-u_e) \right] \quad (5)$$

$$n_{1i} = \frac{n_i^0 e \phi}{T_i} \left[ 1 + \frac{\omega_l}{k_l v_t^i} \sum_n Z \left( \frac{\omega_l - n \omega_i^i}{k_l v_t^i} \right) I_n(u_i) \exp(-u_i) \right] \quad (6)$$

Now, the density perturbation term for the beam is as follows

$$n_{1b} = \frac{-n_b^0 e \phi}{T_b} \left[ 1 + \frac{\omega_l - k_z v_b}{k_l v_t^b} \sum_n Z \left( \frac{\omega_l - k_z v_b}{k_l v_t^b} \right) I_n(u_b) \exp(-u_b) \right] \quad (7)$$

$n_e^0, n_i^0, n_b^0$ , is the initial electron, ion and beam density,  $\omega_c^i \left( = \frac{e B_0}{m_i c} \right)$  is the ion cyclotron frequency, ion mass is  $m_i$ ,  $I_n$  is the modified Bessel's function,  $Z$  denotes the dispersion function of plasma,  $\omega_c^b \left( = \frac{e B_0}{m_b c} \right)$  denotes the beam cyclotron frequency,  $m_b$  is the beam mass,  $(T_i, v_t^i)$  denotes ion temperature and thermal velocity,  $(T_b, v_t^b)$  indicates beam temperature and thermal velocity, respectively.

Now, the density perturbation term for dust is given as

$$n_{1d} = \frac{n_d^0 Q_d^0 \phi}{T_d} \left[ 1 + \frac{\omega_l}{k_l v_t^d} \sum_n Z \left( \frac{\omega_l - n \omega_c^d}{k_l v_t^d} \right) I_n(u_d) \exp(-u_d) \right]$$

Here, the dust is taken to be unmagnetized, i.e.  $\omega_c^d \left( = \frac{Q_d^0 B_0}{m_d c} \right) = 0$ , where  $\omega_c^d$  denotes the dust cyclotron frequency,  $Q_d^0 (= -Z_d e)$  refers to the charge of dust grains,  $Z_d$  is the dust charge state,  $m_d$  is dust grain mass, i.e.  $m_d = 10^{12} m_p$ , where  $m_p$  is the mass of a proton,  $T_d$  refers to the temperature of dust grain,  $v_t^d$  is the dust grain thermal velocity. For simplicity, we have taken  $n = 1$ . As we can see that the dust cyclotron frequency varies inversely with the dust grain mass, it becomes negligible and significantly small. Now, the dust density perturbation term becomes

$$n_{1d} = \frac{n_d^0 Q_d^0 \phi}{T_d} \left[ 1 + \frac{\omega_l}{k_l v_t^d} \sum_n Z \left( \frac{\omega_l}{k_l v_t^d} \right) \right]. \quad (8)$$

In Poisson's Equation, substituting Eqs. (5)–(8),  $\nabla^2 \phi = 4\pi(e n_{1e} - e n_{1i} + Q_d^0 n_{1d} + e n_{1b})$  the linear dispersion relation for the electrostatic mode with beam and dust grains is obtained as

$$\varepsilon = 1 + \chi_e + \chi_i + \chi_b + \chi_d, \quad (9)$$

where  $\chi$  refers to the susceptibilities. Further, Eq. (9) can be simplified as

$$\begin{aligned} \varepsilon = 1 + \frac{2\omega_p^e}{k_l^2 v_t^e} & \left[ 1 + \frac{\omega_l}{k_l v_t^e} \sum_n Z \left( \frac{\omega_l}{k_l v_t^e} \right) I_n(u_e) e^{-u_e} \right] \\ & + \frac{2\omega_p^i}{k_l^2 v_t^i} \left[ 1 + \frac{\omega_l}{k_l v_t^i} \sum_n Z \left( \frac{\omega_l - n \omega_c^i}{k_l v_t^i} \right) I_n(u_i) e^{-u_i} \right] \\ & + \frac{2\omega_p^d}{k_l^2 v_t^d} \left[ 1 + \frac{\omega_l}{k_l v_t^d} \sum_n Z \left( \frac{\omega_l}{k_l v_t^d} \right) \right] \\ & + \frac{2\omega_p^b}{k_l^2 v_t^b} \left[ 1 + \frac{\omega_l - k_z v_b}{k_l v_t^b} \sum_n Z \left( \frac{\omega_l - k_z v_b}{k_l v_t^b} \right) I_n(u_b) e^{-u_b} \right], \end{aligned} \quad (10)$$

where  $\omega_p^e \left( = \frac{4\pi n_e^0 e^2}{m_e} \right)$ ,  $\omega_p^i \left( = \frac{4\pi n_i^0 e^2}{m_i} \right)$ ,  $\omega_p^b \left( = \frac{4\pi n_b^0 e^2}{m_b} \right)$  denotes the plasma frequency of electron, ion, and beam responses, respectively, and  $\omega_p^d \left( = \frac{4\pi n_d^0 Q_d^2}{m_d} \right)$  denotes the dust plasma frequency. For finding the solution of Eq. (10), we assume  $\left( \frac{\omega_l - n \omega_c^{\beta}}{k_l v_t^{\beta}} = \xi \right)$ , where  $\beta = d, \alpha(e, i \text{ and } b)$ , for dust, electron, ion and beam responses, respectively. Using the expression of plasma dispersion function  $Z(\xi)$  from the kinetic theory [35], we obtain  $Z(\xi) = -\frac{1}{\xi} - \frac{1}{2\xi^3} + i\sqrt{\pi}e^{-\xi^2}$ , if  $\xi \gg 1$

or

$$Z(\xi) = -2\xi + \frac{2}{3}\xi^3 + \dots + i\sqrt{\pi}e^{-\xi^2}, \text{ if } \xi < 1,$$

where higher-order terms are neglected. And using conditions  $v_t^d < \frac{\omega_l}{k_l} < v_t^i$ , Eq. (10) can be solved to yield the growth rate and real frequency expression as follows

$$\begin{aligned} \Gamma = -\sqrt{2\pi}\omega_p^b & \left[ \left\{ I_n^2(u_b) e^{-2u_b} \left( X^2 + 16\pi e^{-2\xi^2} \right) + (1 - 2I_n(u_b) e^{-u_b} X) \right\}^{1/2} - \right. \\ & \left. (1 + I_n(u_b) e^{-u_b} X) \right], \end{aligned} \quad (11)$$

$$\text{where } X = \frac{\omega_l}{(\omega_p^b - k_z v_b) \times 0.8 \times \delta} + \frac{\omega_l}{\omega_l - n \omega_c^e} + \frac{\omega_l - k_z v_b}{\omega_l - k_l v_t^b}.$$

$$\omega_r = \omega_p^b \left[ \frac{\left\{ I_n^2(u_b) e^{-2u_b} (X^2 + 16\pi e^{-2\xi^2}) + (1 - 2I_n(u_b) e^{-u_b} X) \right\}^{1/2} + (1 + I_n(u_b) e^{-u_b} X)}{1} \right]^{1/2} \quad (12)$$

The phase velocity is as follows

$$v_{ph} = \frac{\omega_r}{k_l} = \frac{\omega_p^b}{k_l} \left[ \frac{\left\{ I_n^2(u_b) e^{-2u_b} (X^2 + 16\pi e^{-2\xi^2}) + (1 - 2I_n(u_b) e^{-u_b} X) \right\}^{1/2} + (1 + I_n(u_b) e^{-u_b} X)}{1} \right]^{1/2} \quad (13)$$

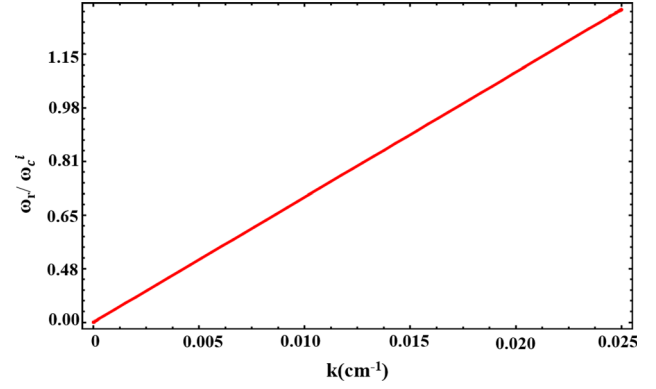
Equating the growth rate Eq. (11) equal to zero, i.e.  $\Gamma = 0$  equates to stability and describes the critical drift velocity  $u_{cd}$  as

$$u_{cd} = \log \left( \frac{16I_n \pi e^{-2\xi^2}}{4X} \right) \quad (14)$$

### 3. Results and discussion

In this paper, the analytical model has been developed to study the excitation of LHWs in a magnetized plasma cylinder with negatively charged dust grains by kinetic treatment with an electron beam. The developed model aims to analyse the relationship between various plasma parameters, i.e. real frequency, growth rate, etc. The effect of the dust grain size, critical drift velocity, and relative density ratio on the wave is studied. The equations that hold accountable for LHWs in Sec. II are used to calculate the real frequency and growth rate for the same. In the present calculations, typical dusty plasma parameters have been used. Using the following parameters, i.e. plasma density of ion  $n_i^0 = 1 \times 10^9 \text{ cm}^{-3}$ , plasma density of electron  $n_e^0 = (1 \times 10^9 - 0.2 \times 10^9) \text{ cm}^{-3}$ , guide magnetic field  $B_0 = 5 \times 10^4$  gauss, ion mass  $m_i^0 = 39 \times m_p$  (Potassium-plasma), where  $m_p = 1.67 \times 10^{-24}$  grams, the electron mass  $m_e = 9.1 \times 10^{-28}$  grams, ion temperature  $T_i = 0.2 \text{ eV}$ , the temperature of the electron  $T_e = 0.2 \text{ eV}$ , the density of the electron beam  $n_b^0 = 10^9 \text{ cm}^{-3}$ , density of the dust grains  $n_d^0 = 5 \times 10^4 \text{ cm}^{-3}$ , dust grain mass  $m_d = 10^{12} m_p$ , and size of the dust grains  $a = 5 \times 10^{-4} \text{ cm}$ , ion sound speed  $C_s = \sqrt{\frac{k_B T_e}{m_i}}$ ,  $k_B$  is the Boltzmann constant. We have plotted the various graphs.

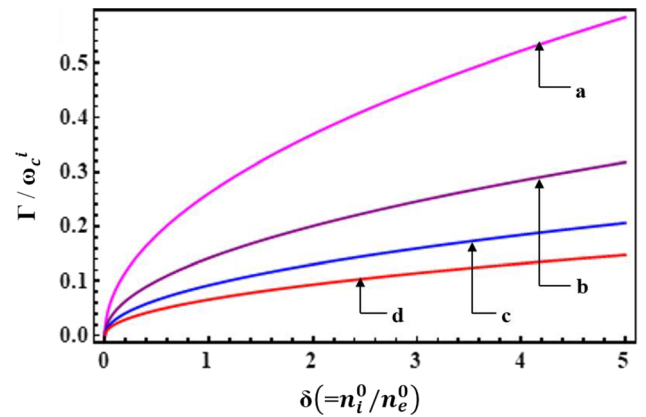
In Fig. 1, the dispersion curve for LHWs has been plotted using Eq. (10) for the parameters mentioned above of lower hybrid beam plasma instability. The graph



**Fig. 1** Dispersion curve in a magnetized dusty plasma of lower hybrid waves

portrays the linear relationship between wavenumber  $k \text{ (cm}^{-1}\text{)}$  and the normalized frequency of LHWs, signifying that they are dependent on each other.

Figure 2 depicts the normalized growth rate variation with a relative density ratio of negatively charged dust grains  $\delta (= n_i^0/n_e^0)$  for different values of dust grain size, (a)  $a = 2 \text{ }\mu\text{m}$ , (b)  $a = 4 \text{ }\mu\text{m}$ , (c)  $a = 6 \text{ }\mu\text{m}$ , (d)  $a = 8 \text{ }\mu\text{m}$ . It is observed from Fig. 2 that the value of growth rate is more significant at the lower value of dust grain size. Because of the small dust grain size, the electron density will increase, which further enhances the growth rate. As we augment the density ratio, the rate of growth also increases. Growth rate of LHWs instability in the existence of dust charge stabilizes more as the size of the dust grain increases with  $\delta (= n_i^0/n_e^0)$ . The rate of growth gradually decreases as the size of the dust grain augment and the curve flattens. Moreover, it can also be observed that the normalized rate of growth of LHWs upsurges by a factor of  $\sim 1.75$ , with the variation of  $\delta$  from 1 to 3 and by a factor of  $\sim 2.04$  with the variation of  $\delta$  from 1 to 4. Hence, the



**Fig. 2** Normalized growth rate variation with  $\delta (= n_i^0/n_e^0)$  for the varying value of dust grain size, i.e. (a)  $a = 2 \text{ }\mu\text{m}$ , (b)  $a = 4 \text{ }\mu\text{m}$ , (c)  $a = 6 \text{ }\mu\text{m}$ , (d)  $a = 8 \text{ }\mu\text{m}$

obtained result is in line with the theoretical observations of Sharma and Walia [36].

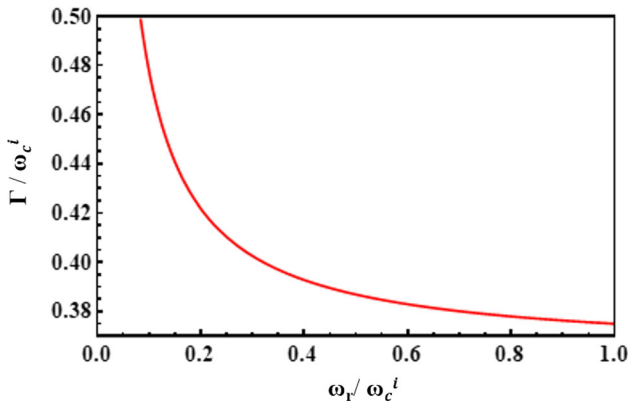
We have drawn the graph of normalized growth rate versus normalized frequency  $\omega_r/\omega_c^i$  of LHWs in Fig. 3 for the same parameters as plotted in Figs. 1 and 2. This graph shows that the normalized growth rate and frequency of LHWs are inversely proportional to each other. With an increase in frequency, the critical drift increases for the excitation of mode and hence the growth rate decreases. When the frequency of LHWs is more, the growth rate of the wave is more stable.

In Fig. 4, the normalized lower hybrid wave frequency variation with the dust grain size of negatively charged dust grains for different relative density ratio  $\delta (= n_i^0/n_e^0)$ , (a)  $\delta = 6$ , (b)  $\delta = 4$ , (c)  $\delta = 2$ , (d)  $\delta = 1$ , is shown. As we can see.

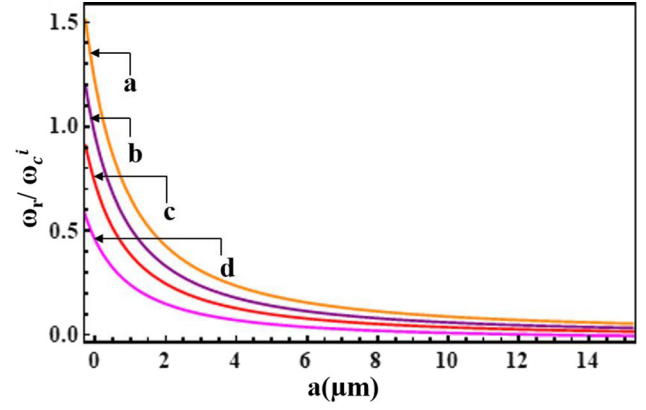
$$\omega_p^d \left( = \frac{4\pi n_d^0 Q_d^{02}}{m_d} \right) \quad (15)$$

where  $m_d$  = density  $\times$  volume and volume  $= \frac{4}{3}\pi a^3$ , where  $a$  is the radius of the dust grains. Here, we assume that the dust grains are spherical. By putting the expression of mass ( $m_d$ ) in Eq. (15), we can say that frequency scales to  $3/2$  power of dust grain size, i.e.  $\omega_p^d \left( = \left(\frac{1}{a}\right)^{3/2} \sqrt{\frac{3n_d^0 Q_d^{02}}{\text{density}}} \right)$ . From this, we can conclude that as the dust grain size increases, the frequency of the LHWs decreases. As the size of the dust grain increases, volume increases, and hence more electrons will stick to the dust grains. As a result, electron density decreases sharply, and consequently, wave frequency decreases. Further, with the increase in  $\delta (= n_i^0/n_e^0)$ , the frequency augments.

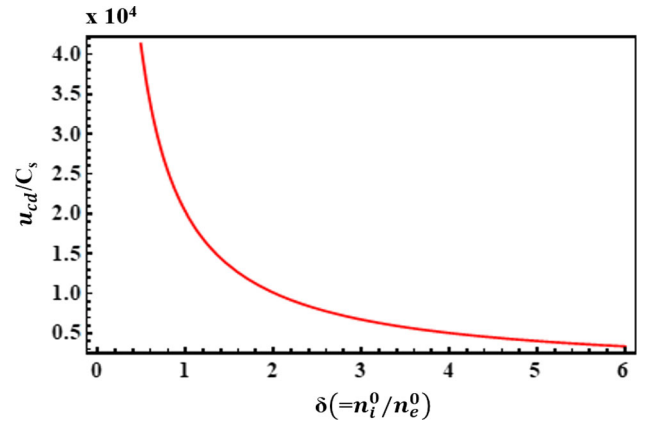
The normalized critical drift velocity versus relative density ratio  $\delta (= n_i^0/n_e^0)$  graph is portrayed in Fig. 5. From Fig. 5, we have inferred that with the increase in relative density ratio, the critical drift velocity decreases and can be



**Fig. 3** Normalized growth rate variation with the normalized frequency of the lower hybrid wave



**Fig. 4** The normalized frequency variation with the dust grain size for different relative density ratio  $\delta (= n_i^0/n_e^0)$ , i.e. (a)  $\delta = 6$ , (b)  $\delta = 4$ , (c)  $\delta = 2$ , (d)  $\delta = 1$



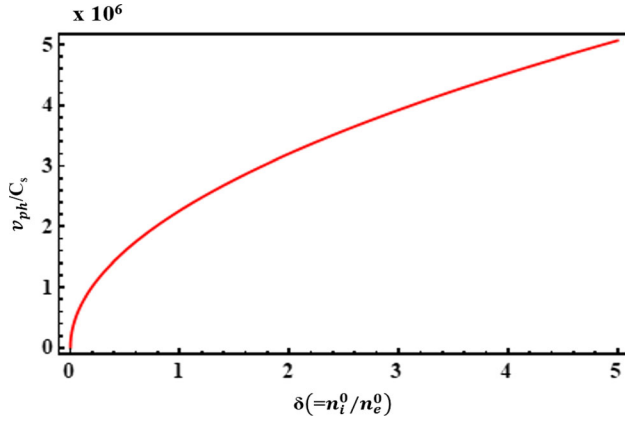
**Fig. 5** The normalized critical drift velocity variation with the relative density ratio  $\delta (= n_i^0/n_e^0)$

seen from Eq. (14) also. From Fig. 5, we can say that as  $\delta$  changes from 4 to 1, the critical drift velocity increases by about 2.97%. Hence, we can say that it is stabilizing the growth rate of the wave. For  $\delta \geq 2$ , the rate of critical drift decreases at a very slow rate.

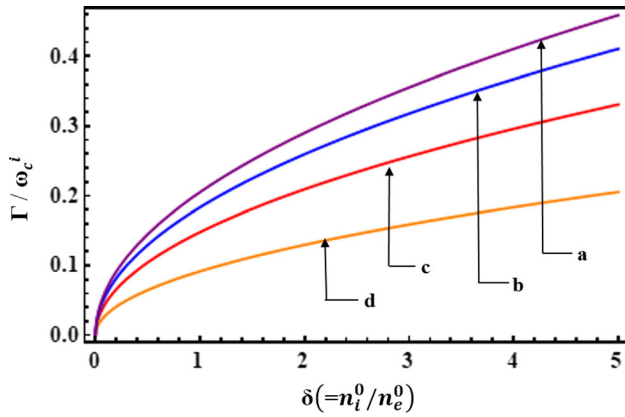
Figure 6 depicts the normalized phase velocity graph of negatively charged dust grains with respect to the relative density ratio  $\delta (= n_i^0/n_e^0)$ . The  $\delta (= n_i^0/n_e^0)$  variation from 1 to 5 is shown. In the presence of negatively charged dust grains of LHWs, the phase velocity augments with the relative density ratio. This is because, with the increase in electron density, the real frequency increases, which in turn, increases the phase velocity of the waves. From Fig. 6, it is clear that as the  $\delta$  value increases, the phase velocity of the wave also increases. This is because the space occupied by electrons on the dust grains becomes large.

Figure 7 depicts the normalized growth rate variation with the relative density ratio  $\delta (= n_i^0/n_e^0)$  for varying





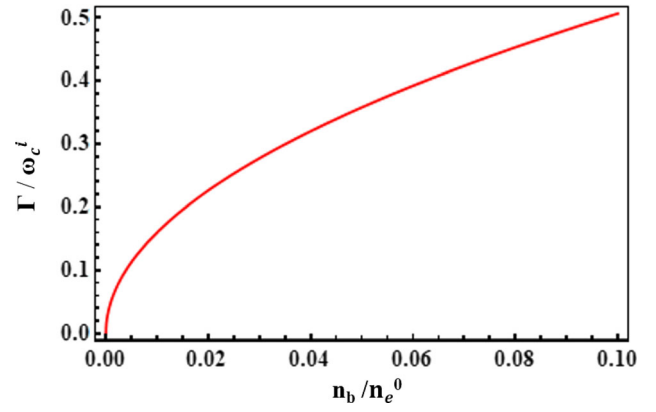
**Fig. 6** Normalized phase velocity  $v_{ph}$  as a function of relative density ratio  $\delta (= n_i^0/n_e^0)$ .



**Fig. 7** Normalized growth rate graph is plotted versus  $\delta (= n_i^0/n_e^0)$  for different beam velocities  $v_b^0$ , i.e. (a)  $v_b^0 = 10 \times 10^8$  cm/sec, (b)  $v_b^0 = 8 \times 10^8$  cm/sec, (c)  $v_b^0 = 5.2 \times 10^8$  cm/sec and (d)  $v_b^0 = 2 \times 10^8$  cm/sec

values of beam velocities  $v_b^0$ , i.e. (a)  $v_b^0 = 10 \times 10^8$  cm/sec, (b)  $v_b^0 = 8 \times 10^8$  cm/sec (c)  $v_b^0 = 5.2 \times 10^8$  cm/sec and (d)  $v_b^0 = 2 \times 10^8$  cm/sec. As we observed, the growth rate of LHWs augments with the velocity of the beam, i.e. beam electrons give energy to the waves, and the wave is growing. As the velocity of the beam augments, the maximum value of the growth rate increases. It is observed that the wave frequency of LHWs increases when the beam velocity and relative density ratio are increased. It can be observed that the normalized growth rate of unstable wave upsurges by a factor of  $\sim 1.68$ , with a variation of  $\delta$  from 1 to 3 and by a factor of  $\sim 1.93$ , with the variation of  $\delta$  from 1 to 4. Our results are in line with Prakash et al. [6].

Figure 8 depicts the growth rate of LHWs with respect to beam density. We can conclude that as beam density increases, so does the wave's growth rate. The rate of growth is determined by the square root of the beam's



**Fig. 8** The normalized growth rate variation with the beam density

density, as shown by Eq. (11). Our theoretical findings are consistent with R.P.H. Chang [9] of experimental observations.

#### 4. Conclusions

In magnetized dusty plasma, the LHWs excitation has been examined by electron beam using the kinetic treatment. The electrostatic LHWs via Cerenkov interaction are driven to instability. The outcomes of this work with different plasma parameters were compared and discussed. In magnetized dusty plasma, we investigated the impact of dust grain size on the growth rate and excitation of LHWs. The rate of growth of LHWs augments with the relative density ratio. The growth rate of LHWs increases significantly with a higher population of dust grains. Our growth rate findings are consistent with Sharma et al. [36] theoretical results and Chang [9] experimental observations. The critical drift velocity for mode excitation decreases as the relative density ratio of dust grains increases, resulting in an increase in the phase velocity. Furthermore, it was shown that the growth rate of the unstable mode is proportional to the one-half power of the density of the beam and increases with beam density. LHWs instability driven by an electron beam in plasma processing of material experiments [37] can alter the plasma-surface interactions. Furthermore, in the plasma processing of materials [38–40], nano-sized dust particles can be incorporated.

Very few studies have been made so far to derive the relationship among various parameters such as growth rate, frequency and size of dust grains particles, etc. using kinetic treatment. This work may serve as a foundation for future studies, where experimentalists and theorists are likely to draw inspiration from its findings. LHWs can be studied in tokamak plasmas [41], lunar dusty plasma [42] and has recently received significant attention for affecting the current drive at a very high density [43].

The fundamental mechanics and characteristics of lower hybrid waves are often best understood using linear theories which we have shown in our present work. In the given plasma model, nonlinear effects are not taken into account which can play an important role in changing the plasma behaviour. For example, the dispersion relation of LHWs. Nonlinear effects can indeed play a crucial role in the generation and propagation of waves, energy transfer, and plasma current-driving abilities of lower hybrid waves [44]. Therefore, it is crucial to investigate and characterize these nonlinear effects in theoretical and experimental studies. In the near future work, the plasma model can be studied after incorporating these effects which will surely help our research move forward.

**Acknowledgements** The SERB-DST grant (grant no. EMR/2016/002699), which provided the necessary funding, is acknowledged by one of the authors.

**Data availability** The findings of this research are supported by the data that are presented in the manuscript. Additionally, the corresponding author will disclose the data supporting the manuscript's conclusions upon reasonable request.

## References

- [1] M M Shoucri and R R J Gagne *J. Plasma Phys.* **19** 281 (1978)
- [2] D N Gupta and A K Sharma *Laser & Particle Beams* **22** 89 (2004)
- [3] S C Sharma and A Gahlot *Phys. Plasmas* **16** 123708 (2009)
- [4] M Yadav and J Sharma *Indian J. Pure Appl. Phys.* **59** 671 (2021)
- [5] Y Pinki, D N Gupta and K Avinash *Phys. Plasmas* **24** 062107 (2017)
- [6] V Prakash, Vijayshri, S C Sharma and R Gupta *Phys. Plasmas* **20** 053701 (2013)
- [7] K Papadopoulos and P Palmadesso *Phys. Fluids* **19** 605 (1976)
- [8] O Koshkarov, A I Smolyakov, A Kapulkin, Y Raites and I Kaganovich *Phys. Plasmas* **25** 061209 (2018)
- [9] R P H Chang *Phys. Rev. Lett.* **35** 285 (1975)
- [10] V Prakash, V Vijayshri, S C Sharma and R Gupta *Phys. Plasmas* **20** 063701 (2013)
- [11] J Sharma, S C Sharma, V K Jain and A Gahlot *Phys. Plasmas* **20** 033706 (2013)
- [12] S C Sharma, M P Srivastava, M Sugawa and V K Tripathi *Phys. Plasmas* **5** 3161 (1998)
- [13] S C Sharma and Ritu Walia *Phys. Plasmas* **15** 093703 (2008)
- [14] Magdi M. Shoucri and Réal R. J. Gagné, *J. Plasma Phys.* **19** 28 (1978)
- [15] R L Stenzel and W Gekelman *Phys. Rev. A* **11** 6 (1975)
- [16] S V Vladimirov, K N Ostrikov and M Y Yu *Phys. Rev. E* **60** 3257 (1999)
- [17] M R Jana, A Sen and P K Kaw *Phys. Rev. E* **48** 3930 (1993)
- [18] S V Vladimirov, K N Ostrikov, M Y Yu and L Stenflo *Phys. Rev. E* **58** 8046 (1998)
- [19] S V Vladimirov, K Ostrikov, M Y Yu and G E Morfill *Phys. Rev. E* **67** 036406 (2003)
- [20] P. V., Bliokh and Yarashenko, V. V. *Soil. Astron., Engl. Transl.* **29** 330 (1985)
- [21] S Seiler and M Yamada *Phys. Rev. Lett.* **37** 700 (1976)
- [22] A Barkan, N D'Angelo and R L Merlino *Planet. Space Sci.* **43** 905 (1995)
- [23] B Song, D Suszcynsky, N D'Angelo and R L Merlino *Phys. Fluids B* **1** 2316 (1989)
- [24] R Gupta, S C Sharma and V Prakash *Phys. Plasmas* **17** 122105 (2010)
- [25] T K Baluku and M A Hellberg *Phys. Plasmas* **22** 083701 (2015)
- [26] N D'Angelo *Planet. Space Sci.* **38** 1143 (1990)
- [27] V N Tsytovich *Phys. Plasmas* **7** 554 (2000)
- [28] V W Chow and M Rosenberg *Planet. Space Sci.* **44** 465 (1996)
- [29] S C Sharma and M Sugawa *Phys. Plasmas* **6** 444 (1999)
- [30] R L Merlino, A Barkan, C Thompson and N D'Angelo *Phys. Plasmas* **5** 1607 (1998)
- [31] P Ricci, G Lapenta, U de Angelis and V N Tsytovich *Phys. Plasmas* **8** 769 (2001)
- [32] J Sharma, S C Sharma and A Gahlot *Phys. Plasmas* **28** 043701 (2021)
- [33] Anshu, S C Sharma and J Sharma *46<sup>th</sup> EPS Conference on Plasma Physics*, EPS (2019)
- [34] C S Liu and V K Tripathi *Physics Reports* **130** 143 (1986)
- [35] M N Rosenbluth and R Z Sagdeev *Handbook of plasma physics*, (Elsevier Science Pub.) (1983)
- [36] C Suresh and R Sharma *Phys. Plasmas* **15** 093703 (2008)
- [37] K Ostrikov, U Cvelbar and A B Murphy *J. Phys. D: Appl. Phys.* **44** 174001 (2011)
- [38] K. Ostrikov and S. Xu, Plasma-Aided Nanofabrication from Plasma Sources to Nanoassembly (2007)
- [39] B Denysenko, S Xu, J D Long, P P Rutkevych, N A Azarenkov and K Ostrikov *J. Appl. Phys.* **95** 2713 (2004)
- [40] A Okita, Y Suda, A Ozeki, H Sugawara, Y Sakai, A Oda and J Nakamura *J. Appl. Phys.* **99** 014302 (2006)
- [41] Z Liu, Z Gao and A Zhao *Phys. Plasmas* **27** 042503 (2020)
- [42] S I Popel, A I Kassem, Y N Izvekova and Lev M Zelenyi *Phys. Lett. A* **384** 126627 (2020)
- [43] Jong Gab Jo, S H Kim, *Fusion Eng. Des.* **190** 113531 (2023)
- [44] S I Popel and V N Tsytovich *Contrib. Plasma Phys.* **32** 77 (1992)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

# Liver-Type Fatty Acid Binding Protein (FABP1) Has Exceptional Affinity for Minor Cannabinoids

Dr. Fred Shahbazi<sup>a,b,\*†</sup>, Sanam Mohammadzadeh<sup>a†</sup>, Dr. Daniel Meister<sup>a,b</sup>, Valentyna Tararina<sup>a,c,d‡</sup>, Vagisha Aggarwal,<sup>a,e,f‡</sup> Dr. John F. Trant<sup>a,b,g\*</sup>

†F.S. and S.M. contributed equally to this work and can order their name in any order for any professional purpose.

‡ V.T. and V.A contributed equally to this work and can order their name in any order for any professional purpose.

<sup>a</sup> Department of Chemistry and Biochemistry, University of Windsor, 401 Sunset Avenue, Windsor ON, N9B 3P4, Canada

<sup>b</sup> Binary Star Research Services, LaSalle ON, N9J 3X8, Canada

<sup>c</sup> Department of Chemistry, Taras Shevchenko National University of Kyiv, 60 Volodymyrska St., Kyiv, Ukraine

<sup>d</sup> Current Affiliation: Chemspace, 85 Winston Churchill St., Kyiv, 02094, Ukraine

<sup>e</sup> Department of Applied Chemistry, Delhi Technological University, Rohini, New Delhi, Delhi, 110042, India

<sup>f</sup> Current Affiliation: Department of Chemistry, Paris Sciences & Lettres, 60 Rue Mazarine, 75006 Paris, France

<sup>g</sup> We-Spark Health Institute, Windsor, ON, N9B 3P4

\* Corresponding authors' email: [farsheed@uwindsor.ca](mailto:farsheed@uwindsor.ca), [j.trant@uwindsor.ca](mailto:j.trant@uwindsor.ca)

## Abstract:

Fatty acid binding protein 1 (FABP1) is a lipid transporter primarily expressed in the liver where it helps move fatty acids between lipid membranes. Inhibition of FABP1 has potential therapeutic implications for nonalcoholic fatty liver disease, metabolic syndrome & obesity, diabetes, and inflammatory & cardiovascular diseases. Curiously, FABP1 is known to bind to both endocannabinoids (ECs) and the major phytocannabinoids (PCs) with moderately high affinities. We have developed an *in-silico* model of the protein and validated it against experimental data. We then employed the model to predict the binding mode and affinities of minor cannabinoids (MCs) to FABP1. Our study predicts that the top ranked MCs **5-acetyl-4-hydroxy-CBG** and **CBGA** bind stronger than fatty acids (FAs), ECs or PCs, and participate in the key interactions used to stabilize FABP1-FA complexes. This makes them promising starting points for the

development of new therapeutics. The implications this has on considering the minor cannabinoids as low entropy isosteres of the fatty acids is also discussed.

**Keywords:** minor cannabinoids, THC, molecular modeling, ligand profiling

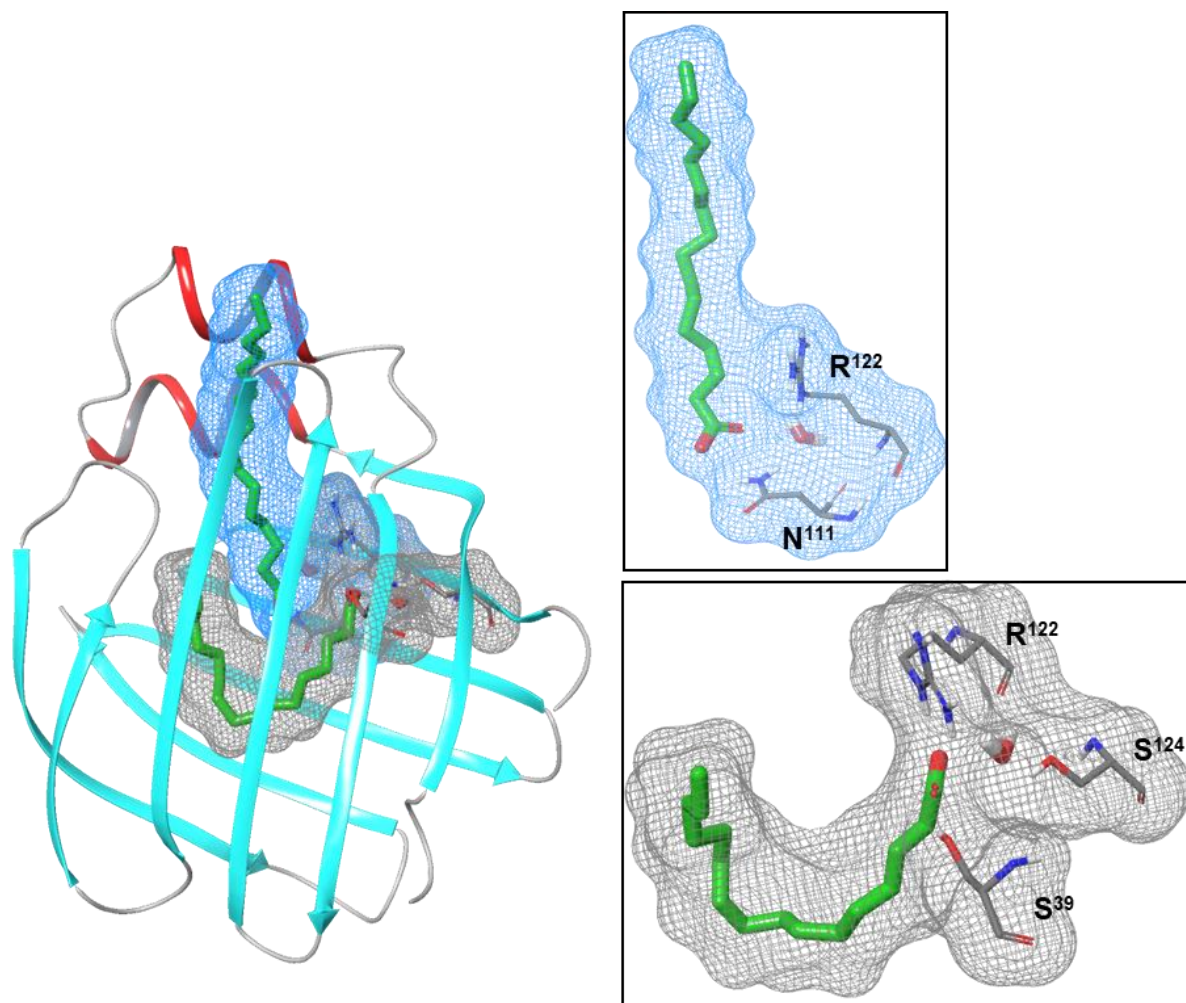
## Introduction

Fatty acid binding proteins transport water-insoluble hydrophobic molecules from membrane to membrane within a cell.<sup>1</sup> Of the ten identified members, FABP1 is the most highly expressed in humans overall, and is found primarily in tissues involved in fatty acid (FA) metabolism where it is implicated in their transport, especially to the mitochondria for oxidation.<sup>2</sup> It consequently regulates lipid metabolism and downstream cellular signaling pathways.<sup>3</sup> The mechanism of how FABP1 affects these signaling pathways is not fully understood, but might be through the PPAR pathway as the two proteins physically interact; regardless, abnormal (either elevated or decreased) levels of FABP1 are indicative of diabetic nephropathy and are predictive clinical markers of renal illnesses.<sup>4</sup> Knockout of the gene leads to mice with significant weight gain,<sup>2, 5</sup> and FABP1 overexpression enhances hepatocyte fatty acid uptake,<sup>6</sup> and is associated with a variety of cancers,<sup>7</sup> including liver, lung, stomach, and colon.

The chemical inhibition of FABP1 phenotypically modifies FA storage in adipose; this changes FA uptake, esterification, oxidation, nuclear targeting, and intracellular transport.<sup>2</sup> Inhibition of FABP1 fatty acid transport has been proposed as a target for preventing or reversing diet-induced obesity and diabetes.<sup>8</sup>

Although FABP1 is structurally related to the other FABPs, it is distinguished by its far larger binding cavity; this means that FABP1 alone can bind two FAs concurrently, and can also interact with other large hydrophobic species such as bilirubin and lysophospholipids.<sup>9</sup> When transporting traditional FAs, the first molecule, with affinity in the nM range, is totally encapsulated within the  $\beta$ -barrel structure of the protein, generally in a U-shaped conformation, locked in both by hydrophobic collapse and the carboxylate forming a series of H-bonds with **R**<sup>122</sup> and **S**<sup>39</sup>, and a water-bridged interaction with **S**<sup>124</sup> (Fig. 1, black mesh surface).<sup>8-9</sup> The second FA binds mostly by hydrophobic forces to the “lipophilic” binding pocket in the portal region with its carboxylic group buried in the protein cavity by polar contacts with **N**<sup>111</sup> and **R**<sup>122</sup> (Fig. 1, blue mesh surface, PDB: 3STK).<sup>8, 10</sup> This portal region can be thought of as the entry gate to the deeper pocket, but is a large enough opening that it can accommodate the second molecule. Of course, dissociation must occur first through the portal-bound molecule, followed by the high affinity bound molecule deep in the pocket. This is only really feasible when FABP1 is associated with a lipid membrane as otherwise the solvation energy is too much to overcome. Each of these two sites demonstrates similar high affinities for saturated FAs, while their affinities for polyunsaturated

fatty acids (PUFA) differs by more than 7-fold.<sup>11</sup> FABP1 has high affinity for *n*-6 polyunsaturated fatty acids (PUFA) such as arachidonic acid (ARA), or its endocannabinoid derivatives (ECs), anandamide (AEA), and 2-Arachidonoylglycerol (2-AG). This is because the internal site does not require a carboxylate group to attain high affinity.<sup>10</sup>



**Figure 1.** Structure of FABP1 bound to two palmitic acid residues (molecules in green, surface in blue and black mesh). The  $\beta$ -sheets of the protein are in cyan, the alpha helices in red. The structure is obtained following a relaxation of the crystal structure, *3STK* in an implicit water model.

In late 2019, Elmes and colleagues released the crystal structure of human FABP1 in complex with THC at 2.5Å resolution; the first crystal structure of a FABP bound to a cannabinoid (PDB: *6MP4*).<sup>12</sup> By promoting FABP1's cytoplasmic trafficking to hepatic CYP450 enzymes, they



showed that FABP1 plays a significant role in regulating THC biotransformation and metabolism. The large THC molecule fills the majority of the FABP1 binding cavity. Its conjugated rings occupy a hydrophobic pocket, its pyran ring's O1 atom forms a hydrogen bond with N<sup>111</sup>, and its carbon chain stretches back towards the pocket entrance. The bulky THC molecule occupies the majority of the FABP1 binding cavity. To support this effort or for experimental reasons, they also acquired a comparison of THC (PDB: 6MP4) and palmitic acid (PDB: 3STM) both bound to FABP1; THC is located away from the fatty acid's more polar environment. THC-FABP1 lacks the ion pair interaction between the carboxylate of the FAs and R<sup>122</sup>; the M<sup>74</sup> sidechain moves substantially to accommodate the large THC rings in the hydrophobic pocket, while the F<sup>50</sup> sidechain rotates to pack against THC's short hydrocarbon tail. However, there are significant steric clashes that are induced by the presence of THC in the binding pocket; indicating that the simultaneous binding of THC with endogenous lipids is unlikely.. Likely only one or the other can bind at any one time.

Stepping back for a moment, let us consider the structures of the two classes of cannabinoids: the endocannabinoids (EC) and the phytocannabinoids (PC, Figure 2). The ECs are structurally unrelated to the PCs: the former are long *cis*-polyunsaturated alkyl chains with a polar head group, the latter a polycyclic monoterpene-resorcinol conjugate, but both act on the cannabinoid CB1 and CB2 receptors in the brain.<sup>13</sup> Clearly, both also interact with FABP1. Although FABP1 is not found in the brain, it still regulates the levels of the ECs by facilitating their metabolism in the liver: FABP1 ablation preferentially elevated brain ECs system AEA and 2-AG outside of the central nervous system.<sup>14</sup> PCs may enhance ECs signaling by competing with FABP1 for absorption and subsequent metabolism.<sup>12</sup> This suggests that the endocannabinoids tend to fold, when in contact with proteins, into a compact surface reminiscent of the phytocannabinoids. As this requires a series of single bonds to all specifically rotate into a non-extended, low-dynamic conformation, there is a significant entropic cost.<sup>15</sup> The PCs do not have to pay this cost as the cyclic structure locks in the positioning. The fact that the endocannabinoids and the phytocannabinoids, despite their radically different structures both adopt binding conformations with the cannabinoid receptors and the unrelated FABPs points to their close conformational arrangement.

Despite the research focus, the diverse attributed bioactivity of the cannabinoids cannot be explained simply from interaction at the CB1 and CB2 receptors.<sup>16</sup> Some of this non-CB activity

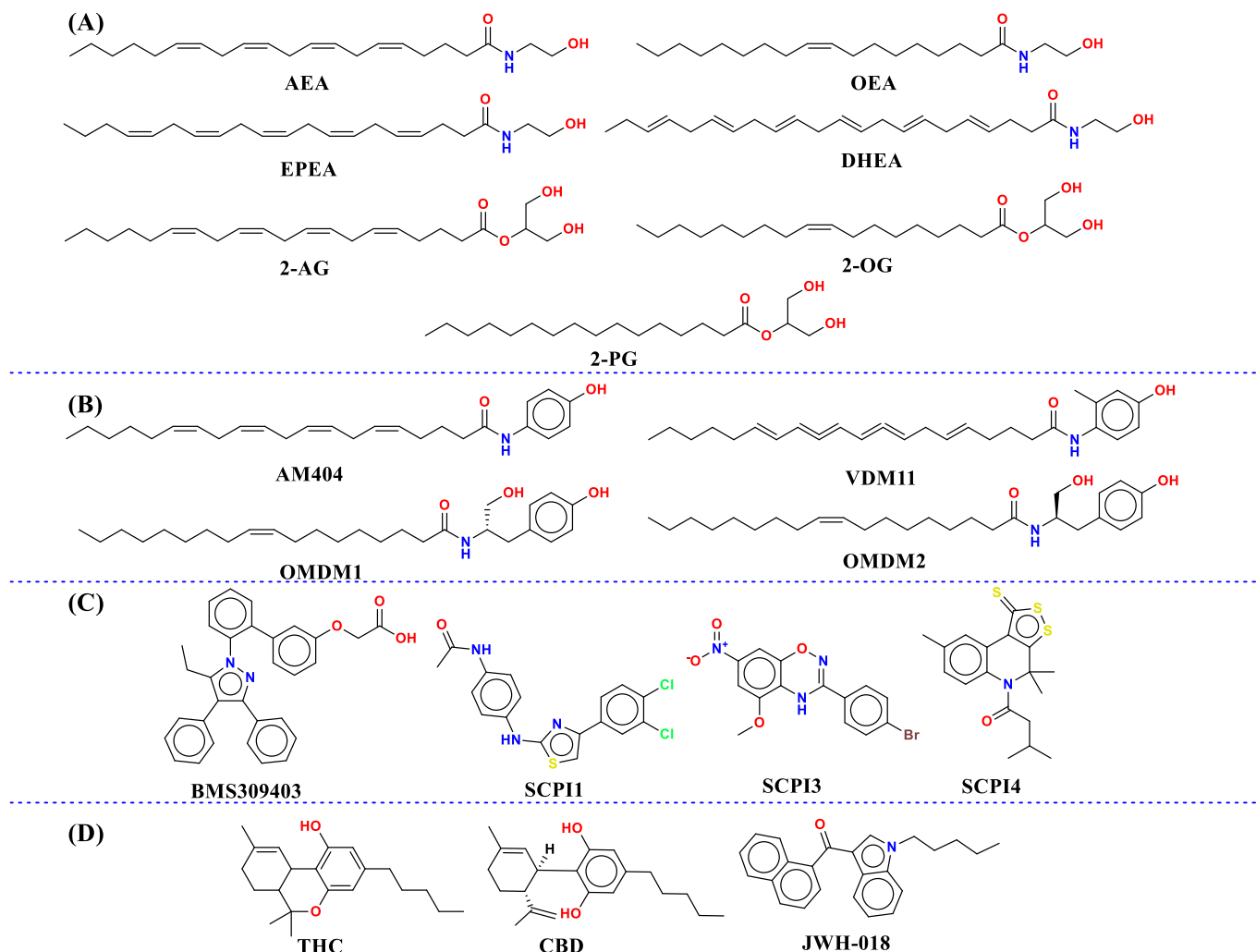
can however, be ascribed to the FABPs. Binding to intracellular FABPs has an effect on ligand-dependent transactivation of PPARs, a receptor known to be targeted by various cannabinoids through direct activation of its orthosteric site.<sup>17</sup> Furthermore, occupation of FABP1 increases AEA and 2-AG levels in the liver and brain (by decreasing the rate of their metabolism). The interactions may have an impact on inducing weight gain (as seen in the mouse models), but may also be important for controlling overexpression diseases.<sup>14</sup> Either way, cannabinoids might prove to be useful tools for probing the biology around FABP1, but, like all cases, the major cannabinoids are likely not the most active. THC and CBD are only the most common of over 200 phytocannabinoids that have been identified in *C. sativa* extracts.<sup>18</sup> The minor cannabinoids are only minor relative to the major ones, as consumers of cannabis may take in gram quantity of material at a time, the minor cannabinoids can be present in sufficient enough quantity to have meaningful biological activity.<sup>19</sup> These minor compounds could be far more potent than the major components. However, many of these minor cannabinoids have only been isolated or identified a few times, and besides a few “trendy” compounds are not readily available. This situation implies that an *in silico* screen would potentially prove valuable as a first step.

In order to develop a useful predictive *in silico* screen, one must build a good model of the protein and derive equations that can relate quantitative computationally-derived molecular interaction strengths with an experimental readout. This requires both a well-defined structure of the target protein, and a self-consistent and broad dataset. No such complete model exists for FABP1, but the required inputs are available. We consequently began our screen of FABP1 and minor cannabinoids with the generation of a corrected 3-D structure of the protein and a mathematical model that can recapitulate known interactions with known binders. If found useful, this can then be extended to screen new molecules to prophesize affinity. Our generation and application of this model is the focus of this current report.

## Results and Discussion

A total of 18 ligands with empirically-measured binding affinity,  $K_i$ , for FABP1, determined using the same fluorescence displacement assay by Huang and co-workers,<sup>14, 20</sup> were used to inform the model. The list includes both N-Acylethanolamides (NAEs) and 2-Monoacylglycerides (2-MGs) **AEA**, **OEA**, **EPEA**, **DHEA**, **2-AG**, **2-OG** & **2-PG**, AEA uptake

Inhibitors **AM404**, **VDM11**, **OMDM1** & **OMDM2**, general preclinical FABP inhibitors **BMS309403**, **SCPI1**, **SCPI3** & **SCPI4**, and **THC**, **CBD** & synthetic cannabinoid **JWH-018** (Figure 2). We used the Ligprep feature in Maestro to prepare the structures of each ligand. We obtained the SMILES notation for each compound from PubChem and entered it into Ligprep's SMILES field for structure preparation. Based on the binding configuration observed for THC in the X-ray crystallography structure (PDB: 6MP4), we employed Schrödinger's Maestro suite of computational tools, including for RRD, IFD, MD simulations, MMGBSA calculations, root-RMSD, and H-bond analysis for computational predictions of how experimental ligands would bind to the binding site. Our goal was to determine whether the computed interaction energies would align with the experimental trends. Details on how the FABP1 structures were prepared for simulations are available in the supplementary information.



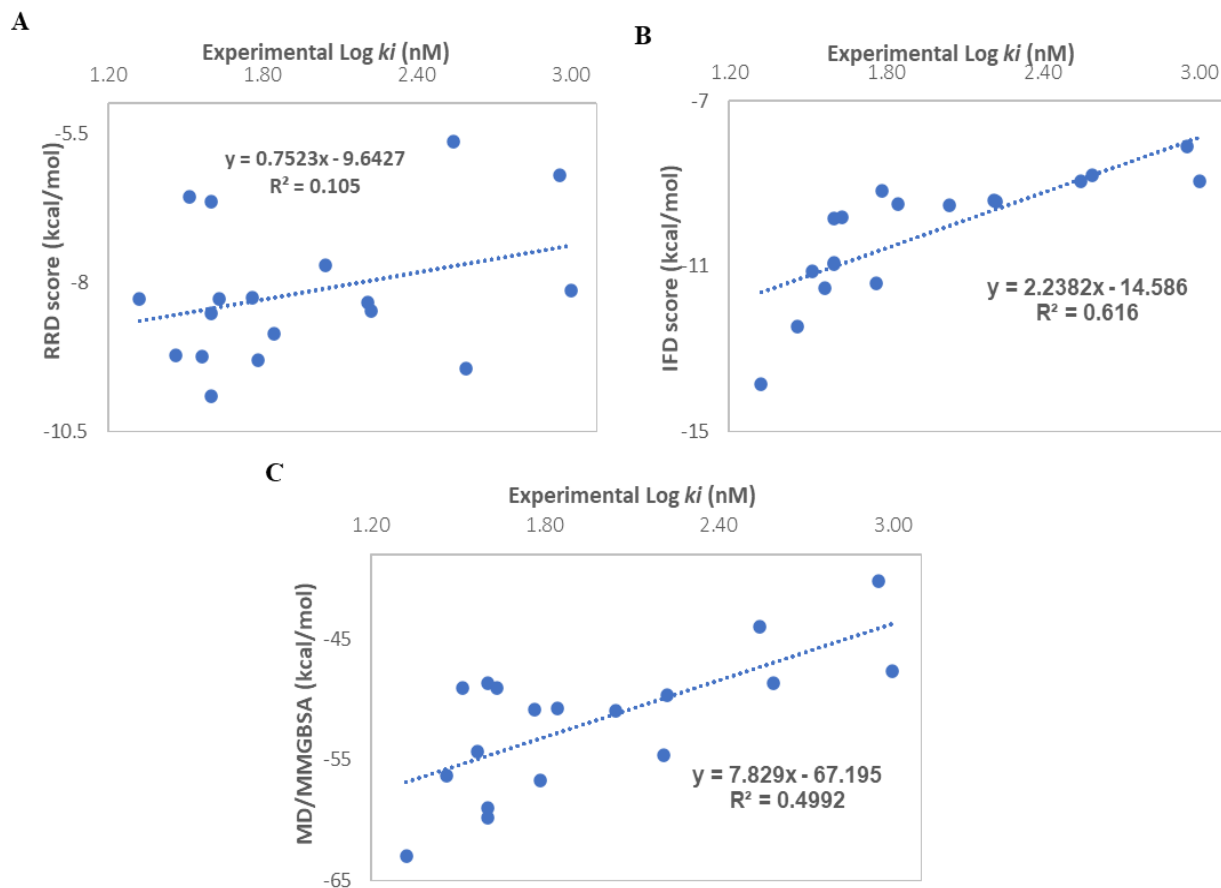
**Figure 2.** Structures of the experimental ligands with known experimental binding affinities used in this study to validate the molecular model of FABP1: A) NAEs and 2-MGs, B) AEA uptake Inhibitors, C) general FABPs inhibitor and D) phyto- and synthetic cannabinoids.

Several in-silico techniques were employed in scoping studies to create a model that gives a reliable correlation between experimental and computational data. First, rigid receptor docking (RRD) model was employed and the van der Waals radii of non-polar atoms was lowered to 0.8 to indicate some of the residue's flexibility.<sup>21</sup> Prime/MM-GBSA calculations were then done on this docked structure and the distance between flexible residues and ligands adjusted to 5.0 Å to better estimate the binding free energy of the most stable docked structure for each ligand.<sup>22</sup> Computationally more intensive induced fit docking (IFD) calculations were then performed on the top hits; IFD enables flexibility in the binding site residues and generally generates a more

reliable complex. The experimental  $K_i$  values, calculated physiochemical properties, rigid docking scores (RRD 0.8), IFD and MD/MM-GBSA were collected and analyzed (Table 1).

**Table 1.**  $K_i$  value, calculated physiochemical properties, RRD scores, IFD and MD/MM-GBSA of the FABP1 ligands. Experimental values were extracted from Huang and coworkers.<sup>14, 20</sup>

Experimental			Physiochemical Properties			Computational (kcal/mol)		
Ligand	$K_i$ (nM)	Log ( $K_i$ )	LogP (O/W)	MW (g/mol)	PSA	RRD 0.8	IFD	MD/MM-GBSA
AEA	111	2.05	4.94	347.54	54.99	-7.71	-9.53	-51.01
OEA	43	1.63	4.46	325.53	54.99	-8.27	-9.81	-49.06
EPEA	390	2.59	4.62	345.52	59.15	-9.46	-8.78	-48.66
DHEA	163	2.21	5.17	371.56	58.76	-8.34	-9.42	-54.64
2-AG	61	1.79	5.34	378.55	79.99	-9.30	-9.16	-56.76
2-OG	40	1.60	4.86	356.55	77.89	-9.92	-9.83	-48.64
2-PG	70	1.85	4.14	330.51	70.56	-8.85	-9.50	-50.73
AM404	29	1.46	6.87	395.58	57.80	-9.23	-12.45	-56.32
VDM11	37	1.57	7.14	409.61	55.70	-9.25	-11.53	-54.37
OMDM1	40	1.60	5.51	431.66	70.41	-8.52	-10.95	-59.03
OMDM2	40	1.60	5.51	431.66	77.06	-6.63	-10.91	-59.79
BMS309403	21	1.32	5.64	474.56	74.59	-8.28	-13.87	-62.96
SCPI1	350	2.54	4.53	378.28	58.05	-5.64	-8.93	-43.99
SCPI3	900	2.95	2.77	364.150	86.775	-6.21	-8.10	-40.17
SCPI4	33	1.52	4.47	363.55	28.94	-6.57	-11.11	-49.04
THC	1000	3.00	5.64	314.47	25.95	-8.14	-8.93	-47.70
CBD	167	2.22	5.38	314.47	38.18	-8.47	-9.42	-49.63
JWH-018	58	1.76	6.17	341.45	24.03	-8.25	-11.40	-50.87



**Figure 3.** Correlation analysis between experimental values, (A) RRD, (B) IFD scores and (C) MD/MM-GBSA in FABP1.

The first effort was simply to plot the correlation between the three methods for predicting binding affinities (rigid docking score, RRD, induced docking score; IFD; and post-MD corrected binding energy, MMGBSA) and the experimental  $K_i$ . There is only a weak correlation between the experimental log  $K_i$  and the RRD score ( $R^2 = 0.11$ , Fig 2A). This however improves markedly when induced-fit docking is used ( $R^2 = 0.62$ , Fig 2B).<sup>23</sup> This makes sense considering that we expect the binding site to undergo significant rearrangement upon the binding of a ligand. This is true for all proteins of course, but especially true for the FABPs where the “roof” of the binding site closes down over the ligand when it enters the protein. The *apo*-form is interesting as a starting point, but a rigid *apo*-form is unsurprisingly a poor mimic for an occupied FABP.

Rigid docking was never likely to return a useful result. Using MM-GBSA analysis of the MD simulations does result in a decrease in the Pearson coefficient to 0.5. Ideally, we would see a



perfect correlation between experimental and theoretical data, but this is never the case. Experimental data can have significant inherent error, and this can exacerbate differences in the correlation. Another complication is that these are all strong binders. We have nothing above 1  $\mu\text{M}$  in the data set, and this means we cover under two orders of magnitude. This essentially makes them similar binders and it can be difficult to differentiate small changes in affinity using medium-throughput computational methods or experimental methods. A Pearson of 0.6 is perfectly acceptable as a starting point. The quality of the analysis is also hindered by the similarity within classes of these ligands, there is not a large conformational, and hence interaction, variety.

For all the NAE and 2-MG ligands, the oxygen of the amide/arachidonic acid group is predicted to directly interact with **M<sup>74</sup>**. All the phenolic AEA uptake inhibitors are predicted to orient towards the inside of the pocket, where the phenolic oxygen would interact with **R<sup>122</sup>** and **S<sup>124</sup>**. **SCPI1** & **SCPI3** are predicted to have an interaction with **M<sup>74</sup>**, **SCPI4** & **CBD** with **M<sup>74</sup>**, and **JWH-013** with **R<sup>122</sup>** (Fig. S1A). These are likely reasonable poses, both because of the match to experimental data and because the IFD docking pose for THC closely reflects the crystallographic result (PDB: *6MP4*, Fig. S1B). This was promising from the simple reorganization of the protein; for a hopefully more accurate reflection of the energies and binding poses, we conducted an MD simulation followed by MMGBSA binding free energy calculations. This resulted in an acceptable correlation between the experimental values ( $\log K_i(\text{nM})$ ) and the MD/MM-GBSA energies ( $R^2 = 0.50$ , Fig 2C).

The experimental and computational results both agree that **BMS309403** has the strongest affinity for the target. However, any displacement or functional assay generally relies on factors additional to simple thermodynamic binding between ligand and the protein, and the way the FABP1 assay works suggests that compartmentalization into a lipid membrane might be an important factor in refining the local concentration of the ligand near the protein.<sup>1, 24</sup> However, all attempts to improve correlation with weighted correction terms representing the  $\log P$ , the molecular weight, or the solubility  $\text{PlogS}$  do not lead to improvements in  $R^2$ . Consequently, the raw induced fit binding score is used as the input, and the relaxed *apo*-form of the protein was used as the model structure.

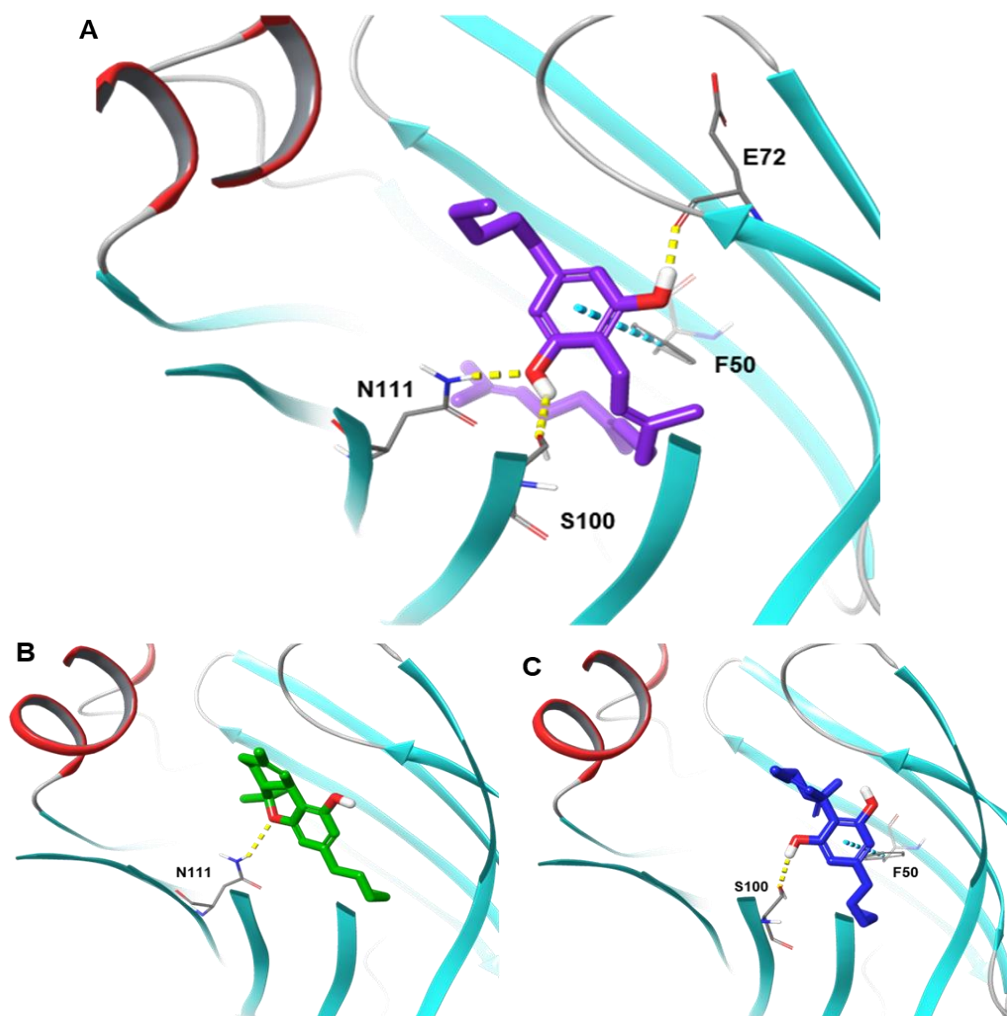
### 3-1. FABP1 and cannabinoids

In this work, a total of 140 cannabinoids were evaluated for their binding with FABP1, including 129 MCs and 11 synthetic cannabinoids. We employed XP RRD with a 0.8 scaling of the van der Waals radii of non-polar atoms for each cannabinoid ligand. After rejecting any pose with root-mean square deviation  $>2 \text{ \AA}$ , the highest scored pose for each ligand was advanced to Prime/MM-GBSA analysis to better determine the free energy of the complex (and consequently the binding energy).<sup>22</sup> Table S1 shows the calculated LogP(O/W), total CNS activity, molecular weight (MW), polar surface area (PSA), and calculated RRD and Prime/MM-GBSA values. As noted in our validation study, we expected rigid docking to not provide a good correlation to experiment, so the top 9 RRD-scored MCs, the top 9 Prime/MM-GBSA (excluding duplicates), the two top RRD-scored synthetic cannabinoids, and the biologically important **PA**, **THC**, **CBD**, and **THCA** comprised the 26 total ligands that were advanced to a more computationally expensive IFD analysis (Table 2).

**Table 2.** ADME, RRD scores, Prime/MM-GBSA and IFD and MD/MMGBSA of 26 selected ligands with FABP1.

Ligand	Physiochemical Properties			Computational (kcal/mol)						
	logPo/w	MW	PSA	RRD 0.8	Prime MM-GBSA	IFD	Receptor residues interaction			MD MM-GBSA
							N <sup>111</sup>	S <sup>100</sup>	M <sup>74</sup>	
THC(1)	5.71	314.47	26.61	-7.58	-59.77	-8.93	√	---	---	-44.07
THCA(2)	5.67	358.48	70.09	-8.24	-41.37	-9.77	√	---	---	-35.38
CBG (29)	5.82	316.48	41.95	-8.07	-69.53	-10.79	√	√	---	-48.68
CBGA (30)	5.54	360.49	84.39	-9.47	-61.18	-12.53	√	√	---	-56.52
Camagerol_1 (36)	4.25	350.50	81.37	-9.28	-69.22	-11.58	√	√	---	-39.54
Sesqui-CBG (39)	7.23	384.60	39.43	-8.32	-80.81	-12.78	√	√	---	-55.47
Sesqui-CBGA (39-2)	7.62	428.61	82.92	-8.02	-73.17	-12.69	√	√	---	-51.52
(5-acetyl-4-hydroxy-CBG) (40)	5.39	374.52	73.58	-9.57	-76.80	-12.26	√	---	---	-58.33
CBC_1 (45)	5.85	314.47	28.83	-8.22	-66.00	-10.96	√	√	---	-43.24
CBCVA_1 (48)	4.82	330.42	67.78	-9.53	-34.44	-11.37	√	---	---	-52.15
CBD (54)	5.29	314.47	39.36	-7.5	-70.53	-9.42	---	√	---	-49.63
ICBT (82)	4.07	346.47	63.12	-9.24	-56.95	-10.07	√	√	---	-44.23
ICBT-OET (85)	5.12	374.52	47.90	-8.58	-71.39	-11.26	---	√	√	-48.43
CBCN (95)	4.14	332.44	77.01	-9.43	-61.05	-10.56	√	√	√	-37.39
CBCON (99)	5.02	328.45	45.81	-8.76	-66.88	-10.87	√	---	√	-51.4
CBC-D (107)	4.53	314.42	36.26	-9.92	-69.38	-11.34	√	---	√	-42.68
THCP(121)	6.50	342.52	26.63	-8.73	-73.97	-11.56	√	√	√	-49.51
Sesqui-THC (124)	6.74	382.59	24.04	-8.5	-91.71	-12.29	---	√	---	-49.25
Sesqui-THCA (125)	7.01	426.60	63.82	-7.97	-68.58	-11.74	---	√	---	-47.87
Sesqui-THCV (126)	6.40	354.53	24.04	-7.75	-85.07	-12.59	---	√	---	-51.49
Sesqui-CBD (127)	6.72	382.59	32.29	-8.13	-78.38	-11.91	---	√	---	-51.65
AEA (a1)	6.67	345.57	37.94	-7.34	-78.07	-9.53	---	---	√	-51.01
2-AG (a2)	6.67	378.55	74.91	-8.68	-84.06	-9.16	---	---	√	-56.76
AM12033 (a7)	5.20	413.60	74.37	-9.23	-65.64	-13.07	√	√	---	-53.45
Nabilone (a11)	5.52	372.55	55.75	-9.89	-70.04	-12.37	√	√	---	-48.39
PA	5.00	256.43	50.00	-6.25	-8.55	-12.53	√	---	---	-56.31

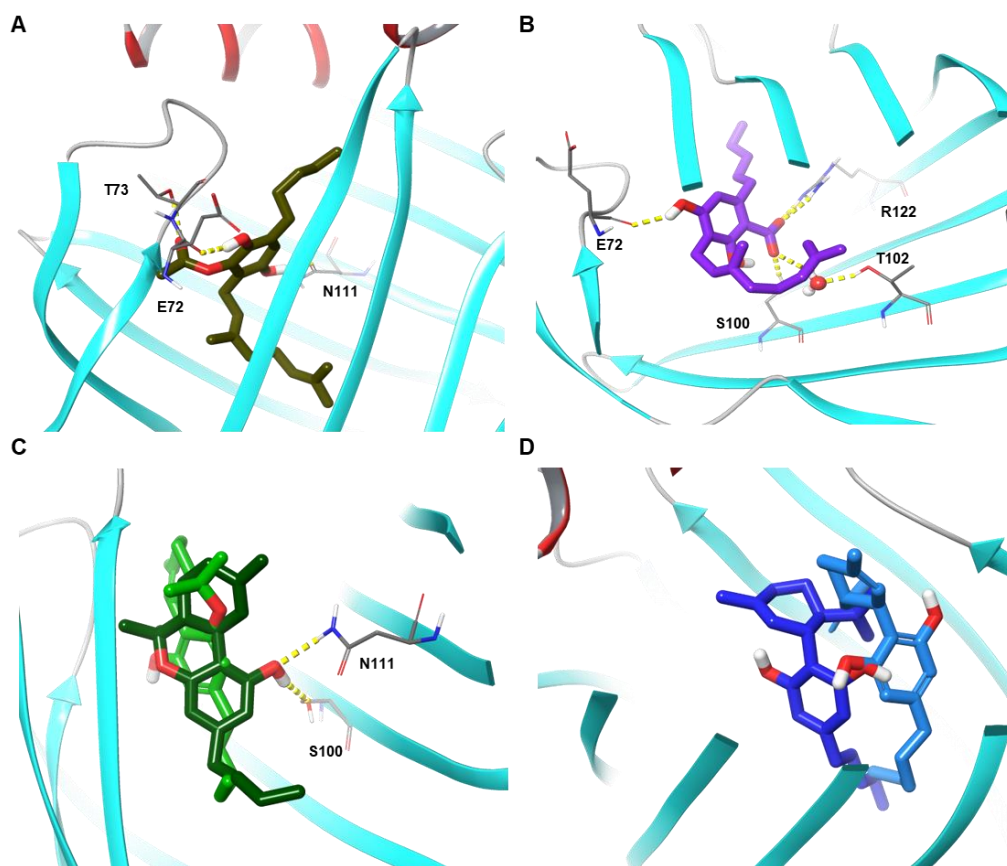
This analysis shows that, by IFD, many phytocannabinoids are predicted to have meaningful affinity for FABP1, ranging from -8.93 to -18.05 kcal/mol; many of these are predicted to be tighter binders than the designed FABP inhibitors (Table 1), and certainly stronger than the **ECs** (Table 2). All 23 examined cannabinoids have strong interactions with at least one of **N<sup>111</sup>**, **S<sup>100</sup>**, or **M<sup>74</sup>** either directly or through a water-mediated hydrogen bond. This interaction is what makes them strong binders. Hydrophobicity of course also plays a role in determining binding to a fatty-acid binding pocket: as the polarity of the ligand increases, the stronger the desolvation penalty, and consequently the weaker the overall binding within the hydrophobic pocket.<sup>12</sup> Another significant class of binders are the longer chain homologues of the common cannabinoids, the sesquicannabinoids which have a 7-membered alkyl chain on their resorcinol in place of **THC** and **CBD**'s pentyl chain. **Sesqui-CBG** forms a  $\pi$ -cation interaction with **F<sup>50</sup>** and the phenolic hydroxyl groups form H-Bonds with **E<sup>72</sup>**, **S<sup>100</sup>** and **N<sup>111</sup>**. Its long alkyl chain extends deeply into the pocket and forms an L-shape hydrophobic interaction with the base  $\beta$ -sheet floor of FABP1 (Fig. 4A). The major cannabinoids are by far the weakest predicted binders of those investigated: for both **THC** and **CBD** the O<sub>5</sub> atom of their ring systems forms H-Bonds with **N<sup>111</sup>**. **CBD** forms a  $\pi$ -cation interaction with **F<sup>50</sup>**, and its phenolic hydroxyl group forms an H-Bond with **S<sup>100</sup>** (Fig. 4B and 4C).



**Figure 4.** IFD binding poses of (A) **Sesqui-CBG** (MD: purple) (B) **THC** (dark green) and (C) **CBD** (dark blue) in complex with FABP1 (PDB ID: 6MP4). The oxygen atoms are in red and nitrogen in blue, H-bonds in yellow dotted lines,  $\pi$ -cation interaction in dark green dotted lines.

Since the docking scores suggest most of the cannabinoids have acceptable binding scores, and all higher than **THC**, an MD simulation was performed for all the cannabinoids in Table 2 to calculate the MM-GBSA binding energy of ligands to the protein. It is important to note that these values cannot be interpreted as absolute affinities and are most useful as relative values; the MM-GBSA calculations do not necessarily consider entropic effects and often overestimate binding energies significantly. But this has generally been found to be a systematic error meaning the tool is useful to rank binders, if not predict their actual energy of binding.<sup>25</sup>

The first and second ranked MCs **5-acetyl-4-hydroxy-CBG** and **CBGA** have MM-GBSA-predicted binding energies of  $-58.3$  and  $-56.5$  kcal/mol respectively. These are in the range of reference values for the experimental ligands (Figure 3, Table 1). These values are close to that of **BMS309403**, the strongest experimentally validated binder to FABP1 ( $-63.0$  kcal/mol). **5-acetyl-4-hydroxy-CBG** fully occupies the binding site and forces it to close around it, as is typical of FABP binders. The phenolic oxygen directly interacts with **E<sup>72</sup>**, **T<sup>73</sup>** & **N<sup>111</sup>** (Fig. 5A). **CBGA**, after MD simulation, moves slightly and its acidic group forms an H-bond with **S<sup>100</sup>**, **T<sup>102</sup>**, & **R<sup>122</sup>**; likewise, its phenolic OH H-bonds with **E<sup>72</sup>** (Fig. 5B). The MD simulation leads to a complete flip for **THC (Fig S1b)**, its binding is now completely different from that observed by IFD, with its phenolic group now forming H-bond interactions with both **S<sup>100</sup>** and **N<sup>111</sup>** (Fig. 5C). Here the phenol is attempting to mimic the hydrogen bond normally formed by a carboxylic acid. It is important to note that it is the IFD structure that aligns with the crystal structure; however, both poses have similar low energies, suggesting both poses are possible. **CBD** only moves very little; its overall structure remains the same as the IFD docking structure, and this also means that it has no specific conserved H-Bond interactions (Fig. 5D). The sesqui-terpene minor cannabinoids maintain good binding energy to FABP1, and are also worthy of further experimental study.



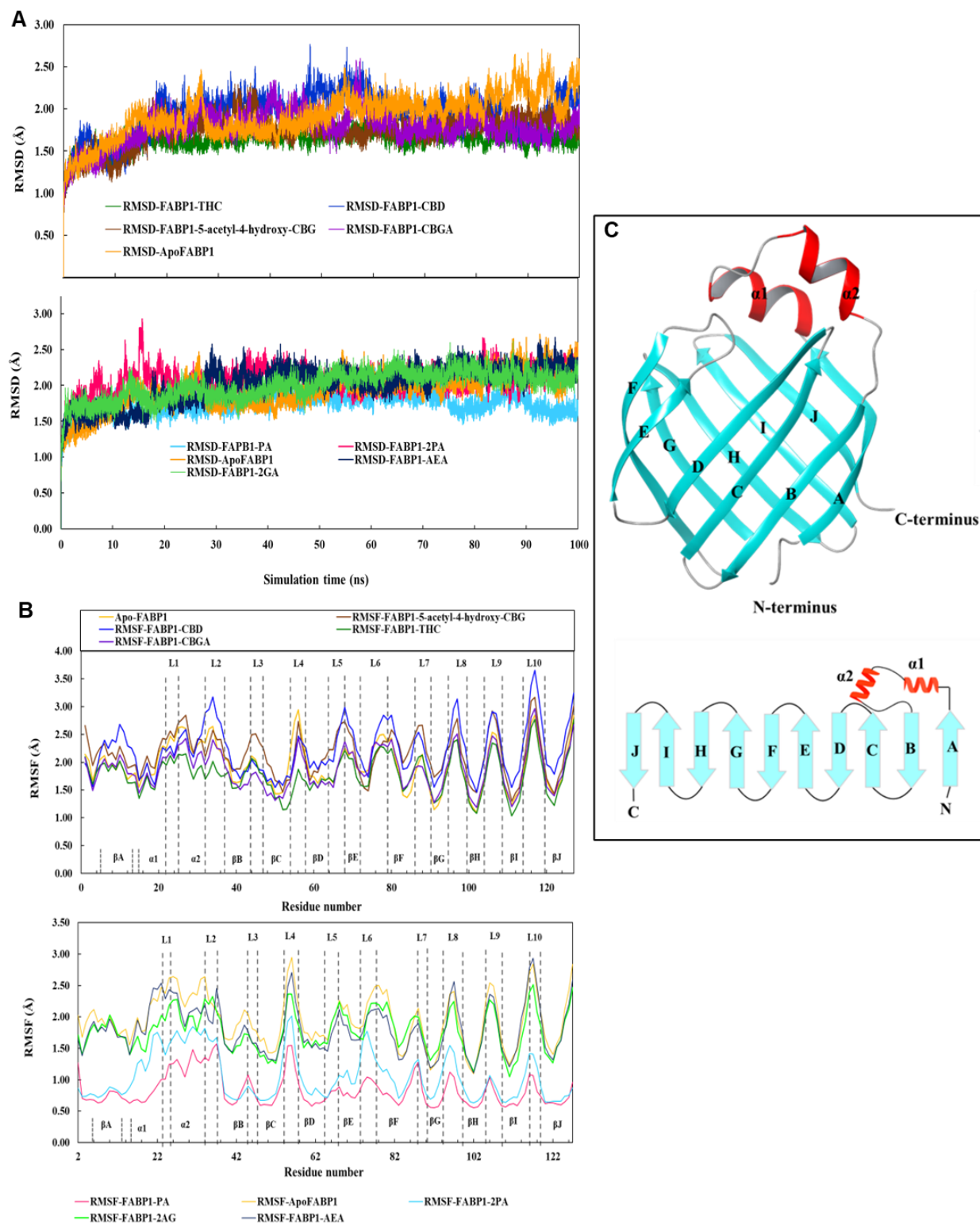
**Figure 5.** The most representative structure of the lowest energy conformation of the complex obtained by clustering of the MD for, (A) **5-acetyl-4-hydroxy-CBG** (brown) (B) **CBGA** (MD: purple) (C) **THC** (dark green) and (D) **CBD** (dark blue) in complex with FABP1 (PDB ID: *6MP4*). The oxygen atoms are in red, nitrogen is in blue, and H-bonds are represented yellow dotted lines.

To validate the stability of the ligand binding mode in the complexes, RMSD of the backbone atoms relative to first frame of the MD simulation were calculated for the first 100 ns (Fig. 6A). All complex structures reached equilibrium after no more than 16 ns. The average RMSD value calculated for FABP1 in complex with **5-acetyl-4-hydroxy-CBG**, **CBGA**, **CBD**, **THC**, **PA**, two molecules of **PA**, **AEA**, **2GA** and apo protein were 1.74, 1.77, 1.94, 1.63, 1.74, 2.07, 2.01, 1.99 and 1.92 angstrom respectively. These reasonable RMSD values strongly imply the complex remains largely stable. These don't meaningfully suggest major differences between the ligands with the possible suggestion that the phytocannabinoids restrict overall motion more than either the ECs or the FAs. However, the RMSD, being calculated over the entire complex can



mask substantial flexibility in some components, or can overstate the movement of other components.

Consequently, per residue root mean square fluctuations (RMSFs) of the protein backbone in FABP1 were calculated (Fig 6B). The protein clearly, in general, becomes less flexible when a ligand is bound, particularly in the portal region, which is composed of  $\alpha$  helix 2,  $\beta$  turn CD, and  $\beta$  turn EF (Figure 6C). **CBGA**, **5-acetyl-4-hydroxy-CBG**, **THC** and **CBD** reduced the flexibility of  $\alpha$  helix 2. Of all the ligands examined, **CBGA** and **THC** are the most effective in reducing the flexibility of  $\alpha$  helix 2.  $\beta$ -turn EF also becomes stabilized upon binding the ligands. This reduction in flexibility of these motifs is consistent with the observed interactions of the ligands with **E**<sup>72</sup>, **T**<sup>73</sup>, **R**<sup>122</sup>, **S**<sup>100</sup> and **N**<sup>111</sup>. Interaction locks them into place relative to one another, and would be expected to reduce their motion, and consequently that of their parent secondary motif.



**Figure 6.** A) Root-mean-square deviations (RMSD) and B) Root mean square fluctuations (RMSFs) of backbone atoms relative to their initial minimized structures as function of time for FABP1.

However, the greatest reduction in flexibility occurs when the endogenous ligand **PA** binds. The reduction in flexibility is even greater when two molecules of **PA** bind to the cavity. The effect is strongest at  $\alpha$  helix 2,  $\beta$  turn CD, and  $\beta$  turn EF. These are the same motifs affected by the ECs and MCs. This effect is likely driven by the conserved hydrogen bonds between the **PA**s and key residues **R**<sup>122</sup>, **S**<sup>39</sup> and **S**<sup>124</sup> (Table 3). The same, although lesser, decreased flexibility of the portal region is observed upon binding of the natural ECs **2-AG** and **AEA**. Analysis of hydrogen bonds between these ligands and FABP1 revealed that **E**<sup>72</sup> and **M**<sup>74</sup> created a hydrogen bond with **2-AG** which may result in reduction of flexibility in portal region of protein. However, no hydrogen bonds between **AEA** and any residues of FABP1 was detected. Comparing hydrogen bonds pattern between ligands in FABP1 in complex with ECs and **CBGA** and **5-acetyl-4-hydroxy-CBG** revealed that the hydrogen bond between ligands and **E**<sup>72</sup> become more stable (Table 3).

**Table 3.** Hydrogen bonds formed between the ligands and FABP1.

Ligand	Donor	Acceptor	Average Distance (Å)	%Occupancy
<b>5-acetyl-4-hydroxy-CBG</b>	5-acetyl-4-hydroxy-CBG@O2	E <sup>72</sup> @O	2.77	60.58
	T <sup>73</sup> @OG1	5-acetyl-4-hydroxy-CBG@O4	2.76	26.5
	5-acetyl-4-hydroxy-CBG@O1	N <sup>111</sup> @OD1	2.75	23.66
	N <sup>111</sup> @ND2	5-acetyl-4-hydroxy-CBG@O1	2.87	17.35
<b>CBGA</b>	R <sup>122</sup> @NH2	CBGA@O4	2.83	63.3
	CBGA@O1	E <sup>72</sup> @O	2.76	61.78
	R <sup>122</sup> @NH1	CBGA@O4	2.78	50.83
	R <sup>122</sup> @NH1	CBGA@O3	2.77	41.56
	S <sup>100</sup> @OG	CBGA@O3	2.61	41.45
	R <sup>122</sup> @NH2	CBGA@O3	2.81	33.74
	CBGA@O2	S <sup>100</sup> @OG	2.8	25.09
<b>THC</b>	THC @O2	E <sup>72</sup> @O	2.73	86.6
<b>PA</b>	R <sup>122</sup> @NH2	PA@O1	2.79	76.5
	S <sup>39</sup> @OG	PA@O1	2.63	73.42
	R <sup>122</sup> @NE	PA@O2	2.82	63.3
	S <sup>124</sup> @OG	PA@O2	2.69	57.21
	R <sup>122</sup> @NH2	PA@O2	2.83	31.1
	S <sup>39</sup> @OG	PA@O2	2.61	17.01
	R <sup>122</sup> @NE	PA@O1	2.81	16.99
	S <sup>124</sup> @OG	PA@O1	2.72	14.75
<b>2PA</b>	R <sup>122</sup> @NH2	PA2*@O1	2.76	67.4
	S <sup>39</sup> @OG	PA1*@O2	2.68	57.3

	R <sup>122</sup> @NH1	PA1@O1	2.8	51.12
	R <sup>122</sup> @NH2	PA1@O2	2.8	43.66
	S <sup>124</sup> @OG	PA1@O1	2.67	40.84
	S <sup>100</sup> @OG	PA2@O2	2.62	40.17
	R <sup>122</sup> @NH1	PA1@O2	2.8	39.65
	R <sup>122</sup> @NH2	PA1@O1	2.7	38.09
	S <sup>39</sup> @OG	PA1@O1	2.66	34.5
	R <sup>122</sup> @NH2	PA2@O2	2.78	22.7
	R <sup>122</sup> @NE	PA2@O2	2.81	21.93
	R <sup>122</sup> @NE	PA2@O1	2.83	19.54
	N <sup>111</sup> @ND2	PA2@O1	2.8	15.44
	S <sup>124</sup> @OG	PA1@O2	2.66	14.67
<b>2-AG</b>	M <sup>74</sup> @N	2-AG@O4	2.87	34.61
	2-AG@O2	E <sup>72</sup> @O	2.74	14.92
<b>AEA</b>	---	---	---	---

\* PA1 and PA2 are first and second molecules of PA

It is curious that such structurally diverse molecules can show similar high affinity for a protein like FABP1.<sup>26</sup> Computational techniques can help us to evaluate molecular similarity. These techniques can be categorized into three methods; in the first method we can conduct QSAR analyses comparing chemical and molecular properties such as lipophilicity (logPo/w), molecular weight (MW) and polar surface area (PSA) of FAs, synthetic ligands, ECs and MCs. Using this approach reveals no significant trends between these parameters and either experimental binding affinity, or computationally calculated IFD or MD/MM-GBSA values. The second method, employing the Tanimoto similarity measure (which provides a coefficient, *Tc*, between 1 to 0) is common in 2D SAR approaches.<sup>27</sup> According to this analysis, **PA** is more similar to the ECs (0.4 similarity between PA and 2-AG) with far less similarity to the MCs (0.13 similarity between PA and CBD). If one was to use these parameters to predict binding to the protein, one would be sorely mistaken: molecules that are structurally dissimilar tend to bind in a dissimilar fashion, and therefore tend to induce different bioactivities. The important words in that sentence are, of course “tend,” and caution must be taken whenever one abstracts a molecule for analysis. A third abstracting approach is the use of interaction fingerprints (IFP) on the bound structure: identifying the residues involved in strong interactions with a given ligand. The key residues identified by an IFP analysis of the FAs binding to FABP1 are **R<sup>122</sup>**, **N<sup>111</sup>**, **S<sup>100</sup>** and **M<sup>74</sup>**. The ECs have very little IFP similarity, but the MCs **5-acetyl-4-hydroxy-CBG** and **GBGA** do share the same residues and have a good IFD match to the FAs (Table 3). However, this type of analysis only works if you

have structural data or a good all atomic simulation. We cannot recommend using any of these approaches to interpret the interactions between the ligands and the receptor, and we strongly recommend that any computational screening technique take all-atomic modelling into account to make any meaningful predictions.

## Conclusion

What is clear from this data is that the ECs can avail themselves of the same transport mechanisms as the FAs. This is well established. The major cannabinoids are known to bind to FABP1, and this activity has been suggested to be responsible for some of their biological activity. However, we show that the strongest affinities likely arise from the interaction of minor cannabinoids with the FABPs. Of course, they are present in lower amounts than THC or CBD, but if one were looking to design an FABP1-inhibiting drug, the minor cannabinoids are likely a better starting point than the major cannabinoids.

Most curiously is that this study provides additional evidence that the PCs imitate the preferred accessible conformation of the ECs. Razdan in 1996,<sup>28</sup> and Howlett in 1998<sup>29</sup> both discussed the pharmacophores of the PCs and ECs, and research since then has tried to combine these two systems into simple cannabinoid receptor agonists and antagonists.<sup>15</sup> However, this similarity is clearly more general and applies to their interactions with other proteins as well. As both we and these authors note, the constrained ring system of the PCs “pays” the entropic cost of binding in a specific conformation up front, increasing binding affinity. This has been clearly discussed in the context of the cannabinoid receptors, but we believe this is the first report suggesting the effect is more general and likely applies to other proteins as well. The effect is also clearly not unique to the ECs. The PCs can imitate far more ubiquitous FAs. Examining other proteins that interact with FAs might prove to be a promising avenue of research to further understand the complex pharmacology of the PCs. Some of this work is currently underway in our lab and will be reported on in due course.

## Supplementary Information and Data Availability:

The supplementary information that accompanies this article includes a more detailed discussion of the computational methodology used to generate the data, complete tables for the rigid binding data for all 131 cannabinoids examined in the article, and 2D interaction plots for FABP1's

endogenous and designer ligands highlighting key interactions. All of the computational input and output geometries, and the required information to recreate the MD trajectories, and consequently all the data in the article, is available from the Borealis repository at: <https://doi.org/10.5683/SP3/9HCGOM>. This also includes the *apo*-FABP1 structure we use as the basis of our model should anyone wish to use it for further drug discovery work. The Borealis repository is a collaboration of the Canadian Universities to facilitate access to research data.

### Author Credit Statement:

Conceptualization, JFT, FS; Funding acquisition JFT; Investigation, FS, SM, DM, VA, VT; Methodology, FS, SM, DM, JFT; Visualization, FS, SM; Project administration, JFT; Supervision, JFT, FS; Writing—original draft, FS; Writing—review and editing, All authors.

### Acknowledgements:

The authors gratefully acknowledge financial support for the project from the Natural Sciences and Engineering Research Council of Canada (JFT: grant # 2018-06338). All authors wish to recognize that this work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: [www.sharcnet.ca](http://www.sharcnet.ca)) and Compute/Calcul Canada, now known as the Digital Alliance of Canada (<https://alliancecan.ca/en>).

### References

1. Storch, J.; McDermott, L., Structural and functional analysis of fatty acid-binding proteins. *J. Lipid Res.* **2009**, *50 Suppl* (Suppl), S126-31.
2. Atshaves, B. P.; Martin, G. G.; Hostetler, H. A.; McIntosh, A. L.; Kier, A. B.; Schroeder, F., Liver fatty acid-binding protein and obesity. *J. Nutr. Biochem.* **2010**, *21* (11), 1015-1032.
3. Wang, G.; Bonkovsky, H. L.; de Lemos, A.; Burczynski, F. J., Recent insights into the biological functions of liver fatty acid binding protein 1. *J. Lipid Res.* **2015**, *56* (12), 2238-2247.
4. Hirowatari, K.; Kawano, N., Association of urinary liver-type fatty acid-binding protein with renal functions and antihyperglycemic drug use in type 2 diabetic nephropathy patients. *Int. J. Urol. Nephrol.* **2023**.
5. Martin, G. G.; Atshaves, B. P.; McIntosh, A. L.; Mackie, J. T.; Kier, A. B.; Schroeder, F., Liver fatty acid binding protein gene ablation potentiates hepatic cholesterol accumulation in cholesterol-fed female mice. *Am. J. Physiol. Gastrointest. Liver Physiol.* **2006**, *290* (1), G36-G48.
6. (a) Wu, Y. L.; Peng, X. E.; Zhu, Y. B.; Yan, X. L.; Chen, W. N.; Lin, X., Hepatitis B virus X protein induces hepatic steatosis by enhancing the expression of liver fatty acid binding protein. *J. Virol.* **2016**, *90* (4), 1729-40; (b) Pi, H.; Liu, M.; Xi, Y.; Chen, M.; Tian, L.; Xie, J.; Chen, M.;



Wang, Z.; Yang, M.; Yu, Z.; Zhou, Z.; Gao, F., Long-term exercise prevents hepatic steatosis: A novel role of FABP1 in regulation of autophagy-lysosomal machinery. *FASEB J.* **2019**, *33* (11), 11870-11883.

7. McKillop, I. H.; Girardi, C. A.; Thompson, K. J., Role of fatty acid binding proteins (FABPs) in cancer development and progression. *Cell. Signalling* **2019**, *62*, 109336.

8. Sharma, A.; Sharma, A., Fatty acid induced remodeling within the human liver fatty acid-binding protein. *J. Biol. Chem.* **2011**, *286* (36), 31924-31928.

9. Chuang, S.; Velkov, T.; Horne, J.; Wielens, J.; Chalmers, D. K.; Porter, C. J. H.; Scanlon, M. J., Probing the fibrate binding specificity of rat liver fatty acid binding protein. *J. Med. Chem.* **2009**, *52* (17), 5344-5355.

10. Chuang, S.; Velkov, T.; Horne, J.; Porter, C. J. H.; Scanlon, M. J., Characterization of the drug binding specificity of rat liver fatty acid binding protein. *J. Med. Chem.* **2008**, *51* (13), 3755-3764.

11. Richieri, G. V.; Ogata, R. T.; Kleinfeld, A. M., Equilibrium constants for the binding of fatty acids with fatty acid-binding proteins from adipocyte, intestine, heart, and liver measured with the fluorescent probe ADIFAB. *J. Biol. Chem.* **1994**, *269* (39), 23918-30.

12. Elmes, M. W.; Prentis, L. E.; McGoldrick, L. L.; Giuliano, C. J.; Sweeney, J. M.; Joseph, O. M.; Che, J.; Carbonetti, G. S.; Studholme, K.; Deutsch, D. G.; Rizzo, R. C.; Glynn, S. E.; Kaczocha, M., FABP1 controls hepatic transport and biotransformation of  $\Delta^9$ -THC. *Sci. Rep.* **2019**, *9* (1), 7588.

13. Shahbazi, F.; Grandi, V.; Banerjee, A.; Trant, J. F., Cannabinoids and cannabinoid receptors: The story so far. *iScience* **2020**, *23* (7), 101301.

14. Martin, G. G.; Chung, S.; Landrock, D.; Landrock, K. K.; Huang, H.; Dangott, L. J.; Peng, X.; Kaczocha, M.; Seeger, D. R.; Murphy, E. J.; Golovko, M. Y.; Kier, A. B.; Schroeder, F., FABP-1 gene ablation impacts brain endocannabinoid system in male mice. *J. Neurochem.* **2016**, *138* (3), 407-22.

15. Chen, J.-Z.; Han, X.-W.; Xie, X.-Q., Preferred conformations of endogenous cannabinoid ligand anandamide. *Life Sci.* **2005**, *76* (18), 2053-2069.

16. Citti, C.; Braghiroli, D.; Vandelli, M. A.; Cannazza, G., Pharmaceutical and biomedical analysis of cannabinoids: A critical review. *J. Pharm. Biomed. Anal.* **2018**, *147*, 565-579.

17. Wolfrum, C.; Borrmann, C. M.; Borchers, T.; Spener, F., Fatty acids and hypolipidemic drugs regulate peroxisome proliferator-activated receptors  $\alpha$  - and  $\gamma$ -mediated gene expression via liver fatty acid binding protein: a signaling path to the nucleus. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98* (5), 2323-8.

18. Tahir, M. N.; Shahbazi, F.; Rondeau-Gagné, S.; Trant, J. F., The biosynthesis of the cannabinoids. *J. Cannab. Res.* **2021**, *3*, 7 (Article Number).

19. Felletti, S.; Compagnin, G.; Krauke, Y.; Stephan, S.; Greco, G.; Buratti, A.; Chenet, T.; De Luca, C.; Catani, M.; Cavazzini, A., Purification and isolation of cannabinoids: Current challenges and perspectives. *LCGC Europe* **2023**, *36* (04), 122-131.

20. Huang, H.; McIntosh, A. L.; Martin, G. G.; Landrock, D.; Chung, S.; Landrock, K. K.; Dangott, L. J.; Li, S.; Kier, A. B.; Schroeder, F., FABP1: A novel hepatic endocannabinoid and cannabinoid binding protein. *Biochemistry* **2016**, *55* (37), 5243-5255.

21. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L., Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47* (7), 1750-9.

22. Lyne, P. D.; Lamb, M. L.; Saeh, J. C., Accurate prediction of the relative potencies of members of a series of kinase inhibitors using molecular docking and MM-GBSA scoring. *J. Med. Chem.* **2006**, *49* (16), 4805-8.
23. Zhong, H.; Tran, L. M.; Stang, J. L., Induced-fit docking studies of the active and inactive states of protein tyrosine kinases. *J. Mol. Graphics Modell.* **2009**, *28* (4), 336-346.
24. Velkov, T.; Chuang, S.; Wielens, J.; Sakellaris, H.; Charman, W. N.; Porter, C. J. H.; Scanlon, M. J., The interaction of lipophilic drugs with intestinal fatty acid-binding protein. *J. Biol. Chem.* **2005**, *280* (18), 17769-17776.
25. (a) Hou, T.; Wang, J.; Li, Y.; Wang, W., Assessing the performance of the MM/PBSA and MM/GBSA Methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model.* **2011**, *51* (1), 69-82; (b) Hou, T.; Wang, J.; Li, Y.; Wang, W., Assessing the performance of the molecular mechanics/Poisson Boltzmann surface area and molecular mechanics/generalized Born surface area methods. II. The accuracy of ranking poses generated from docking. *J. Comp. Chem.* **2011**, *32* (5), 866-877; (c) Rastelli, G.; Rio, A. D.; Degliesposti, G.; Sgobba, M., Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. *J. Comp. Chem.* **2010**, *31* (4), 797-810.
26. Xu, X.; Zou, X. Dissimilar ligands bind in a similar fashion: A guide to ligand binding-mode prediction with application to CELPP studies *Int. J. Mol. Sci.* [Online], 2021.
27. Bajusz, D.; Rácz, A.; Héberger, K., Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **2015**, *7* (1), 20.
28. Thomas, B. F.; Adams, I. B.; Mascarella, S. W.; Martin, B. R.; Razdan, R. K., Structure–activity analysis of anandamide analogs: Relationship to a cannabinoid pharmacophore. *J. Med. Chem.* **1996**, *39* (2), 471-479.
29. Tong, W.; Collantes, E. R.; Welsh, W. J.; Berglund, B. A.; Howlett, A. C., Derivation of a pharmacophore model for anandamide using constrained conformational searching and comparative molecular field analysis. *J. Med. Chem.* **1998**, *41* (22), 4207-4215.

# Localization of broken instruments inside the root canal using pulse-echo mode of ultrasonic signal

Rahul Kumar  
ECE Department  
Delhi Technological University  
New Delhi, India

Rajiv Kapoor  
ECE Department  
Delhi Technological University  
New Delhi, India

**Abstract**— Localization of broken instruments inside the root canal (RC) during root canal Treatment (RCT) is one of the most challenging tasks for an endodontist surgeon. This is due to the lack of real time systems for localization of broken instruments. This paper suggests a method for the identification of the broken instrument inside the RC. In this paper a novel ultrasonics signal recording method for tooth with the CNN architecture of VGG16 as a feature extractor and SVM (support vector machine) for the localization of the broken instruments are used. With this paper, authors localize the broken instrument inside the canal at three positions, i.e., top, bottom, and middle. The accuracy of the system is achieved by 76% with proposed methodology.

**Keywords**— broken instruments, CNN, endodontics, localization, VGG16, SVM.

## I. INTRODUCTION

In endodontics, RCT is a common practice for treatment of RC [1]. In which the dental surgeon used to identify the RC diseases. For the diagnosis of diseases, there are lots of techniques available. Some of the major techniques are x-ray, cold test, bioimpedance meter, apex locator, etc. But all these techniques have major issues like false detection, inappropriate testing procedures, hazardous effects on human tissue, etc. While ultrasound is one of the safest technologies for diagnosis [2]. It is a mechanical wave, i.e., acoustic signal having a frequency greater than 20 kHz, as shown in figure. 1. which vibrates and travels through the medium like liquid, gas, and solid. Sound can be classified based on its frequency range. The frequency range is distinguished according to the range of human hearing as shown in figure 1.

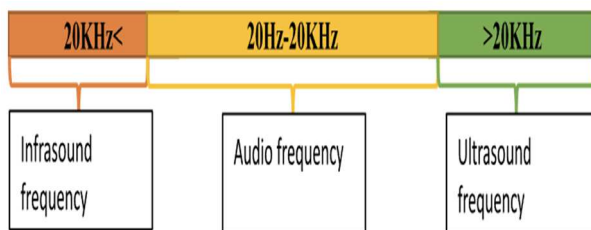


Figure 1: Sound Energy classification

In medical ultrasound, frequencies above 2 MHz are used [3]. The ultrasonography used the refracted and reflected energy of the acoustic signal. These signals carry information about the internal tissue structure. While in dentistry, it has a very limited scope because of the differences in tooth structure and properties. tooth has a wide range of acoustic properties ranging from 1.5Mrayl to 7.8Mrayl. Therefore, conventional ultrasonography techniques are not suited for dental applications.

In [4], authors mention the application of ultrasound in dentistry. The mention applications are used to extract the information from canal, scaling of tooth, cleaning of tooth and etc. Authors also mention that, the ultrasound can be used for imaging of RC. While First-time work of ultrasound in dentistry was observed by Baum et al., [5] in 1963. In this study, authors used a 15 MHz ultrasonic transducer, but they did not get a clear information of internal tooth structure. In [6], authors used ultrasound, in which a customized probe is attached with an image processing unit and an electronics phantom. The shape of the customized probe is similar to dental handpiece equipment for easy movement inside the mouth. In this work, authors used the 25 MHz ultrasound frequency on three different age groups, i.e., 22, 33, and 59-year-old patients with no periodontal disease. Image taken from top part of the gingiva. Ultrasound image quality is good enough. But the authors did not get any evidence of a RC of the tooth. In [7], authors used the 15 MHz, ultrasound frequency for the observation of teeth's internal structure. The image of the US is not very clear. According to them, if the resolution of the system is increased then detection of caries will also increase. In [8], authors proposed a technique for the measurement of thickness of tooth layer using fractional Fourier transformation. In which for signal recording, they used Olympus NDT inc. Waltham MA based probe with 15 MHz supportive ultrasonic frequency. The probe is in contact with the sample. Authors used glycerin as the matching layer. In [9], Kapoor R. et. al, proposed the methods i.e., Modified Sliding Singular Spectrum Analysis (MSSSA), for detecting foreign objects in the RC, in which authors use the continuous mode of ultrasound signal.

The major contributions of this paper as follow:

- Ultrasonics signal generation and acquisition for analysis of RC of vitro teeth.
- Features extraction of recorded ultrasonics signal using VGG16 methodology.
- Localization of broken instruments inside the canal using SVM.

This paper is organized in following sections: Section 3 presents the proposed methodologies, in which novel recoding techniques for localization of broken instruments, features extractor and the used classifier are explained. Section 4 discusses the outcomes of this research work. In section 5, the conclusions are drawn and the future aspects of the proposed scheme are discussed.

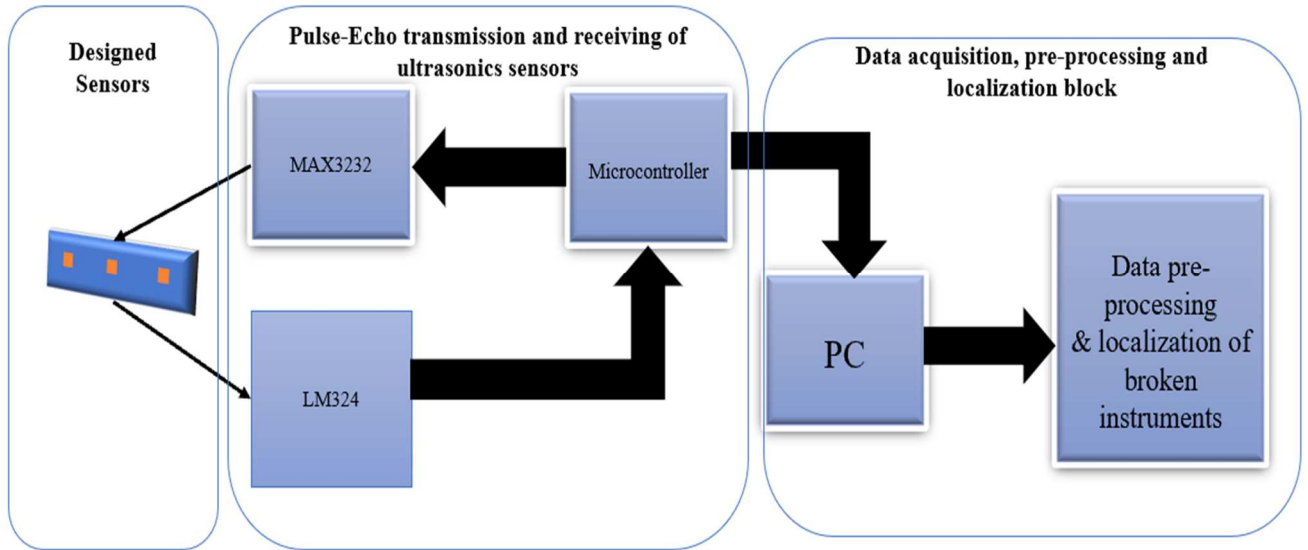


Figure 2: Block diagram of proposed method for localization of broken instruments

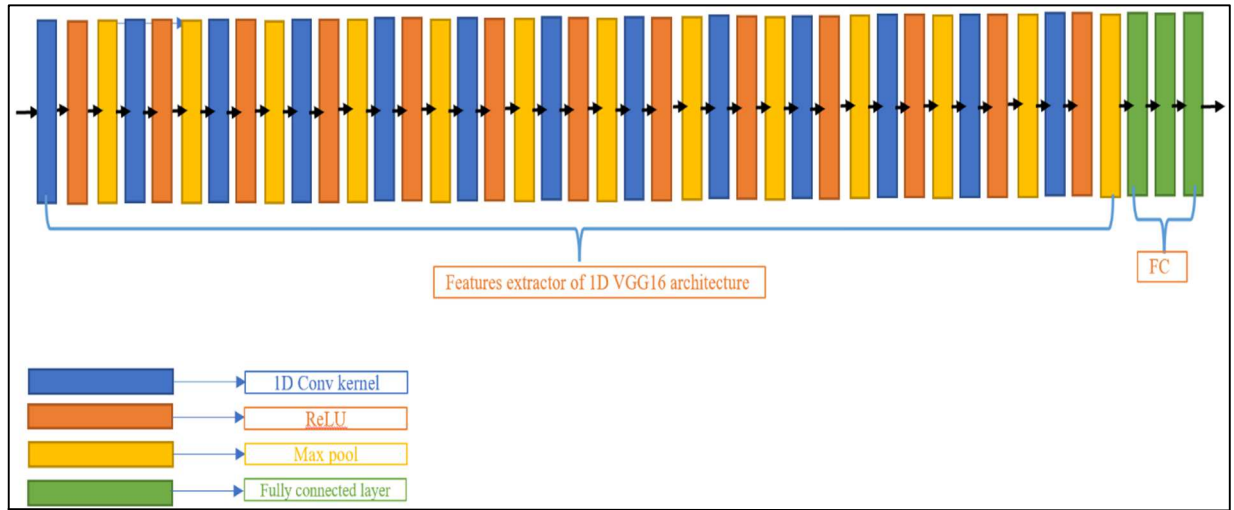


Figure 3: 1D VGG16 as features extractor with batch normalization

## II. PROPOSED METHODOLOGY

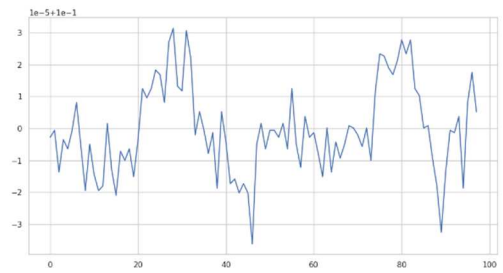
This paper proposed a novel localization methodology for the detection of broken instruments inside the RC as shown in Figure. 2. This paper used 1D VGG16[10] architecture as the feature extractor in conjunction with SVM for localization of broken instruments. The signals are recorded with a novel pulse-echo mode technique.

### A. Data acquisition

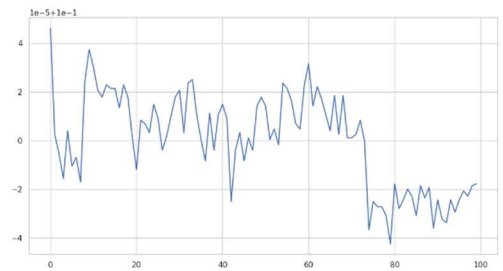
The ultrasonic signal is generated when the transducer vibrates at a frequency above 20 kHz. In this paper, piezoelectric ceramic of 3x3 mm is used. Designed transducers have three arrays, as shown in Figure. 2 (block 1). The designed transducer is used as a transceiver. The proposed scheme of recording set-up is shown in Figure. 2 (block 1 & 2). More than 100 vitro teeth are used for recording the signals for analysis the methodology a frequency of 33 KHz is used. The recorded echo signals are represented in (1).

$$U_e = \sum_{i=0}^N s_i \quad s_i \in \mathbb{R} \quad (1)$$

Where,  $U_e$  is ultrasonic echo signal,  $s_i$  is the instant value of signal at time instant “i”.



(a)



(b)



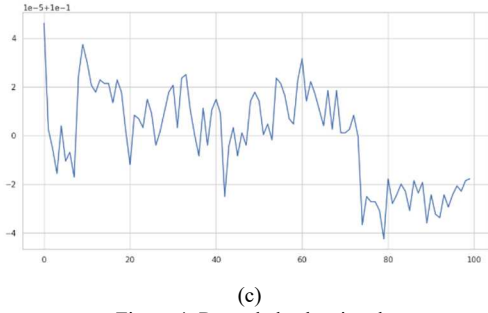


Figure 4: Recorded echo signal

In figure 4 shows the ultrasonics echo signals of designed probe. In which figure. 4 (a), (b), and (c), represents the signal recovered from first, second and third sensors of mentioned probe and the features map of the recorded signal is extracted using mention in (2).

$$f_{U_e} = U_e * CAC \quad (2)$$

The features map of the echo signal using CAC, where CAC is represents the Convolution Architecture of CNN.

### B. Features extraction

The VGG16 is a deep convolutional neural network that has attained significant performance while dealing with various tasks related to computer vision [11]. Another version i.e., 1D VGG16 is designed for the analysis and classification of time-series data [12]. In this paper, 1D VGG16 without batch normalization is used as CAC to extract high-level features from recorded acoustic signals.

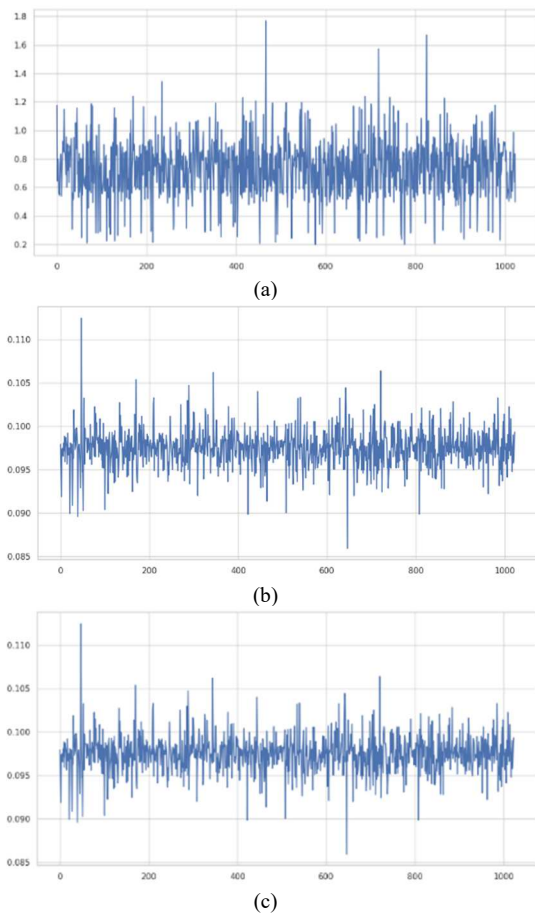


Figure 5: Features map using VGG16

One way to use VGG16 as a feature extractor is to remove the last fully connected layer of the network and use the output

of the layer before it as the feature representation, as shown in Figure. 3. The extracted features shown in figure (5) are further used by SVM for the localization of broken instruments.

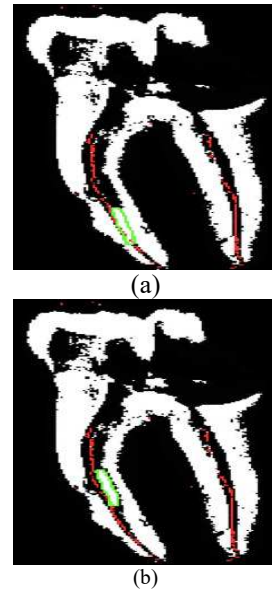
### C. Localization

Support Vector Machine (SVM) is a supervised learning (SL) based classifier for binary as well as multi-class classification applications. SVM is used for One-vs-One (OvO) as well as One-vs-All (OvA) kind of schemes [13]. In our case, the OvO trains three SVM classifiers and during the prediction stage the classifier evaluates instances with respect to other classes. The OvA method also trains three different SVM classifiers where each trained classifier distinguishes one class from the combination of other two classes. During its prediction stage, all three positions are assigned to a test instance to find the class with highest possibility.

In this paper, SVM is used for the localization of broken instruments into three classes, i.e., top, middle, and bottom positions. In the proposed methods, the OvA method is used for the localization of broken instruments. Therefore, this approach gets results faster than OvO and requires less computational memory [14].

## III. RESULTS

Before The results and accuracy of the proposed methodology are going to be explained in this section. The localization of broken instruments inside the RC is predicated based on the proposed methodology. While the predicted results are verified with the labelled x-ray of the tooth and by the endodontic surgeon. The results of broken instruments' localization inside the canal are shown in Figure. 6. The x-ray image of a tooth with broken instruments is displayed according to the prediction of the SVM localizer. The system achieves an accuracy of approximately 76% with proposed methods. The accuracy will improve with a greater number of sensors array and datasets. The proposed methodology is implemented and trained on a system having a configuration of 16GB RAM, a RYZEN 9 (5800 HS) processor, 6GB of NVIDIA RXT 3060 GPU, and a 1TB SSD.





(c)

Figure 6: (a) tooth with broken instruments at bottom, (b) tooth with broken instruments at middle (c) tooth with broken instruments at top.

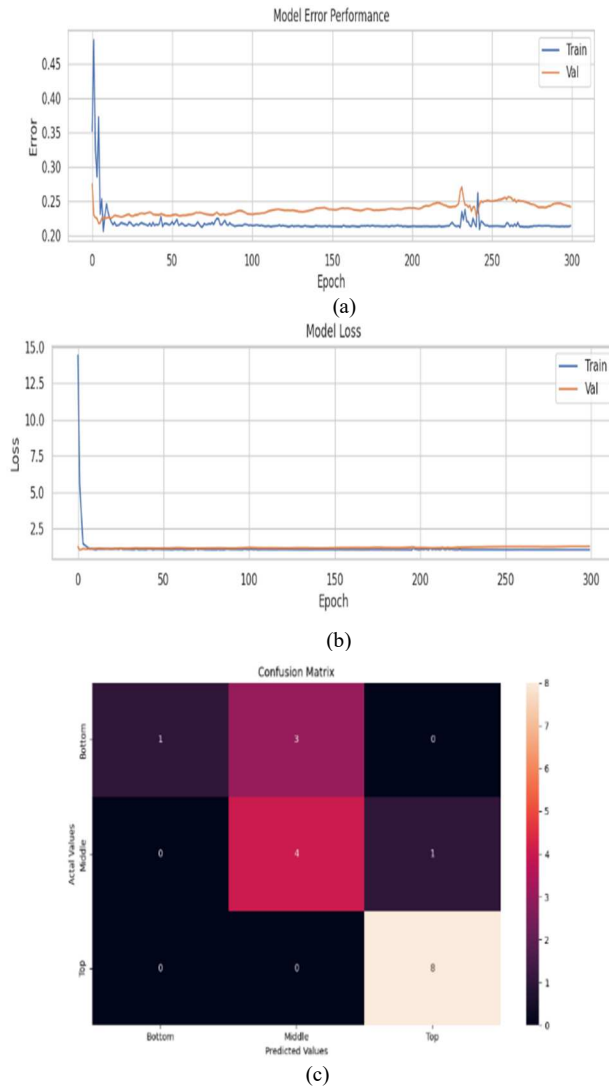


Figure 7: Training parameters

The training parameters of the localizer are shown in figure 7. In which MSE and modal loss for 100 sampled teeth of pulse-echo mode of dental ultrasound are used. the confusion matrix of the proposed methodology is show in figure 7 (c).

#### IV. CONCLUSION

The localization of broken instruments inside the RC is one of the most challenging tasks for the endodontist. With this paper, the authors conclude that broken instrument can be localized using ultrasound signals at a low frequency of 33 kHz. While the accuracy of the detection is about 77% with the initial prototype of pulse-echo mode. This will

increase with the larger number of databases and design modifications.

The work is still in progress for the exact and accurate localization of the broken instrument. This work will continue to detect the broken instruments inside the vivo as well, because in vitro, the hyperparameters are limited. Therefore, the localization of the broken instrument inside the vivo RC is a more challenging task than in vitro.

#### ACKNOWLEDGMENT

Authors pay gratitude to AICTE for funding this project. (Ref No. 8-122/FDC/RPS(Policy-1)/2019-20).

#### REFERENCES

- [1] M. Zehnder and G. N. Belibasakis, "On the dynamics of RC infections—what we understand and what we don't," *Virulence*, vol. 6, no. 3, Landes Bioscience, pp. 216–222, Feb. 05, 2015, doi: 10.4161/21505594.2014.984567.
- [2] K. Maeda and A. Kurjak, "Diagnostic ultrasound safety," *Donald School Journal of Ultrasound in Obstetrics and Gynecology*, vol. 8, no. 2, Jaypee Brothers Medical Publishers (P) Ltd, pp. 178–183, 2014, doi: 10.5005/jp-journals-10009-1353.
- [3] A. Carovac, F. Smajlovic, and D. Junuzovic, "Application of Ultrasound in Medicine," *Acta Informatica Medica*, vol. 19, no. 3, p. 168, 2011, doi: 10.5455/aim.2011.19.168-171.
- [4] S. Ghorayeb, C. Bertoincini, and M. Hinders, "Ultrasonography in dentistry," *IEEE Trans Ultrason Ferroelectr Freq Control*, vol. 55, no. 6, pp. 1256–1266, 2008, doi: 10.1109/TUFFC.2008.788.
- [5] G. Baum, I. Greenwood, S. Slawski, and R. Smirnow, "Observation of internal structures of teeth by ultrasonography," *Science* (1979), vol. 139, no. 3554, pp. 495–496, 1963, doi: 10.1126/science.139.3554.495.
- [6] B. Salmon and D. le Denmat, "Intraoral ultrasonography: Development of a specific high-frequency probe and clinical pilot study," *Clin Oral Investig*, vol. 16, no. 2, pp. 643–649, 2012, doi: 10.1007/s00784-011-0533-z.
- [7] K. T. Szopinski and P. Regulski, "Visibility of dental pulp spaces in dental ultrasound," *Dentomaxillofacial Radiology*, vol. 43, no. 1, 2014, doi: 10.1259/dmfr.20130289.
- [8] P. Kunche and N. Manikantababu, "Fractional Fourier Transform," *SpringerBriefs in Speech Technology*, vol. 58, no. 10, pp. 25–49, 2020, doi: 10.1007/978-3-030-42746-7\_2.
- [9] R. Kapoor, D. Sharma, and A. Kapoor, "Modified Sliding Singular Spectrum Analysis-Based Noncontact-Type Detection of Separated Instrument in RC Therapy Using Low-Frequency Ultrasonic Pattern Transceiver Design," *IEEE Trans Instrum Meas*, vol. 70, 2021, doi: 10.1109/TIM.2021.3094635.
- [10] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mech Syst Signal Process*, vol. 151, Apr. 2021, doi: 10.1016/j.ymssp.2020.107398.
- [11] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Sep. 2014, [Online]. Available: <http://arxiv.org/abs/1409.1556>
- [12] K. H. Cheah, H. Nisar, V. V. Yap, C. Y. Lee, and G. R. Sinha, "Optimizing residual networks and vgg for classification of eeg signals: Identifying ideal channels for emotion recognition," *J Healthc Eng*, vol. 2021, 2021, doi: 10.1155/2021/5599615.
- [13] Chih-Wei Hsu and Chih-Jen Lin, "A comparison of methods for multiclass support vector machines," in *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, March 2002, doi: 10.1109/72.991427.
- [14] A. R. A. Raziff, M. N. Sulaiman, N. Mustapha, and erumal, "Single classifier, OvO, OvA and RCC multiclass classification method in handheld based smartphone gait identification," in *AIP Conference Proceedings*, American Institute of Physics Inc., Oct. 2017, doi: 10.1063/1.5005342.



# Low-Frequency Waves in a Strongly Correlated Collisional Magnetized Dusty Plasma Cylinder

Harender Mor<sup>ID</sup>, Kavita Rani Segwal<sup>ID</sup>, and Suresh C. Sharma<sup>ID</sup>, *Senior Member, IEEE*

**Abstract**—Low-frequency electrostatic modes are studied using the generalized hydrodynamic model (GHD) for a drift-driven strongly correlated magnetized argon dusty plasma cylinder. The frequencies of both drifts-driven compressional and transverse modes tend to increase with normalized wave vector. The compressional wave modes are destabilized with normalized wavenumber; however, transverse mode gets stabilized. The longitudinal modes behave as drift-driven dust-ion acoustic modes, while the transverse mode observed in the presence of magnetic field-aligned currents behaves as dust-modified Alfvén ion waves. The effects of axial magnetic confinement and different geometries have been analyzed for both compressional and transverse modes.

**Index Terms**—Dusty plasma, longitudinal modes, plasma cylinder, strong correlation parameter, transverse shear modes.

## NOMENCLATURE

$n_{0e}$ , $n_{0d}$ , and $n_{0i}$	Unperturbed densities of electrons, dust, and ion, respectively.
$m_e$ , $m_d$ , and $m_i$	Mass of electron, dust, and ions, respectively.
$m_p$	Mass of proton.
$r_d$	Distance between the dust grains.
$v_e$ , $v_d$ , and $v_i$	Collisional frequencies of electron, dust, and ions, respectively.
$v_{ed}$ , $v_{dd}$ , and $v_{id}$	Drift velocities of electrons, dust, and ion, respectively.
$v_{et}$ , $v_{dt}$ , and $v_{it}$	Thermal velocities of electrons, dust, and ions, respectively.
$f_{ep}$ , $f_{dp}$ , and $f_{ip}$	Plasma frequencies of electrons, dust, and ions, respectively.
$f_{ec}$ , $f_{dc}$ , and $f_{ic}$	Cyclotron frequencies of electron, dust, and ions, respectively.
$\lambda_{ed}$ , $\lambda_{dd}$ , and $\lambda_{id}$	Debye lengths of electron, dust, and ions, respectively.
$V_{eA}$ and $V_{iA}$	Electron and ion Alfvén speed, respectively.

$f_r^t$ and $\gamma^t$	Real frequency and growth rate of transverse mode.
$f_r^c$ and $\gamma^c$	Real frequency and growth rate of compressional mode.
$\tau_m$	Relaxation time for the dust particle.
$\Gamma$	Coupling parameter for dust.
$\Gamma_c$	Critical value of the coulomb coupling parameter.
$Q_d = -Z_d e$	Charge on dust particle.
$Q_e = -e$ and $Q_i = e$	Charge on electron and ion, respectively.
$\mu_d$	Compressibility factor of the dust fluid.
$\gamma_d$	Adiabaticity factor.
$K_B$	Boltzmann's constant.
$\eta$	Coefficient of shear viscosity.
$\zeta$	Coefficient of bulk viscosity.
$a$	Radius of dust grains.
$\mathbf{B}$	Applied magnetic field.
$\mathbf{k}$	Wave vector.
$f$	Frequency of perturbed mode.
$\mathbf{u}$	Velocity term corresponding to the convective derivative.
$\delta \mathbf{u}_d$	Perturbed dust velocity.
$\delta \mathbf{E}$	Electrostatic perturbation.
$c$	Velocity of light.
$\phi$	Perturbed potential.

## I. INTRODUCTION

DUST is an omnipresent component in various laboratory and space plasma situations, such as fusion devices, plasma processing of materials, comet tails, planetary rings, solar wind, the earth's ionosphere, and so on [1], [2], [3], [4], [5]. These particles are undesirable during semiconductor wafer manufacture and fusion reactor design [6], [7], whereas dust agglomeration is an essential process in the evolution of protoplanetary disks and in the formation of planetesimals and comets [8].

The micrometer-sized dust particles tend to accumulate a high negative charge due to the high mobility of electrons. The variation of dust charge, mass, and size affects the dust particle dynamics and, hence, the collective behavior of plasma [9], [10], [11], [12], [13]. The vital characteristic of dusty plasma is its ability to have a strong correlation among dust grains. The dust grains are correlated via the screened Coulomb potential,

Manuscript received 15 January 2023; revised 1 June 2023 and 28 July 2023; accepted 24 August 2023. The review of this article was arranged by Senior Editor T. Hyde. (Corresponding author: Suresh C. Sharma.)

Harender Mor and Suresh C. Sharma are with the Department of Applied Physics, Delhi Technological University (DTU), Delhi 110042, India (e-mail: suresh321sharma@gmail.com).

Kavita Rani Segwal is with the Department of Applied Physics, Bhagwan Parshuram Institute of Technology, Delhi 110089, India.

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TPS.2023.3309851>.

Digital Object Identifier 10.1109/TPS.2023.3309851

which is a function of the Coulomb coupling parameter  $\Gamma$ . The Coulomb coupling parameter  $\Gamma$  quantifies the strength of correlation among the dust particles. It is given by the ratio of the electrostatic interaction energy to the dust thermal energy

$$\Gamma = \frac{Z_d^2 e^2}{r_d T_d} \exp(-r_d/\lambda_d)$$

where  $Z_d e$  denotes the dust charge,  $e$  denotes the electron charge,  $r_d$  is the intergrain spacing given by  $= ((4\pi n_{0d})/3)^{-1/3}$ , and  $T_d$  is the dust temperature, in the energy units  $T_d = K_B T$ ,  $\lambda_d$  denotes the Debye length, which accounts for the screening of the interaction. For  $\Gamma \ll 1$ , i.e., when the thermal kinetic energy dominates the electrostatic potential energy, the dust grains are said to be weakly correlated (WC), and when the Coulomb potential energy exceeds the mean kinetic energy, the dust grains become strongly correlated. In the limit  $1 \leq \Gamma \leq \Gamma_c$ , ( $\Gamma_c = 171 \pm 3$ ) given by the Monte Carlo simulations, plasma behaves like a fluid, and above this critical value, it exhibits Wigner crystallization [14], [15], [16], [17], [18]. This ability of strongly coupled complex plasma to co-exist in the fluid and crystalline state in the  $1 \leq \Gamma \leq \Gamma_c$  regime gives rise to the viscoelastic modes.

Since the strong correlation dust particles act as a viscous fluid, this introduces new dispersive corrections in the longitudinal mode along with the possibility of a novel shear mode. Kaw and Sen [19] utilized the generalized hydrodynamic model (GHD) to study the existence of transverse shear waves in strongly coupled dusty plasma [20]. These waves were observed experimentally in the 3-D dusty plasma in a strong coupling regime by Pramanik et al. [21].

Xie and Chen [22] studied the effect of a strong correlation between dust in the presence of a magnetic field and the absence of collisions and drifts of plasma particles. They found a compressional mode like an ion dust hybrid wave with a unique transverse mode similar to the torsional vibration mode. Banerjee et al. [23] studied the dispersion characteristics of the strongly correlated magnetized laboratory dusty plasma in the absence of dust-neutral collisions using the GHD model. It has been studied how strong coupling affects current-driven wave modes in collisional magnetized complex plasma for an infinite geometry [24].

As laboratory and fusion plasma devices exist in a confined geometry, the study of boundary effects on the dispersion characteristics of low-frequency modes is of marked importance. The boundary effect (or finite geometry effect) is significant when the mode wavelength is comparable to the size of the plasma. Additionally, finite Larmor radius effects are important in the case of finite geometry problems [25]. It has been observed that the drifting plasma particles and the quantized radial wave vector tend to modify the wave mode's dispersion properties for a cylindrical plasma geometry [26].

To observe the significance of boundary effects on strongly coupled plasma, we propose a theoretical model for analyzing the drift-driven electrostatic modes in a magnetized strongly correlated collisional dusty plasma cylinder. The effect of dust-dust collision has also been considered in the proposed model. In the proceeding section, the generalized hydrodynamic (GH) model has been applied in the long wavelength

region (where dust dynamics is vital) to derive the dispersion relation for the compressional and transverse mode in the hydrodynamic limit. In Section III, results and discussion are presented followed by the conclusion in Section IV.

## II. MODEL

We have assumed a strongly correlated collisional dusty plasma confined in the cylindrical geometry consisting of electrons, ions, and dust grains. The dust grain radius  $a$  is lesser compared to the intergrain distance  $r_d$  and the electron and ion Debye radii  $\lambda_e$  and  $\lambda_i$ . An axial magnetic field  $B_0$  is considered for the magnetic confinement of the plasma cylinder. The unperturbed densities of species (electrons, ions, and dust grains) are given by  $n_{0e}$ ,  $n_{0i}$ , and  $n_{0d}$ . The charge and mass of species are  $(Q_e, m_e)$ ,  $(Q_i, m_i)$ , and  $(Q_d, m_d)$ .  $v_{ed}$ ,  $v_{id}$ , and  $v_{dd}$  are the drift velocities of the species along the magnetic field. The neutral collisional frequencies of species are  $\nu_e$ ,  $\nu_d$ , and  $\nu_i$ . The electrostatic perturbation is given by  $\sim e^{-i(f t - k r)}$ . The GH momentum equation [19] for dust fluid in the limit  $1 \leq \Gamma \leq \Gamma_c$  is given as

$$\begin{aligned} \left(1 + \tau_m \left(\frac{\partial}{\partial t} + u \cdot \nabla\right)\right) & \left[ m_d n_{0d} \frac{\partial}{\partial t} \delta \mathbf{u}_d + \nabla \delta P \right. \\ & \left. + Z_d e n_{0d} \delta \mathbf{E} + m_d n_{0d} v_d \delta \mathbf{u}_d \right. \\ & \left. + \frac{1}{c} Z_d e n_{0d} \delta \mathbf{u}_d \times \mathbf{B}_0 \right] \\ & = \eta \nabla \cdot \nabla \delta \mathbf{u}_d + \left(\zeta + \frac{\eta}{3}\right) \nabla (\nabla \cdot \delta \mathbf{u}_d) \end{aligned} \quad (1)$$

where  $\delta P$  is pressure,  $c$  is the velocity of light,  $u_d$  is the dust velocity,  $\eta$  and  $\zeta$  are the viscoelastic coefficients, and  $\tau_m$  is the relaxation time for the dust particle.

The compressibility of dust ( $\mu_d$ ) [27] is given as

$$\mu_d = \frac{(\partial_n P_d)_T}{T_d} = 1 + \frac{1}{3} u(\Gamma) + \frac{1}{9} \frac{\partial u(\Gamma)}{\partial \Gamma}. \quad (2)$$

The quantity  $u(\Gamma)$  gives the excess internal energy, which is represented in the strongly correlated regime through the analytical expression given by Slattery et al. [28]

$$u(\Gamma) = -0.89\Gamma + 0.95\Gamma^{1/4} + 0.19\Gamma^{-1/4} - 0.82 \quad \text{for } 1 \leq \Gamma \leq 200. \quad (3)$$

The dependence of viscoelastic coefficients  $\eta$  and  $\zeta$  on  $\Gamma$  is given by the Monte Carlo simulations [28], [29]

$$\eta' = \frac{(\zeta + \frac{4}{3}\eta)}{m_d n_{0d} f_{dp} r_d^2} \approx 0.02\Gamma^{1/2} \quad (4)$$

and

$$\tau_m^* = f_{dp} \tau_m \approx 0.375\Gamma^{1/2} \quad (5)$$

where

$$\tau_m = \left(\zeta + \frac{4}{3}\eta\right) / n_{0d} T_d (1 - \gamma_d \mu_d + 4u(\Gamma)/15) \quad (6)$$

where  $\gamma_d$  is the adiabaticity and  $f_{dp}$  is the dust plasma frequency. The following subsection aims to obtain the dispersion relation for the low-frequency modes by applying the set of GH equations.

### A. Dust Compressional Mode

The compressional low-frequency modes are perpendicular to the undisturbed magnetic field and parallel to the electrostatic perturbation and the wave vector. These modes are supported by the inertia of massive dust particles. The continuity equation for the dust fluid is given by

$$\frac{\partial}{\partial t} \delta n_d + \nabla \cdot (n_d \delta \mathbf{u}_d) = 0. \quad (7)$$

As these waves are accompanied by density fluctuations, the perturbed density of dust particles is found by implementing (2) and (7) into the dust equation of motion given by (1).

The perturbed density of dust particles is readily obtained as

$$\delta n_d = \frac{-Z_d e n_{0d} \nabla_{\perp}^2 \phi}{m_d (f - k_z v_{dd}) A} \quad (8)$$

where

$$A = \left[ \frac{k^2 \gamma_d \mu_d v_{dt}^2}{(f - k_z v_{dd})} - (f - k_z v_{dd} + i v_d) + \frac{k^2 (\zeta + \frac{4}{3} \eta)}{(1 - i \tau_m f) i m_d n_{0d}} \right] + \frac{f_{dc}^2}{\left( \frac{i \eta k^2}{(1 - i \tau_m f) m_d n_{0d}} + (f - k_z v_{dd} + i v_d) \right)} \quad (9)$$

where  $v_{dt} = ((T_d/m_d)^{1/2})$  and  $f_{dc} = [-(Z_d e) B / m_d c]$  are the dust thermal velocity and the dust cyclotron frequency, respectively. Furthermore, the viscoelastic coefficients  $\eta$ ,  $\zeta$ , and  $\tau_m$  have been omitted, taking into consideration the fact that only the dust particles are in a strongly coupled state, while the electrons and ions are in the weakly coupled state. On substituting  $\gamma_d = \mu_d = 1$  for the low compressibility limit, and  $f \ll k_z v_{ed}$ ,  $v_e \ll k v_{et}$  and  $f^2 \ll f_{ic}^2 \ll f_{ec}^2$ ,  $k_z v_{dd} \ll f \ll k_z v_{id}$ , the subsequent equations have been obtained. The perturbed number densities of the ion and electrons may be obtained as

$$\delta n_r = \frac{\mp e n_{0k} \nabla_{\perp}^2 \phi}{m_j [k^2 v_{rt}^2 + f_{rc}^2 - (f - k_z v_{rd} + i v_r)(f - k_z v_{rd})]} \quad (10)$$

where  $r = i$  or  $e$ , and  $v_{rt} = ((T_r/m_r)^{1/2})$  are the thermal velocities of species and  $f_{rc} = \pm (e B_0 / m_r c)$  are the cyclotron frequencies of species. Substituting the values of perturbed densities of species from (8) and (10) in Poisson's equation

$$\nabla^2 \phi = 4\pi (\delta n_e e + z_d e \delta n_d - e \delta n_i). \quad (11)$$

For axially and azimuthally symmetric case, (11) takes the form which is the Bessel equation of order zero

$$\frac{\partial^2 \phi}{\partial r^2} + \frac{1}{r} \frac{\partial \phi}{\partial r} + q^2 \phi = \frac{f_{dp}^2 k_{\perp}^2 \phi}{(f - k_z v_{dd}) A} \quad (12)$$

where

$$q^2 = -k_z^2 - L k_{\perp}^2 \phi \quad (13)$$

and

$$L = \left[ \frac{f_{ep}^2}{[k_{\perp}^2 v_{et}^2 + f_{ec}^2 - (f - k_z v_{ed} + i v_e)(f - k_z v_{ed})]} + \frac{f_{ip}^2}{[k_{\perp}^2 v_{it}^2 + f_{ic}^2 - (f - k_z v_{id} + i v_i)(f - k_z v_{id})]} \right] \quad (14)$$

Here,

$$f_{rp} = \left( \frac{4\pi (Q_r)^2 n_{0r}}{m_r} \right)^{1/2} \quad (15)$$

is the characteristic plasma frequency of oscillation, when they are disturbed from the equilibrium position, where  $(Q_r, n_{0r}, m_r)$  are the charge density, number density, and mass of the plasma species, respectively. Here,  $r = d, i$ , and  $e$  corresponds to the dust particles, ions, and electrons, respectively. In the absence of RHS of (12),  $\phi$  is given by

$$\phi = U J_0(q_x r) \quad (16)$$

where  $U$  is a constant and  $J_0(q_x r)$  is the Bessel function of order zero and argument  $q_x r$ . On the surface of the plasma waveguide, we should have  $J_0(q_x r) = 0$  [30]. Here,  $q_x r = \delta_n$ , where  $\delta_n$  gives the zeroes of the Bessel function. Some of which are given as  $\delta_1 = 2.404$ ,  $\delta_2 = 5.5$ ,  $\delta_3 = 8.7$ , and  $\delta_4 = 11.8$  [31]. When the effect of dust dynamics is considered [i.e., RHS of the (12)  $\neq 0$ ], wave function  $\phi$  can be given by a combination of orthogonal wave functions

$$\phi = \sum_m A_m J_0(q_y r) \quad (17)$$

where the effective radial wave vector  $q_x = 2.404/r_0$  is quantized. Simplifying Bessel's equation for the dominant mode  $x = y$ , in the limit  $f \ll k_z v_{ed}$ ,  $v_e \ll k_{\perp}^2 v_{et}^2 < f_{ec}^2$ , the dispersion relation for the compressional mode is obtained as

$$1 + \frac{1}{\Omega_e^2} + \frac{f_{ip}^2}{[\Omega_i^2 f_{ip}^2 - i v_i (f - k_z v_{id})]} - \frac{f_{dp}^2 I}{(f - k_z v_{dd}) A} = 0 \quad (18)$$

where

$$I = \frac{\int_0^{r_0} r J_0(q_x r) J_0(q_y r) dr}{\int_0^r r J_0(q_x r) J_0(q_y r) dr} \\ \Omega_e^2 = k_{\perp}^2 \lambda_{eD}^2 + V_{eA}^2, \quad \text{and} \quad \Omega_i^2 = (k_{\perp}^2 \lambda_{iD}^2 + V_{iA}^2 - k_z^2 v_{id}^2 / f_{ip}^2) \quad (19)$$

where  $\lambda_{id} = (v_{it} / f_{ip})$ ,  $\lambda_{ed} = (v_{et} / f_{ep})$ ,  $V_{eA} = (f_{ec} / f_{ep})$ , and  $V_{iA} = (f_{ic} / f_{ip})$ .

When the dust drift and magnetic field are neglected, then (18) will yield the same dispersion relation for compressional mode as obtained by Kaw and Sen [19, p. 3556, (28)]. In the hydrodynamic limit ( $f \tau_m \ll 1$ ), (18) gives the dispersion relation for the compressional dust modes. Taking  $f = f_r + i\gamma$  and  $f_r \gg \gamma$  yields the real frequency as

$$f_r^c = k_z v_{dd} + f_{dp} \left[ \frac{P}{Q} I + q_m^2 \lambda_d^2 \gamma_d \mu_d + \left( -\frac{\eta_d^* q_m^2 r_d^2}{4} + \frac{v_d}{4 f_{dp}} \right)^2 + \frac{f_{dc}^2}{f_{dp}^2} \right]^{1/2} \quad (20)$$

where

$$P = f_{ip}^4 \Omega_i^2 (1 + \Omega_i^2) + v_i (f - k_z v_{id})^2$$

and

$$Q = [(1 + \Omega_i^2) f_{ip}^2]^2 + v_i^2 (f - k_z v_{id})^2.$$

The growth rate of the dust compressional mode is given as

$$\frac{\gamma^c}{f_{dp}} = \left[ -\frac{v_d}{2f_{dp}} + \frac{n_d^* q_m^2 r_d^2}{2} \right]. \quad (21)$$

Equations (20) and (21) represent the frequency and growth rate of the compressional mode. In the absence of collision, (21) is similar to Xie and Chen [22, p. 3521, (18)].

### B. Transverse Mode

The presence of the transverse mode is supported by the rigidity provided by the strong correlations among dust particles. Taking the curl on both sides of (1) to get the transverse component, we get

$$(f + i v_d + i \eta_1) \mathbf{k} \times \mathbf{u}_d + i \frac{Z_d e}{m_d} (\mathbf{k} \times \delta \mathbf{E}) + i \frac{Z_d e}{m_d c} (\mathbf{k} \cdot \mathbf{B}) \delta \mathbf{u}_d = 0 \quad (22)$$

where  $\eta_1 = (\eta k^2 / (m_d n_{0d} f_{dp} r_d^2 (1 - i \tau_m f)))$ .

Equation (22) could be used to derive the perturbed velocity of dust fluid transverse to  $\mathbf{B}$

$$\delta \mathbf{u}_{d\perp} = \frac{Z_d e / m_d}{i [f_{dc}^2 - (f + i v_d + i \eta_1)^2]} \times \left[ (f + i v_d + i \eta_1) \delta \mathbf{E} + f_{dc} \frac{(\nabla \times \delta \mathbf{E}_z)}{k_z} \right]. \quad (23)$$

Using (23) in the dust fluid continuity equation, we get the perturbed density of plasma species as

$$\delta n_r = \frac{(f + i \eta_1 + i v_r) Q_r n_{0r} / m_r}{(f - k_z v_{rd}) [f_{rc}^2 - (f + i \eta_1 + i v_r)^2]} \nabla_{\perp}^2 \phi \quad (24)$$

where  $r = d, e, \text{ or } i$  for dust, electron, and ions, respectively.

The term  $\eta_1$ , in numerator and denominator can be ignored for obtaining the perturbed density of electrons and ions.

Using the perturbed densities in Poisson's equation and applying the axial and azimuthal symmetry for simplicity, we get Bessel's equation

$$\frac{\partial^2 \phi}{\partial r^2} + \frac{1}{r} \frac{\partial \phi}{\partial r} + N^2 \phi = -\frac{f_{dp}^2 k_{\perp}^2 \phi (f + i \eta_1' + i v_d)}{(f - k_z v_{dd}) [(f + i \eta_1' + i v_d)^2 - f_{dc}^2]} \quad (25)$$

where

$$N^2 = -k_z^2 + \frac{k_{\perp}^2 f_{ip}^2 (f + i v_i)}{(f - k_z v_{id}) [(f + i v_i)^2 - f_{ic}^2]} + \frac{k_{\perp}^2 f_{ep}^2 (f + i v_e)}{(f - k_z v_{ed}) [(f + i v_e)^2 - f_{ec}^2]}. \quad (26)$$

Now, simplifying (25) as was done earlier, the solution of Bessel's equation to obtain the dispersion relation for the transverse shear mode, we get

$$1 - \frac{k_{\perp}^2}{k^2} \frac{f_{ip}^2 (f + i v_i)}{(f - k_z v_{id}) [(f + i v_i)^2 - f_{ic}^2]} - \frac{k_{\perp}^2}{k^2} \frac{f_{ep}^2 (f + i v_e)}{(f - k_z v_{ed}) [(f + i v_e)^2 - f_{ec}^2]} = f_{dp}^2 \frac{k_{\perp}^2 I(f + i \eta_1 + v_d)}{k^2 (f - k_z v_{dd}) [(f + i \eta_1 + i v_d)^2 - f_{dc}^2]}. \quad (27)$$

Neglecting the effect of drift, collision, bounded geometry, and magnetic field, the dispersion relation is similar to Kaw and Sen [19, p. 3556, (22)]. If magnetized plasma is considered, then it matches well with [22, p. 3522, (33)]. Simplifying (27) in the hydrodynamic limit ( $f \tau_m \ll 1$ ) under the assumptions ( $k_{\perp}^2 / k^2 \ll 1$ ,  $f_{ec}^2 + v_e^2 \gg 1$ , and  $v_{ed} \gg 1$   $f \ll f_{ic}$ ,  $v_i \ll f_{ec}$ ,  $v_e$  and  $f \ll \eta_1 \ll f_{dc}$ ).

We get

$$(f - k_z v_{dd}) [f + i(v_d + \eta_1')] = f_{dp}^2 M \quad (28)$$

where  $M = (k_{\perp}^2 I / (k^2 [1 - i((k_{\perp}^2 f_{ip}^2 v_i) / (k^2 k_z v_{id} (f_{ic}^2 + v_i^2))]))$ .

Equation (28) can be rewritten as  $\varepsilon_r(f, k) + i \varepsilon_i(f, k) = 0$  for obtaining the frequency and the growth rate of the shear mode

$$\frac{f_r}{f_{dp}} = \frac{k_z v_{dd}}{2 f_{dp}} \pm \sqrt{\frac{k_z v_{dd}}{2 f_{dp}} + \eta^* \tau_m^* k^2 r_d^2 + M_r} \quad (29)$$

and

$$\gamma^t = -\frac{v_d}{2} + \frac{v_d k_z v_{dd}}{2 f_r} - \frac{\eta k^2}{2 m_d n_{0d}} + f_{dp}^2 M_i \quad (30)$$

where  $M_r$  and  $M_i$  are the real and imaginary parts of  $M$ .

Equations (29) and (30) represent the frequency and growth rate of the transverse mode. In the absence of dust drift and collision of ions and neutrals, the transverse mode is analogous to the dust-ion Alfvén waves given by Xie and Chen [22].

## III. RESULTS AND DISCUSSION

The effect of a finite geometry in the drift-driven magnetized strongly correlated Argon dusty plasma on the low-frequency modes has been analyzed in the hydrodynamic regime ( $f \tau_m \ll 1$ ). The slightly modified parameters have been used for the strongly correlated magnetized dusty plasma system [32]. The density of various species is given as  $n_{0i} \approx 10^8 \text{ cm}^{-3}$ ,  $n_{d0}/n_{0i} = 5 \times 10^{-4}$ , and  $n_{0e} = 0.5 \times n_{0i}$ . The dust charge state is given by  $Z_d \approx 10^3$ , the electron temperature as  $T_e \approx 1.0 \text{ eV}$ , and  $T_i \sim T_d = 0.14 \text{ eV}$  are the ion and dust temperature, respectively.  $a = 5 \times 10^{-4} \text{ cm}$  is the size (radius) of the dust grains, and their mass is  $10^{12}$  times the proton mass  $m_p$ . An axial magnetic field  $B_0 = 4000 \text{ G}$  is taken, and  $v_i (= 0.5 f_{ic})$  and  $v_d (= 0.2 f_{dp})$  are the collisional frequencies of ions and dust, respectively. The drift velocities of the ions and dust particles are  $v_{id} = 0.5 \times v_{it}$  and  $v_{dd} = 1.0 v_{dt}$ , respectively. The effective quantized radial wave vector is  $k_{\perp} \approx q_n = 2.404/r_0 = 0.8 \text{ cm}$ . The parameters taken in this study are in the range of strongly correlated dusty



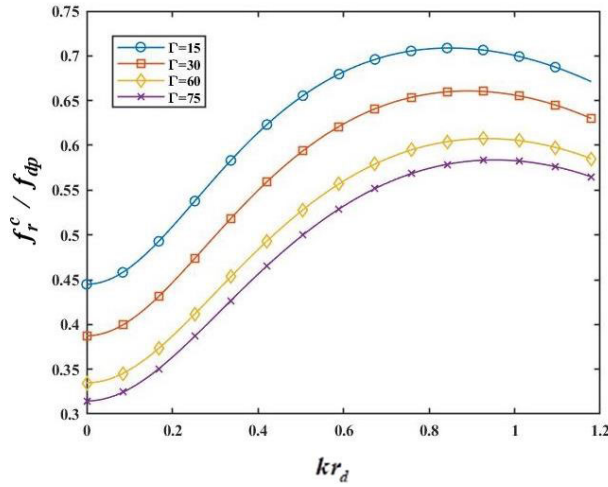


Fig. 1. Variation of normalized real frequency ( $f_r^c/f_{dp}$ ) of compressional mode with normalized wave vector ( $kr_d$ ) for various coupling parameters  $\Gamma$  and for  $v_d = 0.6f_{dp}$ ,  $v_i = 0.5f_{ic}$ ,  $v_{dd} = v_{dt}$ ,  $v_{id} = 0.5v_{it}$ ,  $k_z = 2.0 \text{ cm}^{-1}$ , and  $B_0 = 4000 \text{ G}$ .

one-component plasma limit  $\lambda_d > r_d \gg \lambda_{dd}$ . Here,  $r_d$  is the intergrain spacing and  $\lambda_{dd}$  is the dust Debye length.

For the dust to be magnetized, the small size of the dust particles should be complemented with the use of a high magnetic field. Moreover, the dust Larmor radius  $\rho_d (=v_{dt}/f_{dc})$  should be much smaller than the plasma tube diameter  $2r_0$  and  $f_{dc} \gg v_d$ . This is because magnetizing the complex plasma is technically challenging [32], [33], [34] due to the low charge-to-mass ratio of dust particles. The cyclotron frequency ratio to the neutral momentum exchange frequency ( $f_{rc}/v_{rn} \geq 1$ ) should be greater than one, such that the particle performs full rotation (on average) of the Larmor orbit before collision [35]. To avoid Coulomb crystallization of dust grains, high dust temperature has been considered to retain the fluid characteristic of plasma. We have employed MATLAB (version 2020a Natick, Massachusetts: The MathWorks Inc.) [36] for the solution of the equations and random values of the Coulomb coupling parameter have been considered in the  $1 \leq \Gamma \leq \Gamma_C$  limit.

The variation of normalized real frequency ( $f_r^c/f_{dp}$ ) of compressional mode with normalized wave vector ( $kr_d$ ) for various Coulomb coupling parameter  $\Gamma$  values has been shown in Fig. 1. It has been observed that in the long wavelength region ( $kr_d \leq 1$ ), the mode frequency rises with an increase in wave vector, but anomalous dispersion ( $\partial f/\partial k < 0$ ) is observed in the short wavelength region, which has been confirmed experimentally [37]. It is in fair agreement with the previous studies of Xie and Chen [22], which show a decrement in mode frequency for the strong coupling parameter  $\Gamma$ , and this may be because the viscosity of dust fluid is enhanced by strong correlation, which leads to dampening of the wave.

Fig. 2 depicts the variation of normalized mode frequency with normalized wave vector for different geometries and different values of the strong coupling parameter  $\Gamma = 30, 60$ , and  $125$ , which are below the  $\Gamma_c$ . It is clear from the figure that mode frequency for finite geometry is less than that of infinite geometry, for a particular value of coupling

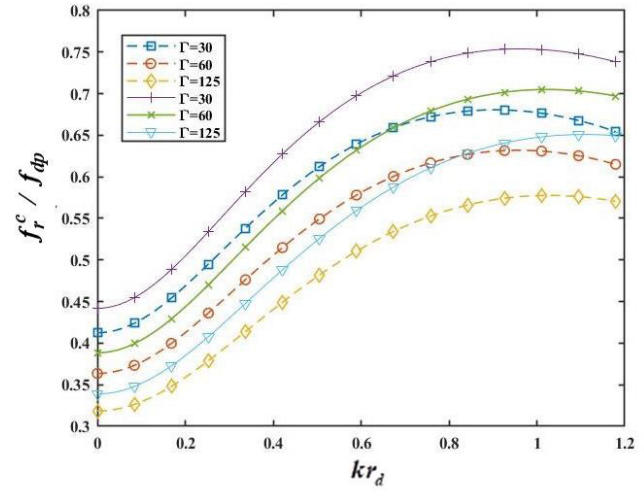


Fig. 2. Variation of normalized real frequency ( $f_r^c/f_{dp}$ ) of compressional modes with normalized wave vector ( $kr_d$ ). (i) Solid lines for infinite geometry with  $\Gamma = 30, 60$ , and  $125$ . (ii) Dashed lines for finite geometry with  $\Gamma = 30, 60, 125$ , and for  $v_d = 0.6f_{dp}$ ,  $v_i = 0.5f_{ic}$ ,  $v_{dd} = v_{dt}$ ,  $v_{id} = 0.5v_{it}$ ,  $k_z = 2.0 \text{ cm}^{-1}$ , and  $B_0 = 4000 \text{ G}$ .

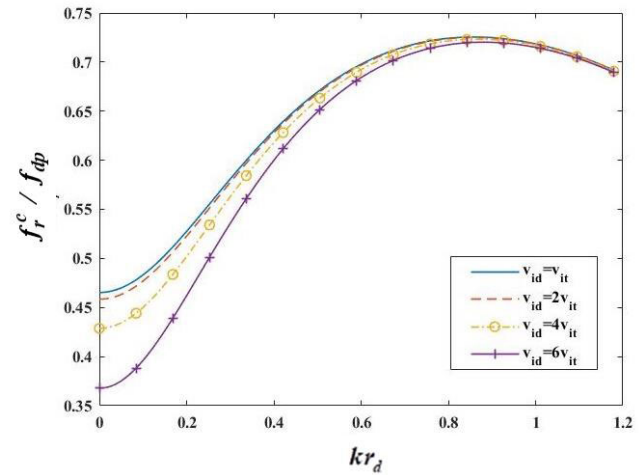


Fig. 3. Variation of normalized real frequency ( $f_r^c/f_{dp}$ ) of compressional modes with normalized wave vector ( $kr_d$ ) for different ion drift speeds (i)  $v_{id} = v_{it}$ , (ii)  $v_{id} = 2v_{it}$ , (iii)  $v_{id} = 4v_{it}$ , and (iv)  $v_{id} = 6v_{it}$  and for  $v_d = 0.6f_{dp}$ ,  $v_i = 0.5f_{ic}$ ,  $v_{dd} = v_{dt}$ ,  $v_{id} = 0.5v_{it}$ ,  $k_z = 2.0 \text{ cm}^{-1}$ , and  $B_0 = 4000 \text{ G}$ .

parameter. The reduction in mode frequency is more for the short wavelength region. The wave frequency is reduced nearly by 10% from that of infinite geometry in the long wavelength region ( $kr_d < 1$ ). This may be due to the reduction in the interaction region for a finite geometry.

Fig. 3 depicts the variation of normalized mode frequency of compressional mode for different ion drift velocities with the normalized wave vector. It has been observed that the mode frequency decreases with the increase in the ion drift velocity in the long wavelength region ( $kr_d \leq 1$ ). This is in line with the observation of Kaw and Singh [38], and for the short wavelength region, it is independent of the drift of the ions.

This behavior agrees well with the experimental observation of Molotov et al. [39], where the increase in the drift motion

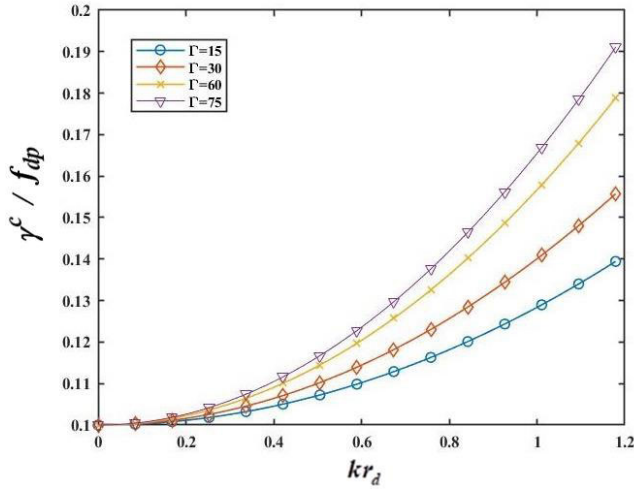


Fig. 4. Normalized growth rate ( $\gamma^c/f_{dp}$ ) of compressional mode with normalized wave vector ( $kr_d$ ) for various Coulomb coupling parameter (i)  $\Gamma = 15$ , (ii)  $\Gamma = 30$ , (iii)  $\Gamma = 60$ , and (iv)  $\Gamma = 75$  and for  $v_d = 0.6f_{dp}$ ,  $v_i = 0.5f_{ic}$ ,  $v_{dd} = v_{dt}$ ,  $v_{id} = 0.5v_{it}$ ,  $k_z = 2.0 \text{ cm}^{-1}$ , and  $B_0 = 4000 \text{ G}$ .

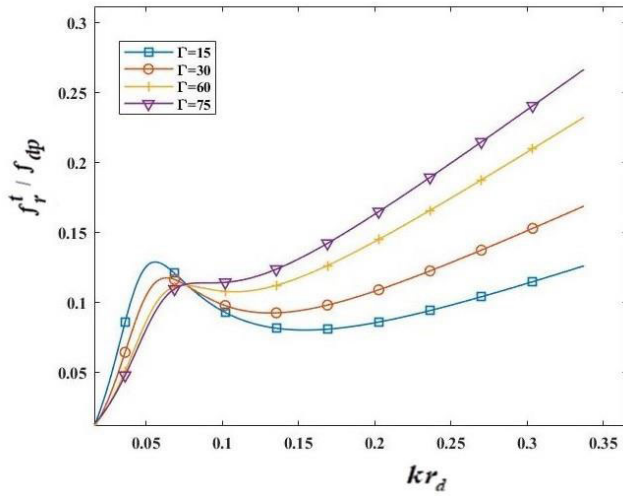


Fig. 5. Dependence of the normalized mode frequency ( $f_r^t/f_{dp}$ ) on the normalized wave vector ( $kr_d$ ) of transverse mode for different values of Coulomb coupling parameter  $\Gamma$  (i)  $\Gamma = 15$ , (ii)  $\Gamma = 30$ , (iii)  $\Gamma = 60$ , and (iv)  $\Gamma = 75$  and for  $v_d = 0.6f_{dp}$ ,  $v_i = 0.5f_{ic}$ ,  $v_{dd} = 0.5v_{dt}$ ,  $v_{id} = 0.5v_{it}$ ,  $k_\perp = 0.8 \text{ cm}^{-1}$ ,  $k_z = 0 - 25 \text{ cm}^{-1}$ , and  $B_0 = 4000 \text{ G}$ .

of ions as compared to their thermal speed suppresses the dust acoustic instability.

Fig. 4 depicts the variation of the normalized growth rate ( $\gamma^c/f_{dp}$ ) of compressional modes versus normalized wave vector for different values of Coulomb coupling parameter  $\Gamma$ . The growth rate rises with the increasing wave vector but diminishes with the increase in strong coupling parameter  $\Gamma$ . This may be associated with the enhanced viscosity of system, which on increase in strong correlation among dust provides inertia to the massive dust particles.

To analyze the transverse mode, we have shown the dependence of the normalized mode frequency ( $f_r^t/f_{dp}$ ) on the normalized wave vector ( $kr_d$ ) for different values of strong coupling parameter  $\Gamma$  in Fig. 5. The mode frequency

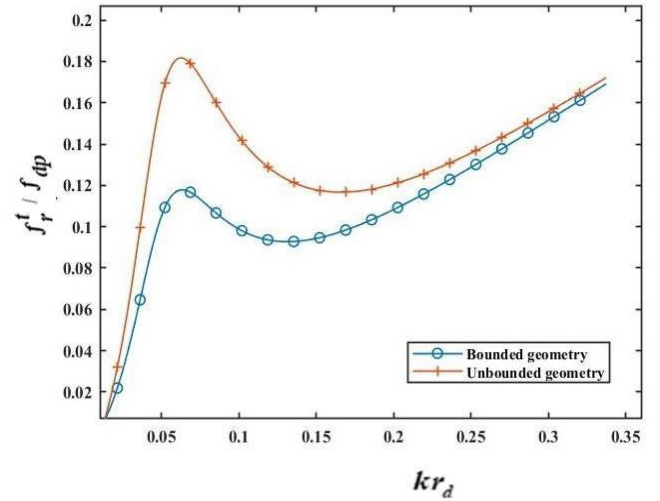


Fig. 6. Dependence of the normalized mode frequency ( $f_r^t/f_{dp}$ ) on the normalized wave vector ( $kr_d$ ) of transverse mode for Coulomb coupling parameter  $\Gamma = 30$  (i) unbounded geometry, (ii) bounded geometry, and  $v_d = 0.6f_{dp}$ ,  $v_i = 0.5f_{ic}$ ,  $v_{dd} = 0.5v_{dt}$ ,  $v_{id} = 0.5v_{it}$ ,  $k_\perp = 0.8 \text{ cm}^{-1}$ ,  $k_z = 0 - 25 \text{ cm}^{-1}$ , and  $B_0 = 4000 \text{ G}$ .

increases with increasing wave vector with the increase in strong coupling parameter showing a slight decrement between  $k_z(0.05-0.1)$ . This behavior is similar to dust-modified Alfvén modes [40] and well in agreement with the results of streaming and collisional instabilities in the absence of magnetic field and strong coupling [41, p. 4413], Figs. 3 and 4]. This is in accordance with the fact that strong coupling supports the transverse shear mode by providing solid like rigidity and is well in agreement with the observations of Khrapak et al. [42].

It has been observed that the cylindrical geometry leads to an effective quantized wavenumber. The effect of different geometries on the mode frequency for the transverse mode in strong coupling regime has been studied (see Fig. 6). It is seen that the mode frequency is nearly halved in the case of finite geometry at  $kr_d \simeq 0.05$  in the long wavelength region. This difference in frequency is significant in the long wavelength region up to  $k < 0.1$ . The corresponding increment in mode frequency for bounded plasmas is less as compared to the infinite geometry in the long wavelength region.

The normalized growth rate of transverse modes with normalized wavenumber for different values of Coulomb coupling parameter  $\Gamma$  has been plotted in Fig. 7. A negative growth rate has been observed, which shows modes are purely damped. This result is well in agreement with the previous results [19]. However, the small positive growth rate may be observed for a weakly collisional regime for small values of strong coupling parameter  $\Gamma$ , and this behavior is similar to Alfvén waves [40]. The growth rate shows a decrease with the increase in the Coulomb coupling parameter  $\Gamma$  as a result of viscous damping.

The comparison of normalized growth rate of transverse mode for finite and infinite cases, as a function of normalized wave vector, is depicted in Fig. 8. The growth rate for the finite geometry is lesser than that of infinite geometry for a minuscule region in long wavelength regime ( $k < 0.1$ ), while it remains unaffected in short wavelength region. However, the mode is damped for both finite and infinite geometries.



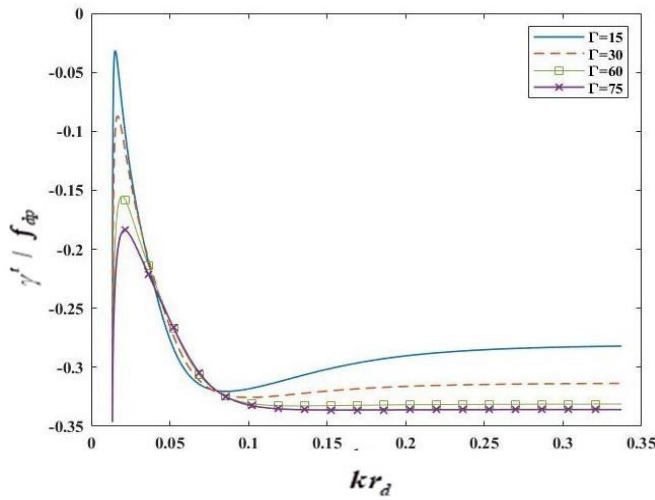


Fig. 7. Variation of normalized growth rate ( $\gamma^t/f_{dp}$ ) of transverse mode with normalized wave vector ( $kr_d$ ) for various values of Coulomb coupling parameter (i)  $\Gamma = 15$ , (ii)  $\Gamma = 30$ , (iii)  $\Gamma = 60$ , and (iv)  $\Gamma = 75$  and  $v_d = 0.6f_{dp}$ ,  $v_i = 0.5f_{ic}$ ,  $v_{dd} = 0.5v_{dt}$ ,  $v_{di} = 0.5v_{it}$ ,  $k_{\perp} = 0.8 \text{ cm}^{-1}$ ,  $k_z = 0-20 \text{ cm}^{-1}$ , and  $B_0 = 4000 \text{ G}$ .

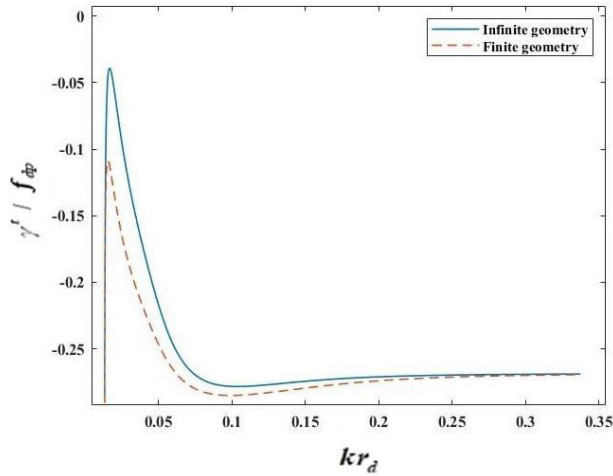


Fig. 8. Variation of normalized growth rate ( $\gamma^t/f_{dp}$ ) of transverse mode for different plasma geometries with normalized wave vector ( $kr_d$ ) for  $\Gamma = 30$  (i) for the infinite geometry, (ii) for finite geometry, and  $v_d = 0.6f_{dp}$ ,  $v_i = 0.5f_{ic}$ ,  $v_{dd} = 0.5v_{dt}$ ,  $v_{di} = 0.5v_{it}$ ,  $k_{\perp} = 0.8 \text{ cm}^{-1}$ ,  $k_z = 2 \text{ cm}^{-1}$ , and  $B_0 = 4000 \text{ G}$ .

We have subsequently analyzed the effect of the magnetic field on the frequency (see Fig. 9) and growth rate (see Fig. 10) of transverse mode for different values of radial distance  $r$  and observed that in the long wavelength region, and the mode frequency increases with increasing magnetic field. At a particular magnetic field strength, mode frequency is more for the larger values of radial distance from the axis of plasma cylinder. The magnetic field provides tensile strength to the plasma fluid, making it more crystalline and hence giving solid-like elasticity for transverse mode propagation. The finite Larmor radius effect is significant when the wavelength of the wave is less than or comparable to the gyroradius (i.e., the ratio of electron gyroradius to perpendicular wavelength).

The growth rate increases with the magnetic field and stabilizes for the higher value of the magnetic field, and this may be due to the fact that the low-frequency modes are difficult sustain at a very high magnetic field.

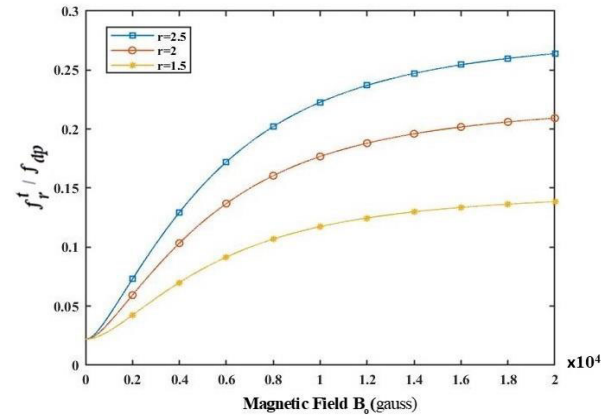


Fig. 9. Normalized frequency of transverse mode versus magnetic field for different values of  $r$ , and the radial distance from the axis of plasma cylinder for Coulomb coupling parameter  $\Gamma = 30$  (i)  $r = 2.5$ , (ii)  $r = 2$ , and (iii)  $r = 1.5$  and for  $k_z = 2 \text{ cm}^{-1}$ ,  $v_d = 0.6f_{dp}$ ,  $v_i = 0.5f_{ic}$ ,  $v_{dd} = 0.5v_{dt}$ ,  $v_{di} = 0.5v_{it}$ , and  $k_{\perp} = 0.8 \text{ cm}^{-1}$ .

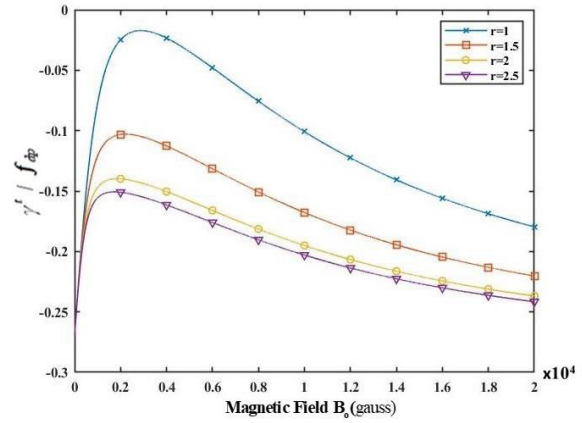


Fig. 10. Normalized growth rate of transverse mode versus magnetic field for different values of  $r$ , and the radial distance from the axis of plasma cylinder for Coulomb coupling parameter  $\Gamma = 30$  (i)  $r = 1$ , (ii)  $r = 1.5$ , (iii)  $r = 2$ , and (iv)  $r = 2.5$  and for  $k_z = 2 \text{ cm}^{-1}$ ,  $v_d = 0.6f_{dp}$ ,  $v_i = 0.5f_{ic}$ ,  $v_{dd} = 0.5v_{dt}$ ,  $v_{di} = 0.5v_{it}$ , and  $k_{\perp} = 0.8 \text{ cm}^{-1}$ .

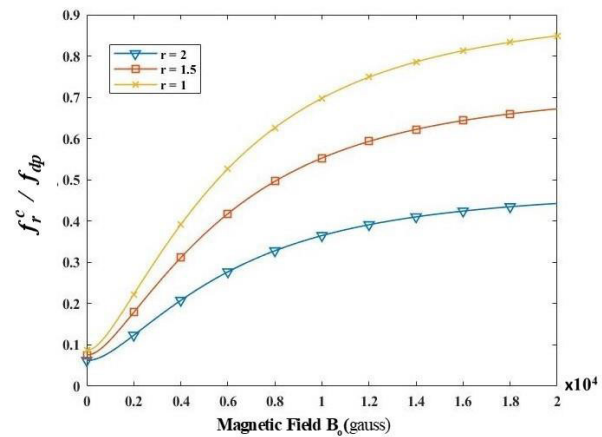


Fig. 11. Normalized frequency of compressional mode versus magnetic field for different values of  $r$ , and the radial distance from the axis of plasma cylinder for Coulomb coupling parameter  $\Gamma = 30$  (i)  $r = 1$ , (ii)  $r = 1.5$ , and (iii)  $r = 2$  and for  $k_z = 2 \text{ cm}^{-1}$ ,  $v_d = 0.6f_{dp}$ ,  $v_i = 0.5f_{ic}$ ,  $v_{dd} = 0.5v_{dt}$ ,  $v_{di} = 0.5v_{it}$ , and  $k_{\perp} = 0.8 \text{ cm}^{-1}$ .

It is also observed that the frequency of compressional mode also grows with increasing magnetic field strength. Moreover, as the radial distance from the axis of the plasma cylinder

increases, mode frequency shows decrement (see Fig. 11). If the radius of confinement is doubled, the frequency nearly is reduced to half as magnetic field strength  $B_0$  increases from 1 to 4 T. It shows an increase in mode frequency with the increase in axial confinement of plasma, which means that to observe the effect of magnetized dust on low-frequency modes in the strongly coupled plasma, the plasma confinement radius should be small enough. At higher values of the magnetic field, frequency saturates, which is in accordance with the experimental observation of the dust acoustic waves in magnetized plasma [34], [43]. Moreover, the growth rate of compressional mode is found to be unaffected with the change in magnetic field in a strongly correlated collisional dusty plasma.

#### IV. CONCLUSION

In this article, we have analyzed the excitation of low-frequency electrostatic modes in the hydrodynamic regime ( $f\tau_m \ll 1$ ) by virtue of the drift produced for ions and dust particles for a plasma cylinder. We have obtained a set of equations for the growth rate and frequency of modes, which applies to the cylindrical geometry. It can be reiterated that the plasma geometry affects the dispersion characteristics of both the wave modes. Anomalous dispersion is observed for the compressional modes in the higher wavenumber limit. Transverse modes are purely damped modes, whose behavior is similar to dust-modified Alfvén ion modes in a magnetized plasma cylinder [22]. It has been observed that the frequency of both modes enhances with an increase in magnetic field strength. The growth rate of low-frequency transverse modes is reduced by increasing the field strength, while the growth rate of compressional modes remains unaffected. These waves were found to be predominant in the low-collisional and long wavelength regions of the strongly coupled laboratory plasmas. In the operating regime of fusion devices, these waves are insignificant and undesirable [44]. The study of strongly coupled dust grains in addition to the magnetic field is important in the case of the outer layer of neutron stars [45].

Our model may find an application, where the plasma is magnetically confined along the axis of plasma cylinder [26], [46], [47], [48], [49], [50]. The model may be validated by studying the excitation of low-frequency waves in magnetized dusty plasma having strongly correlated dust grains in the MDPX [32], provided that the plasma radius is small enough to have an optimum magnetic field strength below 0.2 T. In fusion devices, where a large magnetic field is required, the low-frequency oscillations are nearly damped for all the magnetic field strengths provided the regime, and ( $f\tau_m \ll 1$ ) is maintained during the plasma confinement.

#### ACKNOWLEDGMENT

Harender Mor is thankful to the Council for Scientific and Industrial Research (CSIR), Government of India, for providing necessary financial support under its fellowship. The authors are also thankful to DTU for providing infrastructural facilities.

#### REFERENCES

- [1] M. Y. Pustynnik, A. A. Pikalev, A. V. Zobnin, I. L. Semenov, H. M. Thomas, and O. F. Petrov, "Physical aspects of dust-plasma interactions," *Contributions Plasma Phys.*, vol. 61, no. 10, Nov. 2021, Art. no. e202100126, doi: [10.1002/ctpp.202100126](https://doi.org/10.1002/ctpp.202100126).
- [2] R. Merlino, "Dusty plasmas: From Saturn's rings to semiconductor processing devices," *Adv. Phys. X*, vol. 6, no. 1, Jan. 2021, Art. no. 1873859, doi: [10.1080/23746149.2021.1873859](https://doi.org/10.1080/23746149.2021.1873859).
- [3] P. K. Shukla and A. A. Mamun, "Introduction to dusty plasma physics," *Plasma Phys. Control. Fusion*, vol. 44, no. 3, p. 395, 2002, doi: [10.1088/0741-3335/44/3/701](https://doi.org/10.1088/0741-3335/44/3/701).
- [4] A. Melzer and J. Goree, "Fundamentals of dusty plasmas," in *Low Temperature Plasmas: Fundamentals, Technologies and Techniques*. Weinheim, Germany: Wiley, 2008, pp. 157–206. [Online]. Available: [http://wsx.lanl.gov/RSX/PPSS\\_2006/lectures/Goree\\_LANL\\_PPSS07.pdf](http://wsx.lanl.gov/RSX/PPSS_2006/lectures/Goree_LANL_PPSS07.pdf)
- [5] V. M. Donnelly and A. Kornblit, "Plasma etching: Yesterday, today, and tomorrow," *J. Vac. Sci. Technol. A, Vac., Surf., Films*, vol. 31, no. 5, Sep. 2013, Art. no. 050825, doi: [10.1116/1.4819316](https://doi.org/10.1116/1.4819316).
- [6] J. Winter, "Dust: A new challenge in nuclear fusion research?" *Phys. Plasmas*, vol. 7, no. 10, pp. 3862–3866, Oct. 2000, doi: [10.1063/1.1288911](https://doi.org/10.1063/1.1288911).
- [7] M. Rubel et al., "Dust generation in tokamaks: Overview of beryllium and tungsten dust characterisation in JET with the ITER-like wall," *Fusion Eng. Design*, vol. 136, pp. 579–586, Nov. 2018, doi: [10.1016/j.fusengdes.2018.03.027](https://doi.org/10.1016/j.fusengdes.2018.03.027).
- [8] J. Blum, "Dust evolution in protoplanetary discs and the formation of planetesimals: What have we learned from laboratory experiments?" *Space Sci. Rev.*, vol. 214, no. 2, p. 52, Mar. 2018, doi: [10.1007/s11214-018-0486-5](https://doi.org/10.1007/s11214-018-0486-5).
- [9] M. K. Islam, Y. Nakashima, K. Yatsu, and M. Salimullah, "On low-frequency dust-modes in a collisional and streaming dusty plasma with dust charge fluctuation," *Phys. Plasmas*, vol. 10, no. 3, pp. 591–595, Feb. 2003, doi: [10.1063/1.1539474](https://doi.org/10.1063/1.1539474).
- [10] M. R. Jana, A. Sen, and P. K. Kaw, "Collective effects due to charge-fluctuation dynamics in a dusty plasma," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 48, no. 5, pp. 3930–3933, Nov. 1993, doi: [10.1103/PhysRevE.48.3930](https://doi.org/10.1103/PhysRevE.48.3930).
- [11] N. N. Rao, P. K. Shukla, and M. Y. Yu, "Dust-acoustic waves in dusty plasmas," *Planetary Space Sci.*, vol. 38, no. 4, pp. 543–546, 1990, doi: [10.1016/0032-0633\(90\)90147-I](https://doi.org/10.1016/0032-0633(90)90147-I).
- [12] M. Rosenberg, "Ion- and dust-acoustic instabilities in dusty plasmas," *Planet. Space Sci.*, vol. 41, no. 3, pp. 229–233, Mar. 1993, doi: [10.1016/0032-0633\(93\)90062-7](https://doi.org/10.1016/0032-0633(93)90062-7).
- [13] M. Rosenberg, "Ion-dust streaming instability in processing plasmas," *J. Vac. Sci. Technol. A, Vac., Surf., Films*, vol. 14, no. 2, pp. 631–633, Mar. 1996, doi: [10.1116/1.580157](https://doi.org/10.1116/1.580157).
- [14] J. H. Chu, J.-B. Du, and I. Lin, "Coulomb solids and low-frequency fluctuations in RF dusty plasmas," *J. Phys. D, Appl. Phys.*, vol. 27, no. 2, pp. 296–300, Feb. 1994, doi: [10.1088/0022-3727/27/2/018](https://doi.org/10.1088/0022-3727/27/2/018).
- [15] J. B. Pieper, J. Goree, and R. A. Quinn, "Experimental studies of two-dimensional and three-dimensional structure in a crystallized dusty plasma," *J. Vac. Sci. Technol. A, Vac., Surf., Films*, vol. 14, no. 2, pp. 519–524, Mar. 1996, doi: [10.1116/1.580118](https://doi.org/10.1116/1.580118).
- [16] S. Nunomura, J. Goree, S. Hu, X. Wang, and A. Bhattacharjee, "Dispersion relations of longitudinal and transverse waves in two-dimensional screened Coulomb crystals," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 65, no. 6, p. 66402, Jun. 2002, doi: [10.1103/PhysRevE.65.066402](https://doi.org/10.1103/PhysRevE.65.066402).
- [17] S. Ichimaru, "Strongly coupled plasmas: High-density classical plasmas and degenerate electron liquids," *Rev. Mod. Phys.*, vol. 54, no. 4, pp. 1017–1059, Oct. 1982, doi: [10.1103/RevModPhys.54.1017](https://doi.org/10.1103/RevModPhys.54.1017).
- [18] J. H. Chu and L. I., "Direct observation of Coulomb crystals and liquids in strongly coupled RF dusty plasmas," *Phys. Rev. Lett.*, vol. 72, no. 25, pp. 4009–4012, Jun. 1994, doi: [10.1103/PhysRevLett.72.4009](https://doi.org/10.1103/PhysRevLett.72.4009).
- [19] P. K. Kaw and A. Sen, "Low frequency modes in strongly coupled dusty plasmas," *Phys. Plasmas*, vol. 5, no. 10, pp. 3552–3559, Oct. 1998, doi: [10.1063/1.873073](https://doi.org/10.1063/1.873073).
- [20] P. K. Kaw, "Collective modes in a strongly coupled dusty plasma," *Phys. Plasmas*, vol. 8, no. 5, pp. 1870–1878, May 2001, doi: [10.1063/1.1348335](https://doi.org/10.1063/1.1348335).
- [21] J. Pramanik, G. Prasad, A. Sen, and P. K. Kaw, "Experimental observations of transverse shear waves in strongly coupled dusty plasmas," *Phys. Rev. Lett.*, vol. 88, no. 17, Apr. 2002, Art. no. 175001, doi: [10.1103/PhysRevLett.88.175001](https://doi.org/10.1103/PhysRevLett.88.175001).
- [22] B.-S. Xie and Y.-P. Chen, "Low frequency modes in strongly coupled magnetized dusty plasmas," *Phys. Plasmas*, vol. 11, no. 7, pp. 3519–3524, Jun. 2004, doi: [10.1063/1.1756586](https://doi.org/10.1063/1.1756586).

- [23] D. Banerjee, J. S. Mylavaram, and N. Chakrabarti, "Viscoelastic modes in a strongly coupled, cold, magnetized dusty plasma," *Phys. Plasmas*, vol. 17, no. 11, Nov. 2010, Art. no. 113708, doi: [10.1063/1.3515897](#).
- [24] K. R. Segwal and S. C. Sharma, "A nonlocal theory of current-driven low-frequency modes in a magnetized strongly coupled collisional dusty plasma," *IEEE Trans. Plasma Sci.*, vol. 47, no. 7, pp. 3087–3099, Jul. 2019, doi: [10.1109/TPS.2019.2906035](#).
- [25] C. S. Liu and V. K. Tripathi, "Parametric instabilities in a magnetized plasma," *Phys. Rep.*, vol. 130, no. 3, pp. 143–216, 1986, doi: [10.1016/0370-1573\(86\)90108-0](#).
- [26] G. R. Tynan et al., "Observation of turbulent-driven shear flow in a cylindrical laboratory plasma device," *Plasma Phys. Controlled Fusion*, vol. 48, no. 4, pp. S51–S73, Apr. 2006, doi: [10.1088/0741-3335/48/4/S05](#).
- [27] S. Ichimaru, H. Iyetomi, and S. Tanaka, "Statistical physics of dense plasmas: Thermodynamics, transport coefficients and dynamic correlations," *Phys. Rep.*, vol. 149, nos. 2–3, pp. 91–205, May 1987, doi: [10.1016/0370-1573\(87\)90125-6](#).
- [28] W. L. Slattery, G. D. Doolen, and H. E. DeWitt, "Improved equation of state for the classical one-component plasma," *Phys. Rev. A, Gen. Phys.*, vol. 21, no. 6, pp. 2087–2095, Jun. 1980, doi: [10.1103/PhysRevA.21.2087](#).
- [29] W. L. Slattery, G. D. Doolen, and H. E. DeWitt, "N dependence in the classical one-component plasma Monte Carlo calculations," *Phys. Rev. A, Gen. Phys.*, vol. 26, no. 4, pp. 2255–2258, Oct. 1982, doi: [10.1103/PhysRevA.26.2255](#).
- [30] A. F. Alexandrov, L. S. Bogdankevich, and A. A. Rukhadze, *Principles of Plasma Electrodynamics*, vol. 9. Berlin, Germany: Springer, 1984.
- [31] M. Abramowitz, I. A. Stegun, and R. H. Romer, "Handbook of mathematical functions with formulas, graphs, and mathematical tables," *Amer. J. Phys.*, vol. 56, no. 10, p. 958, Oct. 1988, doi: [10.1119/1.15378](#).
- [32] E. Thomas et al., "The magnetized dusty plasma experiment (MDPX)," *J. Plasma Phys.*, vol. 81, no. 2, pp. 1–21, Apr. 2015, doi: [10.1017/S0022377815000148](#).
- [33] E. Thomas et al., "Preliminary characteristics of magnetic field and plasma performance in the magnetized dusty plasma experiment (MDPX)," *J. Plasma Phys.*, vol. 80, no. 6, pp. 803–808, Dec. 2014, doi: [10.1017/S0022377814000270](#).
- [34] M. Choudhary, R. Bergert, S. Mitic, and M. H. Thoma, "Influence of external magnetic field on dust acoustic waves in a capacitive RF discharge," *Contributions Plasma Phys.*, vol. 60, no. 2, Feb. 2020, Art. no. e201900115, doi: [10.1002/ctpp.201900115](#).
- [35] A. Piel, *Plasma Physics: An Introduction to Laboratory, Space, and Fusion Plasmas*, 1st ed. Berlin, Germany: Springer, 2010, doi: [10.1007/978-3-642-10491-6](#).
- [36] L. F. Shampine and M. W. Reichelt, "The MATLAB ODE suite," *SIAM J. Sci. Comput.*, vol. 18, no. 1, pp. 1–22, Jan. 1997, doi: [10.1137/S1064827594276424](#).
- [37] P. Bandyopadhyay, G. Prasad, A. Sen, and P. K. Kaw, "Experimental observation of strong coupling effects on the dispersion of dust acoustic waves in a plasma," *Phys. Lett. A*, vol. 368, no. 6, pp. 491–494, Sep. 2007, doi: [10.1016/j.physleta.2007.04.048](#).
- [38] P. Kaw and R. Singh, "Collisional instabilities in a dusty plasma with recombination and ion-drift effects," *Phys. Rev. Lett.*, vol. 79, no. 3, pp. 423–426, Jul. 1997, doi: [10.1103/PhysRevLett.79.423](#).
- [39] V. I. Molotkov, A. P. Nefedov, V. M. Torchinskii, V. E. Fortov, and A. G. Khrapak, "Dust acoustic waves in a DC glow-discharge plasma," *J. Exp. Theor. Phys.*, vol. 89, no. 3, pp. 477–480, Sep. 1999, doi: [10.1134/1.559006](#).
- [40] M. C. D. Juli, R. S. Schneider, L. F. Ziebell, and V. Jatenco-Pereira, "Effects of dust-charge fluctuation on the damping of Alfvén waves in dusty plasmas," *Phys. Plasmas*, vol. 12, no. 5, p. 52109, May 2005, doi: [10.1063/1.1899647](#).
- [41] A. A. Mamun and P. K. Shukla, "Streaming instabilities in a collisional dusty plasma," *Phys. Plasmas*, vol. 7, no. 11, pp. 4412–4417, Nov. 2000, doi: [10.1063/1.1315305](#).
- [42] S. A. Khrapak, A. G. Khrapak, N. P. Kryuchkov, and S. O. Yurchenko, "Onset of transverse (shear) waves in strongly-coupled Yukawa fluids," *J. Chem. Phys.*, vol. 150, no. 10, Mar. 2019, Art. no. 104503, doi: [10.1063/1.5088141](#).
- [43] E. Thomas, U. Konopka, R. L. Merlino, and M. Rosenberg, "Initial measurements of two- and three-dimensional ordering, waves, and plasma filamentation in the magnetized dusty plasma experiment," *Phys. Plasmas*, vol. 23, no. 5, Mar. 2016, Art. no. 055701, doi: [10.1063/1.4943112](#).
- [44] S. I. Krasheninnikov, R. D. Smirnov, and D. L. Rudakov, "Dust in magnetic fusion devices," *Plasma Phys. Controlled Fusion*, vol. 53, no. 8, Aug. 2011, Art. no. 083001, doi: [10.1088/0741-3335/53/8/083001](#).
- [45] A. Y. Potekhin, "The physics of neutron stars," *Physics-Uspekhi*, vol. 53, no. 12, pp. 1235–1256, Dec. 2010, doi: [10.3367/UFNe.0180.201012c.1279](#).
- [46] C. Rapozo, "RF heating by cylindrical plasma waveguide modes," *Revista Brasileira de Física*, vol. 17, no. 2, pp. 1–14, 1987.
- [47] N. Buzarbaruah, N. J. Dutta, D. Borgohain, S. R. Mohanty, and H. Bailung, "Study on discharge plasma in a cylindrical inertial electrostatic confinement fusion device," *Phys. Lett. A*, vol. 381, no. 30, pp. 2391–2396, Aug. 2017, doi: [10.1016/j.physleta.2017.05.029](#).
- [48] M. Amberg, J. Geerk, M. Keller, and A. Fischer, "Design, characterisation and operation of an inverted cylindrical magnetron for metal deposition," *Plasma Devices Oper.*, vol. 12, no. 3, pp. 175–186, Sep. 2004, doi: [10.1080/1051999042000237988](#).
- [49] J. Badziak et al., "Formation of a supersonic laser-driven plasma jet in a cylindrical channel," *Phys. Plasmas*, vol. 16, no. 11, Nov. 2009, Art. no. 114506, doi: [10.1063/1.3270520](#).
- [50] T. Windisch, O. Grulke, V. Naulin, and T. Klinger, "Intermittent transport events in a cylindrical plasma device: Experiment and simulation," *Plasma Phys. Controlled Fusion*, vol. 53, no. 8, Aug. 2011, Art. no. 085001, doi: [10.1088/0741-3335/53/8/085001](#).



**Harender Mor** was born in Narwana, Haryana, India. He received the B.Sc. degree from Kurukshetra University Kurukshetra, Thanesar, Haryana, India, in 2015, and the M.Sc. degree from DCRUST Murthal, Murthal, India, in 2017. He is currently pursuing the Ph.D. degree with the Department of Applied Physics, Delhi Technological University (DTU), Delhi, India.

He is currently working in the field of strongly coupled dusty plasma physics.

Mr. Mor is a recipient of Junior Research Fellowship (JRF) and Senior Research Fellowship (SRF) from the CSIR Government of India.



**Kavita Rani Segwal** received the M.Sc. and M.Phil. degrees in physics from Maharishi Dayanand University, Rohtak, Haryana, India, in 2007 and 2008, respectively, and the Ph.D. degree from Delhi Technological University, New Delhi, India, in 2019.

She is an Associate Professor with the Department of Engineering and Technology, Delhi, India. Her current research interests include plasma physics, dusty and strongly coupled dusty plasmas, and condensed matter physics.



**Suresh C. Sharma** (Senior Member, IEEE) is working as a Professor with the Department of Applied Physics, Delhi Technological University (DTU), Delhi, India, and also held administrative responsibility of a Dean (Acad-PG) and a HoD (Applied Physics). He established Plasma and Nanosimulation research Laboratory; EMT, Antenna and Propagation Laboratory, and Microwave Engg. Laboratory at DTU. He has guided 15 Ph.D. students and several M.Tech. and B.Tech. students. He has published 207 research articles in journals of international and national repute and proceedings of international and national conferences.

Dr. Sharma was a recipient of the Young Scientist project as a Principal Investigator by the Department of Science and Technology (DST), Government of India for two years (1997–1999). He was a Monbusho Post-Doctoral Fellow under Japanese Government fellowship, Department of Physics, Faculty of Science, Ehime University, Matsuyama, Japan, from October 1997 to March 1999. In addition, he has been a JSPS (Invitation) Post-Doctoral Fellow and a Visiting Researcher with Center for Atomic and Molecular Technologies (CAMT), Osaka University, Japan, from May 2004 to October 2005. Also, he was awarded as a Senior Research Associate under the Scientist's Pool Scheme by CSIR, Government of India for three years (1999–2002). He was awarded commendable Research Award for Excellence in Research by DTU for six consecutive years, i.e., March 2018, March 2019, March 2020, February 2021, March 2022, and April 2023. He is a member of the American Physical Society (APS), USA, and many more.



2. Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

# Miniaturized Quad-Port Conformal Multi-Band (QPC-MB) MIMO Antenna for On-Body Wireless Systems in Microwave-Millimeter Bands

Manish Sharma<sup>1</sup>, (SMIEEE), Prabhakara Rao Kapula<sup>2</sup>, Shailaja Alagrama<sup>3</sup>, Kanhaiya Sharma<sup>4</sup>, Ganga Prasad Pandey<sup>5</sup>, Dinesh Kumar Singh<sup>6</sup>, Milind Mahajan<sup>7</sup>, Anupma Gupta<sup>8</sup>

<sup>1</sup>Manish Sharma, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India, manishengineer1978@gmail.com

<sup>2</sup>Prabhakara Rao Kapula, Department of Electronics & Communication, B V Raju Institute of Technology, Narsapur, Telangana 502313, India  
Prabhakar.kapula@bvrit.ac.in

<sup>3</sup>Information Technology, University of the Cumberland's, Williamsburg, Kentucky, USA, 40769 shaila25@me.com

<sup>4</sup>Kanhaiya Sharma, Department of Computer Science and Engineering, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India, sharmakanhaiya@gmail.com

<sup>5</sup>Ganga Prasad Pandey, Department of Information Communication & Technology, School of Technology, Pandit Deendayal Energy University, Gujarat India.

<sup>6</sup>Dinesh Kumar Singh, Department of Electronics and Communication Engineering, G L Bajaj Institute of Technology and Management, Greater Noida, UP, India.

<sup>7</sup>Milind Mahajan, Scientist at Space Applications Centre, ISRO, Ahmedabad Space Applications Centre, ISRO, Ahmedabad, Gujarat, India, mb\_mahajan@sac.isro.gov.in

<sup>8</sup>Anupma Gupta, Department of Interdisciplinary Courses in Engineering, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India, anupmagupta31@gmail.com

**ABSTRACT** In this endeavor, miniaturized quad-port MIMO<sub>MB</sub> (multiple-input-multiple-output multi-band) antenna with an overall area of 400 mm<sup>2</sup> offering wider-impedance-bandwidth of 8.31GHz-36.14GHz is reported which covers multiple bands including X-band (8.00-12.0 GHz), Ku-band (12.0-18.0 GHz), K-band (18.0-27.0 GHz), partial Ka-band (27.0-40.0 GHz), FR2: n257 (26.50-29.50 GHz), n258 (24.25-27.50) and n261 (27.50-28.35 GHz) is reported for wideband infrastructure. The proposed antenna radiating EM energy is printed on very thin Rogers RTDuroid5880 substrate with thickness 0.254mm. The radiating EM wave patch consists of hexagonal geometry which is etched with a circular-rectangular slot and partial-ground etched by a beveled shape patch for matching of impedance. The conformal capability of the proposed antenna is verified by S-for single-port, dual-port, and proposed four-port antenna. The time-domain analysis confirms the faithful reception of transmitted signals in far-field regions. The diversity performance including ECC<sub>MB</sub>, DG<sub>MB</sub>, TARC<sub>MB</sub>, CCL<sub>MB</sub>, and MEG<sub>MB</sub> is below permissible standard values with a maximum peak gain of 7.17dBi with stable 2-D radiation patterns in principal planes. The Specific-Absorption-Rate<sub>MB</sub> (SAR) analysis is carried out at different operating frequencies in Microwave-Millimeter wave bands with SAR ≤1.60W/Kg in human phantom tissue and makes it suitable for on-body wireless applications.

**INDEX TERMS** Conformal MIMO patch, multiband, Microwave-Millimeter wave bands, thin substrate, ECC<sub>MB</sub>, DG<sub>MB</sub>, TARC<sub>MB</sub>, CCL<sub>MB</sub> and MEG<sub>MB</sub>, SAR, human phantom tissue

## I. INTRODUCTION

The development of planar technology in the last few decades related to the designing of antennae has been able to attract several researchers. In today's scenario, the motherboard has become very compact thereby reducing the size of the devices. It will be an advantage if the antenna can be more compact, and conformal capability with acceptable Specific-absorption-rate (SAR) that can be used for multiband applications. This literature discusses

various microwave-millimeter wave application antennae with conformal capabilities and controllable SAR. A compact 28.0GHz antenna with an arc-shaped patch provides an impedance bandwidth of 25.83GHz-30.24GHz [1]. Also, a slotted square patch with an embedded T-shaped stub and defective ground resonates at 28.0GHz with a gain of 11.50dBi and 94% efficiency [2]. Utilizing a two-electromagnetically coupled patch offers a dual band of operation at 38.0GHz and 60.0GHz millimeter-Wave

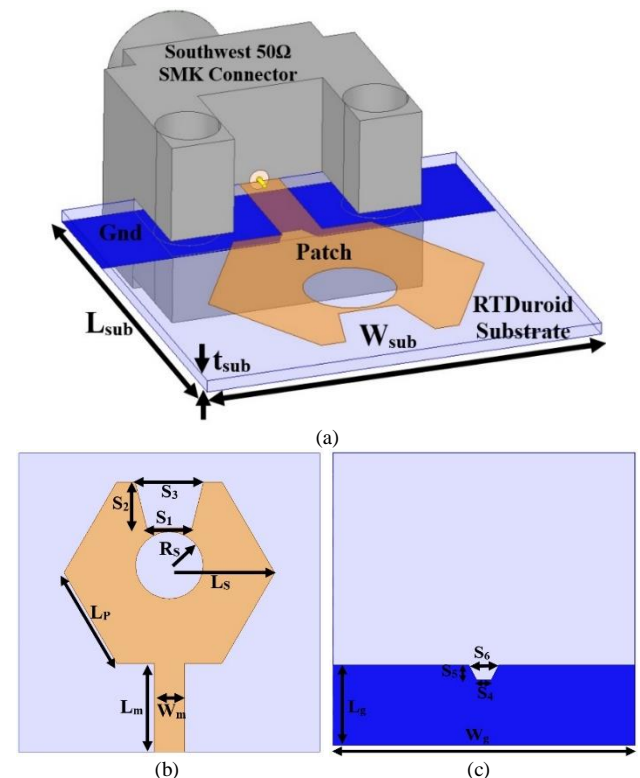
band [3]. A very compact size antenna that resonates at 60.0GHz millimeter-Wave is applicable for body area networks [4] and a genetic-algorithm-based four-band single-port antenna is designed with a population size of 30 with genes in chromosomes equivalent to 36 [5]. A compact antenna [6] for body-centric relay-mode communication is designed for 5.0GHz with high directivity of 7.18dBi. A single-port guitar-shaped patch produces wide operating bandwidth of 2.76GHz-35.93GHz [7], [11], [21], [24] and a low profile two-port dual band is achieved by using a dome-shaped patch with CPW-fees and T-shaped connected stub helps in achieving high isolation [8]. A hexagonal-ring-shaped stub with tapered feed produces dual bands resonating at 28.0GHz/38.0GHz mm-Wave bands [9], [22], [23]. A wide gap between the inter-spaced element of 52.0mm with the combination of circular patch fed by tapered-feed offers more than 10:1 ratio bandwidth between 3.10GHz-20.0GHz [10]. Triple-resonating bands are achieved by utilizing DCOLR (defective-complimentary-open-loop-resonator) which also includes defective feed inclusive of dual-stepped resonator(s) with partial-ground [12]. Better diversity characteristics inclusive of  $\text{isolation} \geq 16.0\text{dB}$  are achieved in 2-ports UWB-monopole antenna [13]-[14]. Four-port MIMO antenna with bandwidth 24.55GHz-26.50GHz is designed for a 24.0GHz center frequency mm-Wave band and the gain is achieved by using a  $9 \times 6$  circular Split Ring (CSR) [15]. A low-frequency 5G-NR band with four-port accommodates n77, n78, and n79 bands [16], [17], [18], [19]. A super wideband antenna with four notched bands offers operational bandwidth of 1.15GHz-40.0GHz and maintains isolation by arranging the truncated elliptical self-complimentary patch in an orthogonal fashion [20]. 8-port narrow-band (5G applications) [25] and super wideband [26] utilizes flower-shaped patch with side L-shaped stub attached to the ground providing maximum radiation in the desired direction. A review on the conformal antenna is reported [27], [28], [29] which utilizes substrates such as PET, and thin Rogers substrate for bending of antenna applications. A conformal antenna designed on an FR4 substrate utilizes a very thin dielectric thickness of 0.15mm and produces applications within 7.20GHz-9.20GHz [30], [33]. A deployable wearable antenna is modeled on a body with a thickness of skin=2.00mm, fat=4.00mm, and muscle=10.0mm producing a maximum specific-absorption rate of 0.512 W/Kg at 2.45GHz [31]. A breast-cancer antenna is capable of detecting tumors which is traceable between a bandwidth of 8.50GHz-10.50GHz [32], [34]. A detailed investigation of SAR and thermal effects is studied by using patch-antenna when deployed on the body and also checking on the fabric-cotton wearable applications with antenna providing resonance at 1.80GHz, 2.40GHz, and 5.00GHz and 8.90GHz. Classical theory of characteristics modes is applied on conformal MIMO antenna which is designed in accordance with the electromagnetic properties of the implantable antenna with

I-type patches between dual-radiators and slot-structure in ground achieving good isolation [37]. Utilizing partially reflecting-surface with Dielectric-Resonator-Antenna (DRA) improves the gain of the antenna [38]-[39]. In this work, a slotted hexagon patch-geometry in combination with beveled ground below the patch forms wideband antenna with multi-band applications designed and analyzed on low permittivity 2.20 flexible substrate. The single-antenna is transformed to two-port and four-port MIMO configuration for higher data rate of transmission with reduction of fading. The calculated SAR at different resonance confirms the value below 1.60 W/Kg make it more feasible for wearable and hand-held devices. The detailed study is discussed in the sections given below

## II. CONFORMAL ANTENNA

Wireless communication has emerged as an impressive technology that has developed to a large extent in delivering data with higher data rates and low latency through the implementation of 5G technology. A multiband antenna is desirable with very compact which can be useful for multiple-wireless applications. In achieving the above objective with multiple bands embedded in single-antenna, an ultra-compact antenna is proposed with inheriting the capability of conformal characteristics which can be used for wearable devices, and the design methodology with supporting results is discussed below

### (a) SINGLE-PORT CONFORMAL ANTENNA

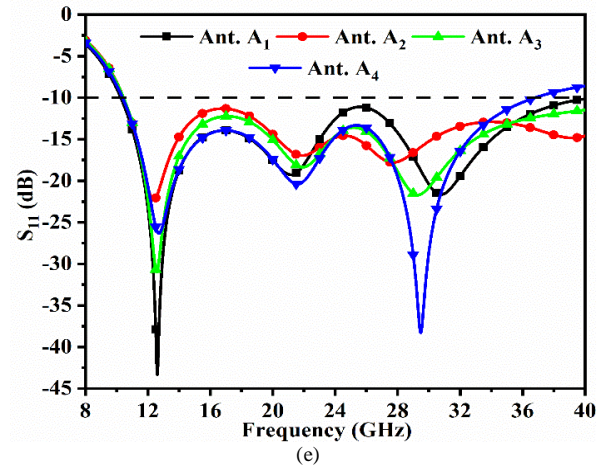
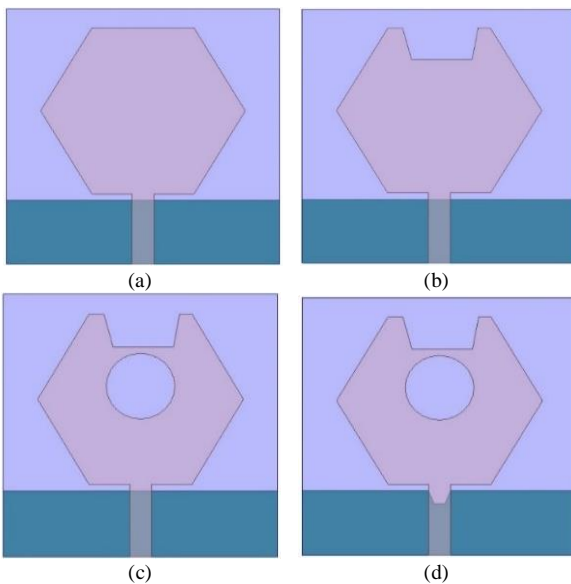


Details of the multiband-antenna<sub>MB</sub> are depicted in Fig. 1 shows which is printed on a thin RogersRT™ Duroid dielectric. The top plane of the dielectric is printed with a radiator patch which is connected to a 50Ω-microstrip feedline with an overall antenna dimension of  $L_{sub} \times W_{sub} \times t_{sub}$  mm<sup>3</sup>. The designed radiator which is capable of generating wider impedance-bandwidth is united with optimized feed and is connected to matched SMK connector for signal input between 5.0GHz-40.0GHz. The opposite plane is printed with fractal-ground with a bent slot to achieve wider impedance bandwidth with the matching of impedance as shown in Fig. 1(a). Fig. 1(b) illustrates the front view of the slotted-hexagonal radiator which consists of a polygon shape with side-length  $L_p$  mm. The radiating patch is carved with a circular slot of radius  $R_s$  (in mm) and a trapezoidal slot of area  $((S_1+S_3)/2 \times S_2)$  mm<sup>2</sup>. On the opposite face of the dielectric, partially-imprinted ground includes a beveled-shaped slot as observed in Fig. 1(c). The partial ground occupies an area of  $L_g \times W_g$  mm<sup>2</sup> and gap  $(L_m-L_g)$  mm is responsible for -10dB wide operational bandwidth. The optimal dimensions extracted from the EM-simulation Ansys HFSS are recorded in Table 1 given below (Fig. 1)

**Table 1:** Optimal values tabulated from Fig. 1.

Single Port					
Dimension	mm	Dimension	mm	Dimension	mm
$L_{sub}$	10.0	$W_{sub}=W_g$	10.0	$t_{sub}$	0.254
$L_m$	2.75	$W_m$	0.80	$L_p$	3.75
$R_s$	1.25	$L_s$	3.75	$S_1$	2.20
$S_2$	1.30	$S_3$	1.25	$S_4$	0.40
$S_5$	0.75	$S_6$	0.50	$L_g$	2.50
Dual Port					
Dimension	mm	Dimension	mm	Dimension	mm
$W_{g1}$	20.0	$L_{g1}$	10.0	$K$	9.00
$S_L$	7.50	$S_W$			0.50
Four Port					
Dimension	mm	Dimension	mm	Dimension	mm
$L_{g2}$	20.0	$W_{g2}$	20.0	$K$	9.00

## (b) EVOLUTION



**Fig. 2.** Configuration of Single-Port Conformal (a) Ant. A<sub>1</sub> (b) Ant. A<sub>2</sub> (c) Ant. A<sub>3</sub> (d) Ant. A<sub>4</sub> (e)  $S_{11}$  result of Ant. A<sub>1</sub>-Ant. A<sub>4</sub>.

Fig. 1 antenna is achieved by carrying out several iterations and resulting in an optimal version. However, the evolution of the final version of the single-element is obtained by four iterations and the respective antenna is named Ant. A<sub>1</sub>, Ant. A<sub>2</sub>, Ant. A<sub>3</sub> and Ant. A<sub>4</sub> respectively. Ant. A<sub>1</sub> as shown in Fig. 2(a) is printed with a polygon patch and partial-rectangular ground on respective opposite planes. The polygon patch with side  $L_p$  mm is calculated by following equations [40]

$$L_p = \frac{c}{4 \times f} \sqrt{1/(1 + \epsilon_{eff})} \quad (1)$$

where  $c$  ( $c=3 \times 10^8$  m/s: the light speed with vacuum as a medium),  $f$  is the center design frequency in GHz, and  $\epsilon_{eff}$  is the effective permittivity which is calculated from Equation 2 given below [40]

$$\epsilon_{eff} = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2} \left[ 1 + \frac{12 \times t_{sub}}{W_m} \right]^{-2} \quad (2)$$

This iteration provides the impedance bandwidth of 10.28GHz-40.0GHz but with a considerably acceptable matching of impedance. Further, the polygon patch is etched by the slot of the area given by following Equation 3

$$\left( \frac{S_1 + S_3}{2} \right) \times S_2 = \left( \frac{2.20 + 1.25}{2} \right) \times 2.75 = 4.74 \text{ mm}^2 \quad (3)$$

Which helps in improving the operating bandwidth and provides single-resonance at 12.36GHz with  $S_{11}=-22.29$ dB. The third iteration is the addition of a circular slot with a radius  $R_s$  mm on the radiating patch (Ant. A<sub>3</sub>) helps in improving the matching of impedance at resonance 12.56GHz ( $S_{11}=-30.97$ dB) and 29.17GHz ( $S_{11}=-21.67$ dB) respectively. The final version of the single-port antenna is achieved by etching a beveled-shaped slot in the ground with an impedance bandwidth of 10.38GHz-36.79GHz. This version of the proposed antenna, however, the Ant. A<sub>4</sub> which is the final version of the proposed antenna, offers better impedance matching beyond 23GHz with marginally reduced bandwidth at higher cut-off frequency and does not affect the design objective. With three resonances within the impedance bandwidth and are centered at 12.71GHz ( $S_{11}=-26.34$ dB), 21.58GHz ( $S_{11}=-20.41$ dB), and 29.48GHz ( $S_{11}=-38.30$ dB).



### (c) PARAMETRIC STUDY

The final version of the mono-radiator represented by Fig. 1 which is optimized, is achieved by changing the different values of the dimension.

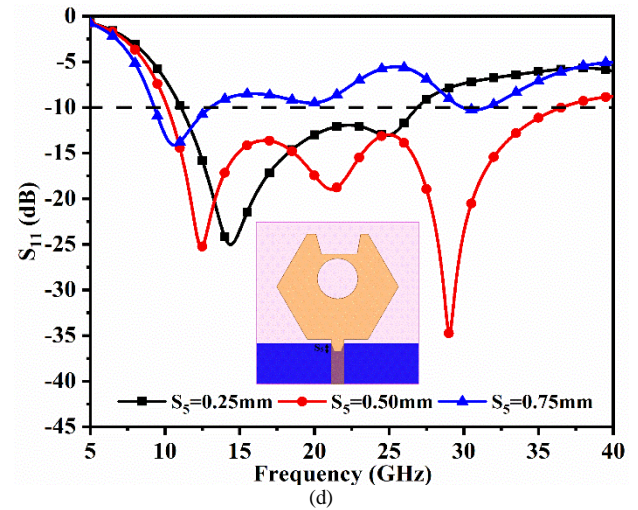
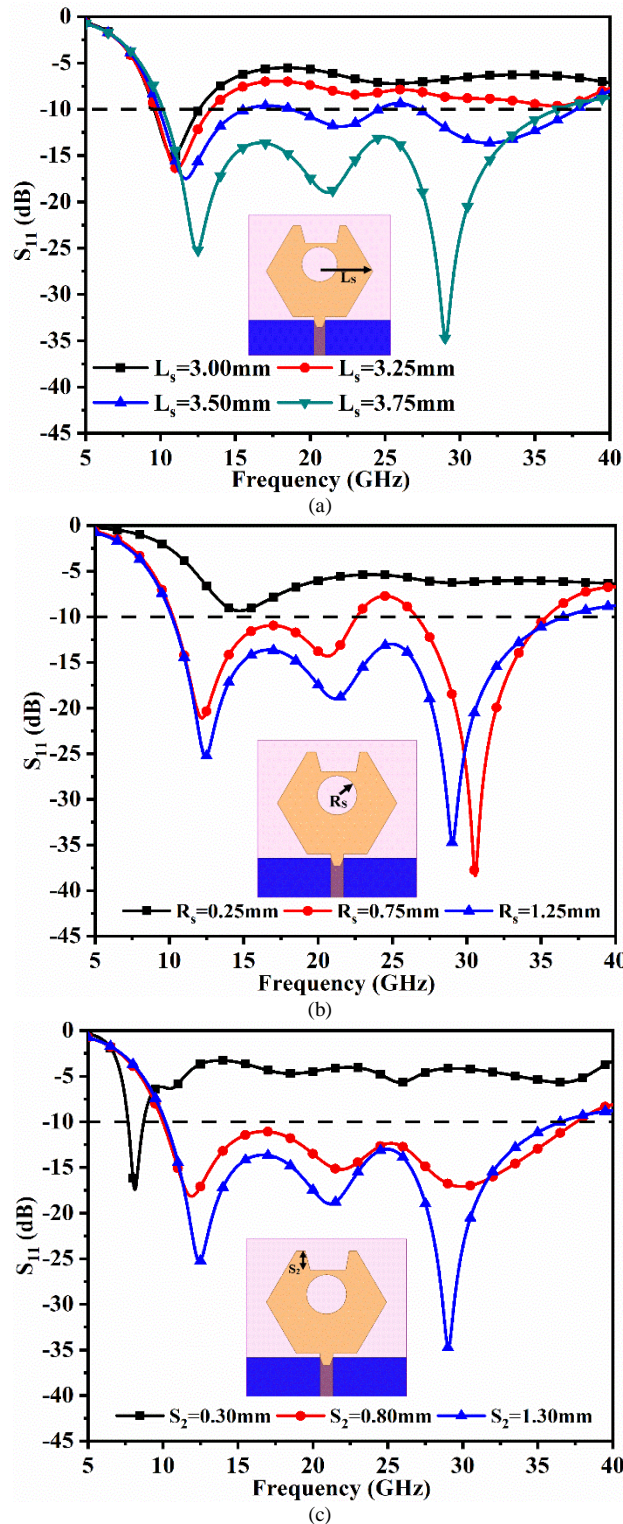


Fig. 3. Parametric Study of key parameter (a)  $L_s$  (b)  $R_s$  (c)  $S_2$  (d)  $S_5$ .

This is done by applying parametric variation to a key parameter such as  $L_s$  (radius of equivalent circle encircling the polygon patch),  $R_s$  (etched circular slot on the patch),  $S_2$  (height of etched trapezoidal slot), and  $S_5$  (height of the etched-beveled slot in the ground). The parameter  $L_s$  (in mm) which is the equivalent radius of the hexagon patch encircling it is changed from 3.00mm to 3.75mm with step size  $\Delta L_s = 0.25$ mm. The value of  $L_s = 3.00$ mm, which signifies the smaller area of the radiating patch offers narrow bandwidth of 9.59GHz-12.59GHz with resonance centered at 10.83GHz. Further increase in values of  $L_s$  at 3.25mm and 3.50mm helps in improving the matching of bandwidth. For the optimized value of  $L_s = 3.75$ mm, the bandwidth of 26.41GHz is obtained. The second key parameter  $R_s$ , which is an etched circular slot on a polygon patch observes a large swing of the  $S_{11}$  bandwidth for values of  $R_s$  between 0.25mm to 1.25mm. The lower values of  $R_s$  observe the narrow bandwidth but for  $R_s = 1.25$ mm, the objective of large-bandwidth multi-band applications is achieved as shown in Fig. 3(b) another important parameter,  $S_2$  which is the height of the trapezoidal slot etched on top of hexagonal-patch, observes improved impedance-bandwidth with the value of  $S_2$  from 0.30mm to 1.30mm. Good matched-impedance bandwidth is achieved for  $S_2 = 1.30$ mm. The beveled slot in the rectangular ground with a height of  $S_5 = 0.75$ mm achieves the highly matched impedance-bandwidth and the remaining values of  $S_5 = 0.25$ mm and 0.50mm are discarded.

### (d) IMPEDANCE GRAPH, SURFACE-CURRENT-DENSITY-DISTRIBUTION, CONFORMAL CHARACTERISTICS

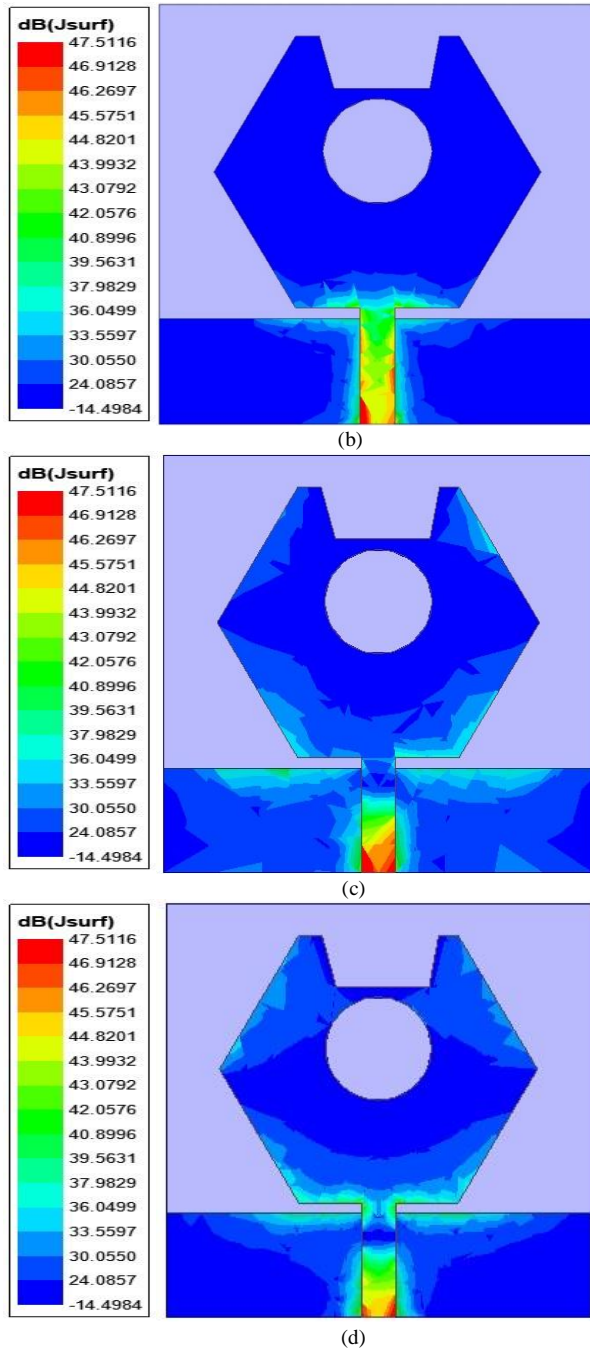


Fig. 4. Surface-Current-Density distribution at (a) 12.40GHz (b) 24.0GHz (c) 28.0GHz.

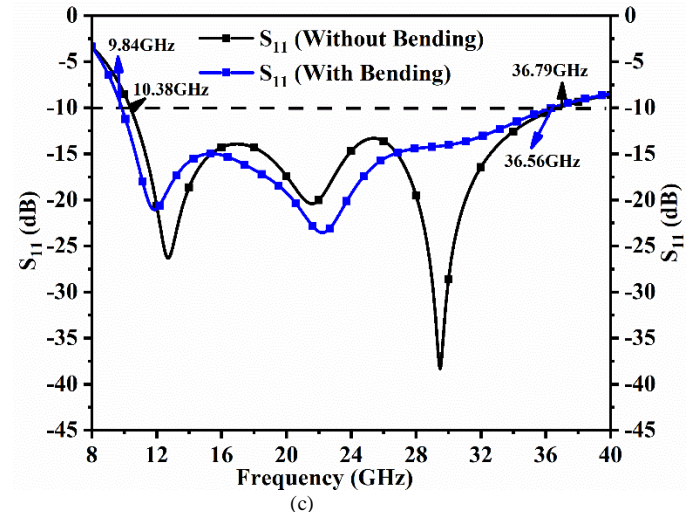
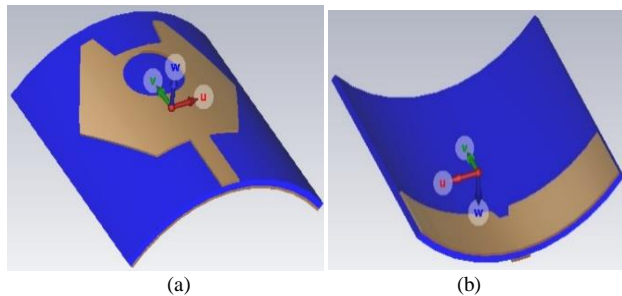


Fig. 5. Conformal Capability of Single-Port Antenna (a) Bend with a front view (b) Bend with a ground view (c) Comparison of S-parameter.

Fig. 4(a)-(c) shows the distribution of  $SCD_{MB}$  for the frequency values corresponding to 12.40GHz, 24.0GHz, and 28.0GHz respectively. In all three cases, it can be concluded that maximum  $SCD_{MB}$  is concentrated within the microstrip feed line which forms the bridge of transmission of a signal from the input port to the radiating patch. This indicates that at these frequency values, all the signal is fed to the radiating patch and the patch radiates the input signal rather than storing it. Thus, the correlation between the net impedance and radiation characteristics is easily established at these frequency values providing high gain and efficiency.

The proposed single-port antenna discussed occupies the wider-impedance bandwidth with multiple-band including X-band, 24.0GHz, and 28.0GHz bands. The proposed work utilizes a thin Rogers-Dielectric with a thickness of 0.254mm and can be easily useable for conformal applications regardless of whether the bandwidth should not be compromised. Fig. 5 shows the conformal configuration with Fig. 5(a) representing bent at 45° with front and ground view. The planar antenna (without bending) occupies an impedance bandwidth of 10.38GHz-36.79GHz and with bending corresponds to  $|S_{11}| < -10.0dB$  bandwidth of 9.84GHz-36.43GHz, thereby, the operating impedance bandwidth is not compromised and can be easily integrated for wearable multiband-applications.

### III. DUAL AND FOUR-POT CONFORMAL MIMO ANTENNA

Section II has seen the analysis of a single-port conformal antenna producing wide impedance bandwidth of 10.0GHz-39.38GHz without bending of the antenna at and 10.0GHz-36.56GHz with conformal capability at 45°. However, when they are deployed in a live wireless communication application environment, will suffer from multiple-path fading of the transmitted signals and hence, the distorted pulse will be received. This effect also

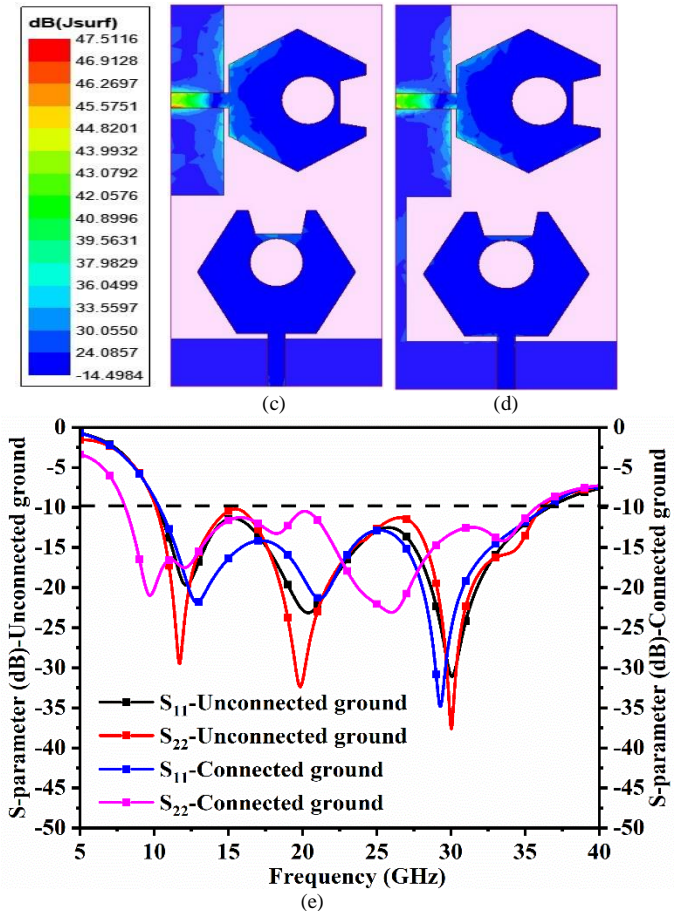
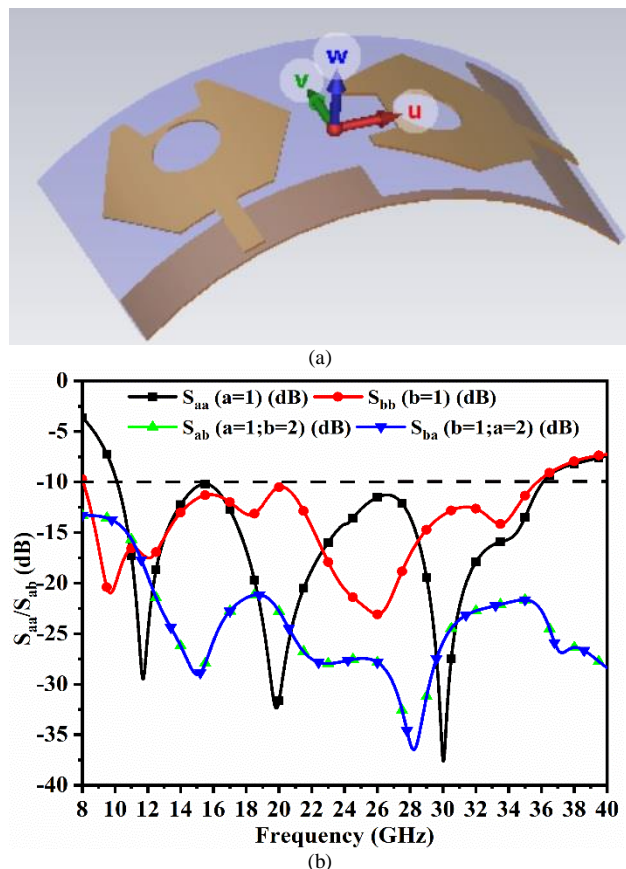


reduces the efficiency which will decrease the working bandwidth. To improve the bandwidth and to control multiple-path fading, multiple-radiating elements achieving several diversity schemes such as spatial, radiation, or polarization diversity need to be implemented while designing multiple-input-multiple-output antenna (MIMO) configuration. The MIMO configuration will enhance the bandwidth and ensures the receiving of the signal efficiently when the signal impinges at any angle on the receiver antenna. The Shannon-Hartley theorem on channel capacity is given by [41]

$$Ch.Cap_{M \times S} = D_{M \times S} \Delta B \log_2 \left( 1 + \frac{S}{N} \right) \quad (4)$$

where  $Ch.Cap_{M \times S}$  is the Channel-Capacity (CC) given by b/s/Hz,  $\Delta B$  is the matched operational bandwidth,  $D_{M \times S}$  is the integral multiple factors for MIMO configuration and corresponds to the ratio between the signal and the additional noise (S/N). Additional radiators forming MIMO<sub>MB</sub> configuration also ensure the enhancement of channel capacity thereby reducing multi-path fading issues.

Fig. 6 shows the two-port conformal MIMO antenna configuration which will reduce not only multi-path fading effects but also preserves the required impedance bandwidth for multi-band applications. Fig. 6(a) shows the simulation configuration of the conformal antenna with the bent of 45° angle and Fig. 6(b) corresponds to reflection and transmission coefficients.



**Fig. 6.** Conformal Capability of Dual-Port Antenna (a) Bend capability (b) S-parameter; Surface-current distribution at 28.0GHz (c) Unconnected ground (d) Connected ground (e) S-parameter: Unconnected and Connected ground.

As per the observations, the proposed conformal antenna maintains a bandwidth of 10.12GHz-36.10GHz (S<sub>11</sub>) for radiator 1 and 8.06GHz-35.76GHz (S<sub>22</sub>) for radiator 2.

The MIMO configuration is obtained by placing the identical modified slotted hexagon patch orthogonally with respective partial rectangular ground connected with a thin stub of dimension 7.5mm×0.50mm. Fig. 6(b) also shows the transmission coefficient or isolation between the two ports with isolation being more than 17.50dB (S<sub>12</sub>, S<sub>21</sub>) for both radiating elements. Fig. 6(c)-(d) shows the distribution of current density on the surface for unconnected and connected ground. In both cases, the required isolation is maintained and the antenna radiates efficiently. However, there is a need for common ground in MIMO-antenna configuration [35] because the signal needs to have a common-ground plane which helps in interpreting all the levels of signal properly based on the reference-zero ground level. Fig. 6(e) shows the comparison of S-parameter graph of MIMO antenna configuration with unconnected and connected ground. The impedance bandwidth in both the cases is tabulated in Table 2 given below

**Table 2:** S-parameter comparison Fig. 6(e).

S-parameter (dB)	Unconnected Ground Bandwidth (GHz)	Connected Ground Bandwidth (GHz)
$S_{11}$	10.23-36.61	10.12-36.11
$S_{22}$	10.35-36.43	8.06-35.76

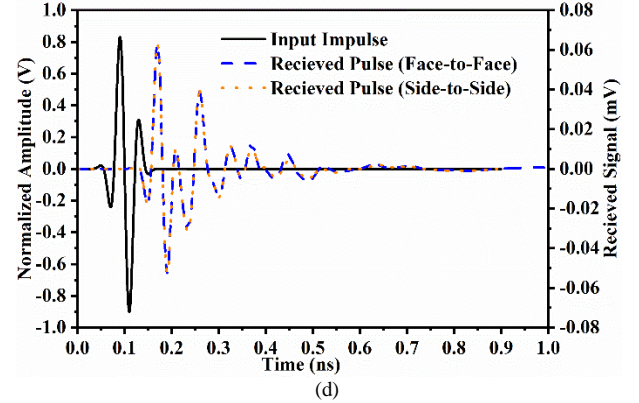
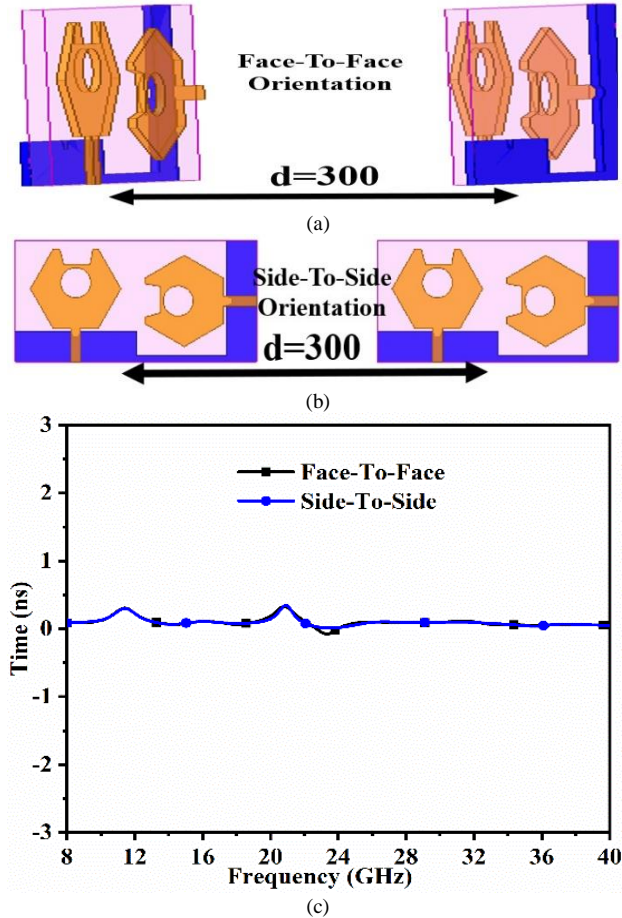
In both the cases, the MIMO antenna configuration covers the required multi-band applications bandwidths.

The proposed two-port MIMO antenna provides wide-impedance bandwidth and hence, the transmission of signal at this bandwidth needs to be evaluated which is carried out by studying time-domain performance. Fig. 7 shows the evaluation of the time-domain performance of the MIMO antenna when subjected to pulse as the input. The simulation environment is set up with identical MIMO antenna placed in face-to-face and side-to-side orientations as shown in Fig. 7(a)-(b). The minimal distance of 300mm is maintained to ensure the far-field condition given by [40]

$$\text{Far Field region} > \frac{2 \times 2 \times w_{sub}^2}{\lambda_c} \quad (5)$$

where  $2 \times W_{wub} = 20.0\text{mm}$  is the maximum dimension of the antenna and  $\lambda_c$  is the cut-off frequency in GHz.

$$\phi = -\frac{d\theta(\omega)}{d\omega} \quad (6)$$

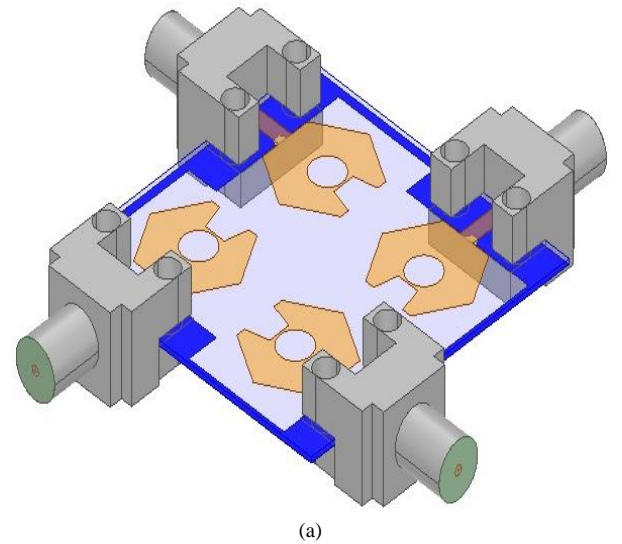


**Fig. 7.** Time-domain analysis (a) Face-to-Face alignment (b) Side-to-Side alignment (c) Group delay (d) Impulse response.

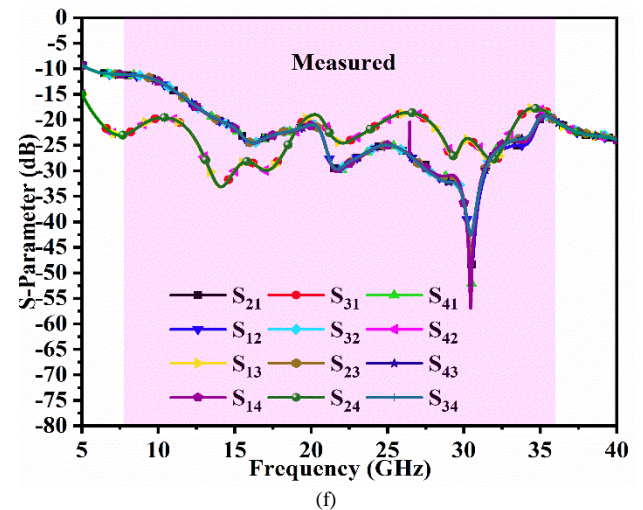
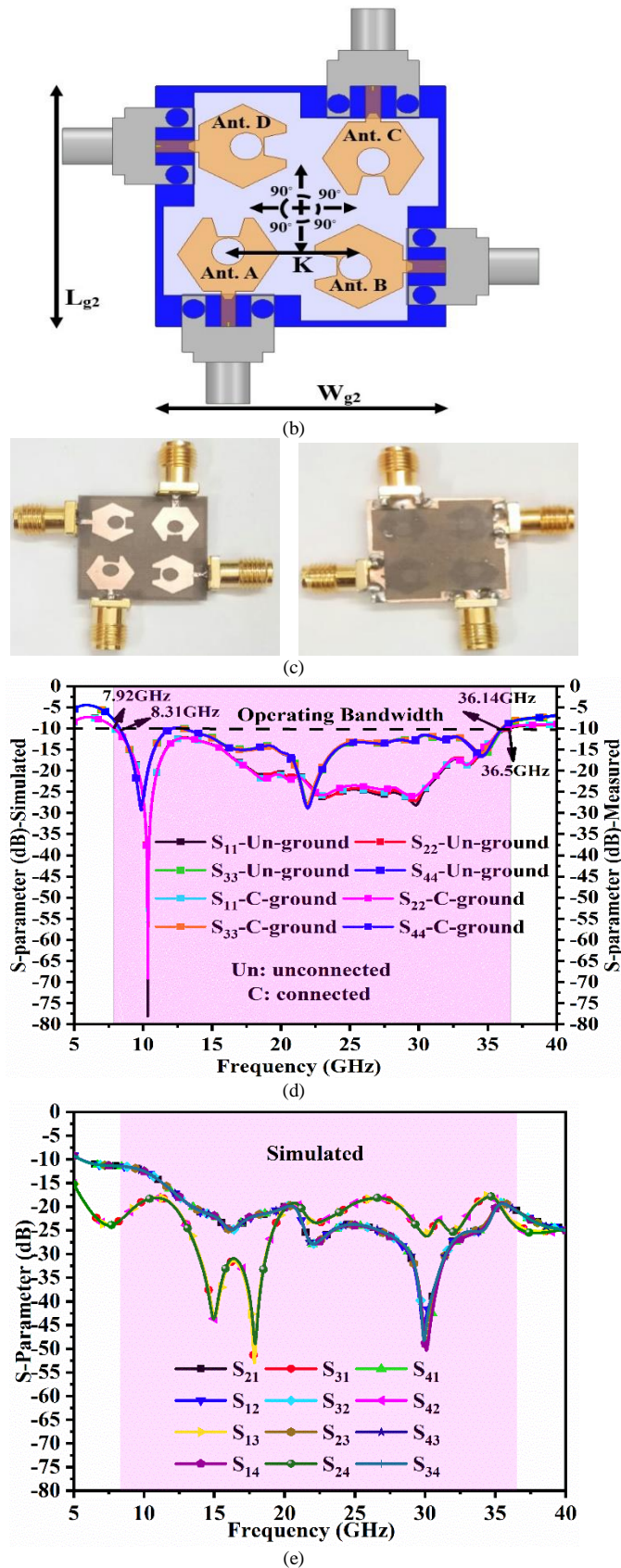
$$\rho = \max_{\phi} \left[ \frac{\int S_{Tx}(t) S_{Tx}(t-\phi) dt}{\sqrt{S_{Tx}^2(t)} \sqrt{S_{Rx}^2(t)}} \right] \quad (7)$$

Fig. 7(c) shows the graph of group delay which is calculated by Equation 6. The group delay in general is used to study the time response between two antenna systems and is the measure of the delay in received signal concerning phase distortion.

This ensures that the received signal shape of the transmitted pulse is preserved without much pulse distortion. The group delay as observed in Fig. 7(c) notes the variation between  $\pm 0.1\text{ns}$  which satisfies the ideal condition. Fig. 7(d) shows the impulse response of the MIMO-antenna in both orientations. It can be observed that the face-to-face and side-to-side received pulses overlap without distortion in comparison to the input impulse. However, there are additional harmonics with very low amplitude that can be easily suppressed at the receiver. Also, the received signal strength (mV) can be easily increased by the integration of a low-noise amplifier.





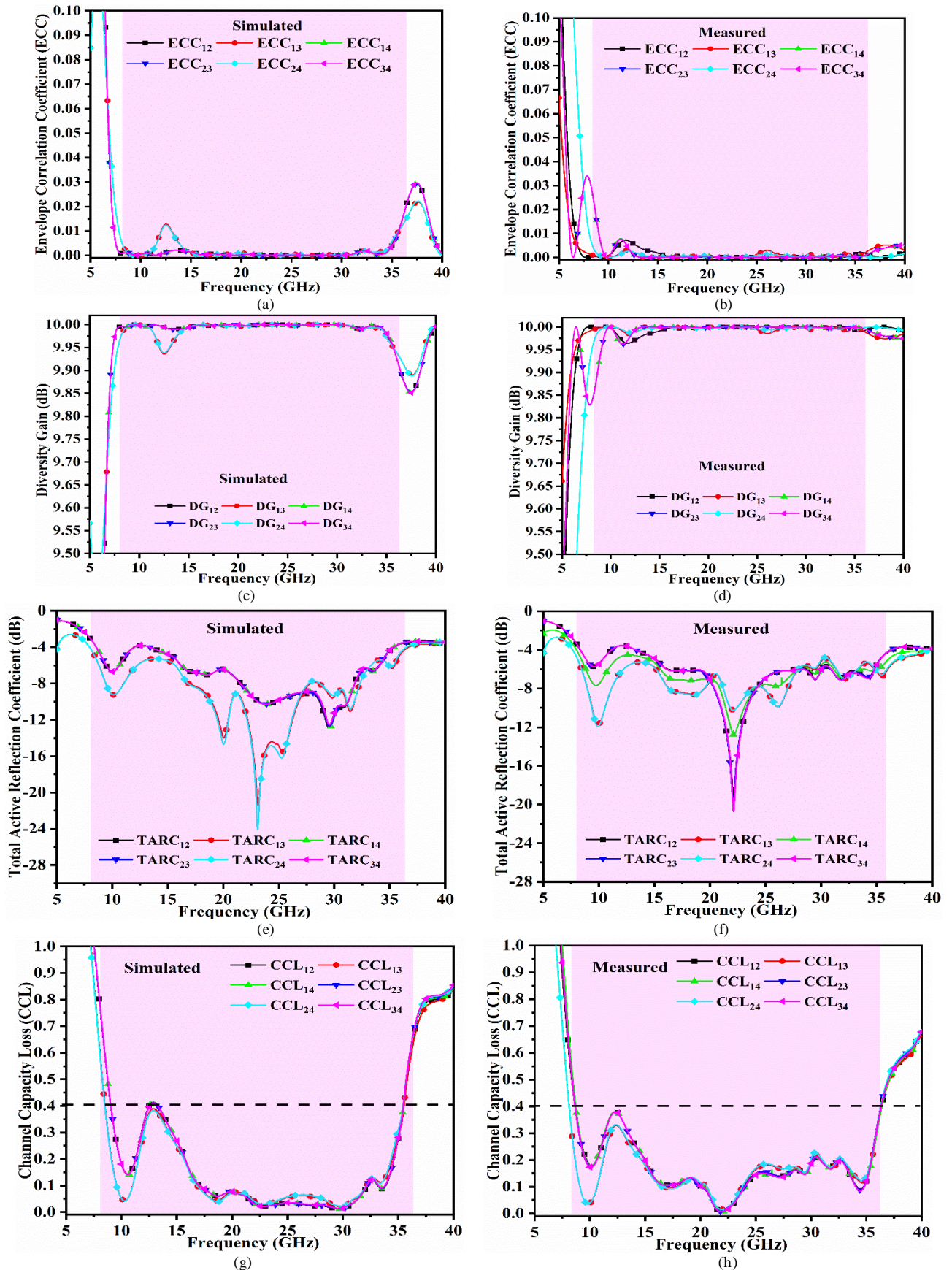


**Fig. 8.** Four port MIMO configuration (a)-(b) Perspective and transparent front view with printed radiating patch and ground in the simulation environment (b) Fabricated prototype (d) Simulated and Measured reflection coefficient (e)-(f) Simulated and Measured transmission coefficient.

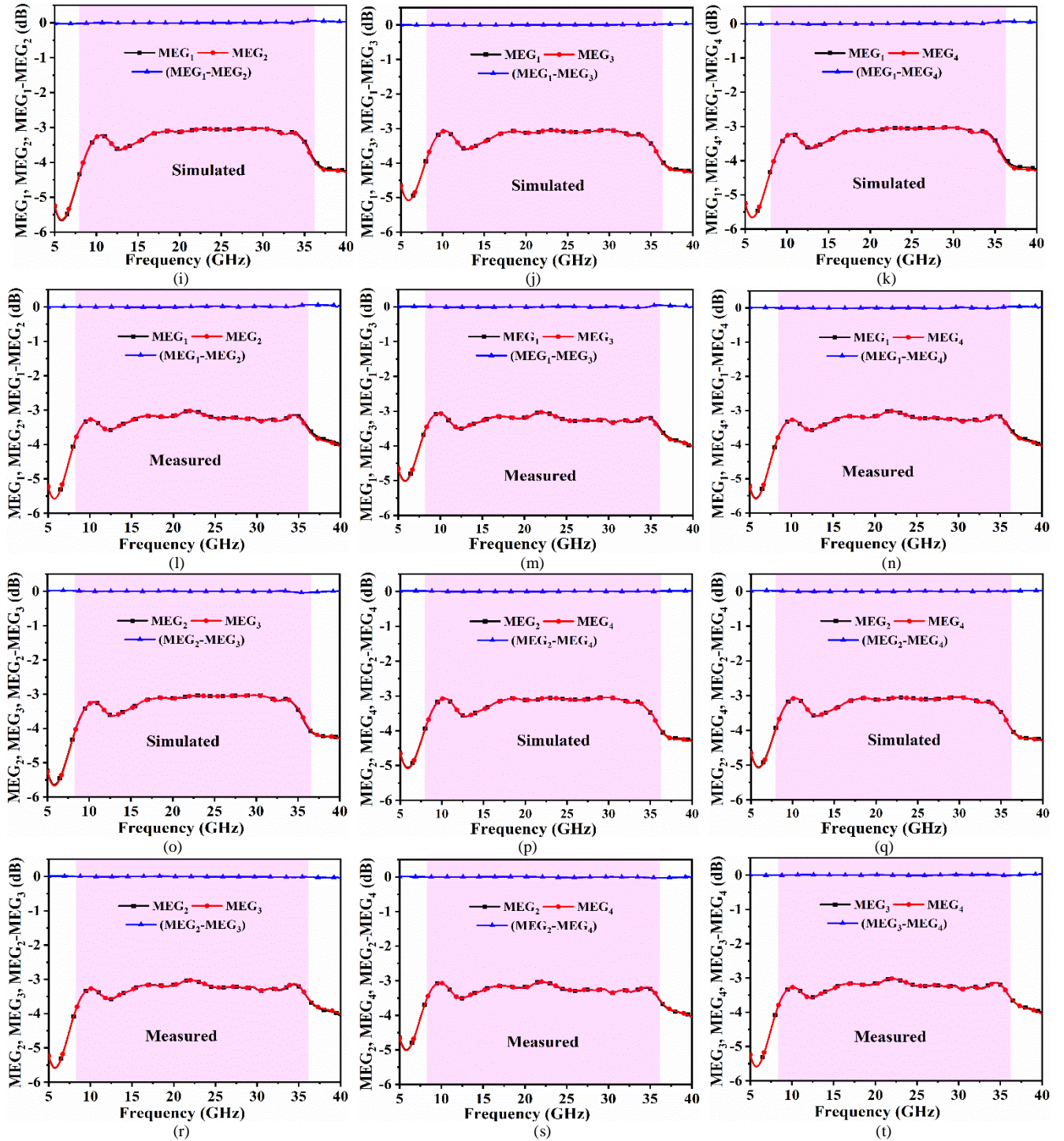
The advantage of transforming single-port to two-port MIMO configuration with common connected ground is studied which reduces the multi-path fading. The increase in the number of radiating elements in the MIMO configuration from 2-Port to 4-Port will further enhance the reduction in fading of the signal. Fig. 8 shows the four-port configuration of the MIMO antenna. Fig. 8(a)-(c) shows the Isometric and transparent front view of the proposed work where all the radiating antennae are connected to modeled SMK2.92mm connector for signal input. The radiating elements are aligned orthogonally ( $90^\circ$  with each other) and spacing of  $K=9.00\text{mm}$  between them to ensure proper isolation. Fig. 8(c) shows the developed prototype by a conventional method with fine-quality substrate and connectors to achieve accurate results. Fig. 8(d) records the simulated and measured reflection coefficients with a shaded portion indicating the operating bandwidth. The simulated bandwidth corresponds to 7.92GHz-36.50GHz and the measured bandwidth is 8.31GHz to 36.14GHz. Fig. 8(e)-(f) corresponds to the simulated and measured transmission coefficient with isolation of more than 12.50dB in a simulation environment and more than 15.0dB for measured values. The simulation environment provides exact  $50\Omega$  matching between the port and the feedline. Also, the finite-element-method (FEM) do provide good approximation calculation of maxwell's equations. Whereas, in measurement conditions, the quality of SMK connector used, quality of the substrate, precision in calibration of VNA with accuracy in measurement are the few factors which results in deviation of S-parameter in comparison of the simulation results.

#### IV. DIVERSITY PERFORMANCE AND FAR-FIELD RESULT DISCUSSION COMPARISON OF PRESENT STATE-OF-THE-ART MIMO ANTENNA









**Fig. 9.** Four port MIMO configuration with simulated and measured diversity parameters (a)-(b) Simulated and measured  $ECC_{2-Port}$  (c)-(d) Simulated and measured  $DG_{2-Port}$  (e)-(f) Simulated and measured  $TARC_{2-Port}$  (g)-(h) Simulated and measured  $CCL_{2-Port}$  (i)-(k) Simulated  $MEG_{12}$ ,  $MEG_{13}$ ,  $MEG_{14}$  (l)-(n) Measured  $MEG_{12}$ ,  $MEG_{13}$ ,  $MEG_{14}$  (o)-(q) Simulated  $MEG_{23}$ ,  $MEG_{24}$ ,  $MEG_{34}$  (l)-(n) Measured  $MEG_{23}$ ,  $MEG_{24}$ ,  $MEG_{34}$ .

The single-port multiband antenna in the proposed work with conformal capability exhibits good far-field results and occupies multi-band bandwidth which finds applications in several wireless applications. However,

when the radiating patch is increased in number say  $d \times k$  with  $d=1,2,3,4$  and  $k=1,2,3,4$  offers the following s-matrix with 16-S-parameters given below [41]

$$[S] = \begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{21} & S_{22} & S_{23} & S_{24} \\ S_{31} & S_{32} & S_{33} & S_{34} \\ S_{41} & S_{42} & S_{43} & S_{44} \end{bmatrix} \quad (8)$$

The S-parameter shown in the above matrix corresponds to reflection coefficients ( $S_{11}$ ,  $S_{22}$ ,  $S_{33}$ ,  $S_{44}$ ) and transmission coefficients ( $S_{12}$ ,  $S_{13}$ ,  $S_{14}$ ,  $S_{21}$ ,  $S_{23}$ ,  $S_{24}$ ,  $S_{31}$ ,  $S_{32}$ ,  $S_{34}$ ,  $S_{41}$ ,  $S_{42}$ ,  $S_{43}$ ). The above S-parameters need to maintain the required bandwidth ( $S_{aa}$ ;  $a=1,2,3,4$ ) and minimum required isolation ( $S_{ab}$ ;  $a=1,2,3,4$ ;  $b=1,2,3,4$ ). When the input signal is fed to the antenna. Due to the property of radiating os signals, the inter-radiating signals need to be isolated from each other to avoid destructive interference, and hence, the parameter related to diversity performance known as Envelope-Correlation-Coefficient<sub>MB</sub> ( $ECC_{MB}$ ; MB-multiband) needs to be calculated. The modulus values of  $ECC_{MB}$  vary between 0 and 1 with 0 indicating nill interference of radiation between inter-spaced elements and 1 with higher interference. The  $ECC_{MB}$  for a four-port MIMO configuration is evaluated either by radiation pattern or by extracted S-parameters. Equation 9-Equation 14 shows the calculation of  $ECC_{MB}$  by using the radiation pattern method considering the  $m^{th}$  and  $s^{th}$  port in the MIMO configuration [41].

$$\gamma_c = \frac{\int_0^{2\pi} \int_0^\pi ((XP_{RG_{\theta m}}(\theta, \phi) E_{\theta s}^*(\theta, \phi) P_{\theta}(\theta, \phi) + E_{\phi m}(\theta, \phi) E_{\phi s}^*(\theta, \phi) P_{\phi}(\theta, \phi)) \sin \theta d\theta d\phi}{\sqrt{\gamma_m^2 \gamma_s^2}} \quad (9)$$

Where  $\gamma_m^2$ ,  $\gamma_s^2$  signifies the variance of corresponding ports and mathematically is written by following sets of equations Mathematically [41]

$$\int_0^{2\pi} \int_0^\pi ((XP_{RG_{\theta m}}(\theta, \phi) P_{\theta}(\theta, \phi) + G_{\phi m}(\theta, \phi) P_{\phi}(\theta, \phi)) \quad (10)$$

$$G_{\theta m}(\theta, \phi) = E_{\theta m}(\theta, \phi) E_{\theta s}^*(\theta, \phi) \quad (11)$$

$$G_{\phi m}(\theta, \phi) = E_{\phi m}(\theta, \phi) E_{\phi s}^*(\theta, \phi) \quad (12)$$

$$G_{\theta s}(\theta, \phi) = E_{\theta s}(\theta, \phi) E_{\theta m}^*(\theta, \phi) \quad (13)$$

$$G_{\phi s}(\theta, \phi) = E_{\phi s}(\theta, \phi) E_{\phi m}^*(\theta, \phi) \quad (14)$$

where  $E_{\theta m}$ ,  $E_{\phi m}$ ,  $E_{\theta s}$  and  $E_{\phi s}$  are complex-electric fields with  $(\theta, \phi)$  for the  $m^{th}$  and  $s^{th}$  antenna respectively. The ECC values are real values which are plotted in Figure 9(a)-(b) showing the amplitudes of the signals at antennas and in the Rayleigh fading channel, the amplitude of the envelope correlation coefficient is given by [41]

$$\rho_e = |\rho_c|^2 \quad (15)$$

The ECC can be calculated from 3-D radiation and S-parameters. Assuming a uniform multipath environment the  $ECC_{MB}$  is evaluated from the following Equations using S-parameters [41]

$$ECC = \rho_e(m, s, N) = \frac{|\sum_{n=1}^N S_{m,n}^* S_{n,s}|^2}{\pi_{k=(m,s)} [1 - \sum_{n=1}^N S_{m,n}^* S_{n,k}]} \quad (16)$$

Where  $\rho_e(m, s, N)$  is the  $ECC_{MS}$  between the  $m^{th}$  and  $s^{th}$  port of the N-Element system. The  $ECC_{MB}$  for any two-port MIMO network is given by [41]

$$ECC_{MB} = \frac{|S_{mm}^* S_{ms} + S_{sm}^* S_{ss}|^2}{(1 - |S_{ii}|^2 - |S_{sm}|^2)(1 - |S_{ss}|^2 - |S_{ms}|^2)} \quad (17)$$

And for two-port and four-port,  $ECC_{MB}$  is given by

$$ECC_{MB(Two Port)} = \frac{|S_{11}^* S_{12} + S_{12}^* S_{22}|^2}{(1 - |S_{11}|^2 - |S_{21}|^2)(1 - |S_{12}|^2 - |S_{22}|^2)} \quad (18)$$

$$ECC_{MB(Four Port)} = \frac{|S_{11}^* S_{12} + S_{12}^* S_{22} + S_{13}^* S_{32} + S_{14}^* S_{42}|^2}{(1 - |S_{11}|^2 - |S_{21}|^2 - |S_{31}|^2 - |S_{41}|^2)(1 - |S_{12}|^2 - |S_{22}|^2 - |S_{32}|^2 - |S_{42}|^2)} \quad (19)$$

The ideal value of  $ECC_{MB}$  is 0 indicating the independent working of the antenna radiating elements and the standard or acceptable values are  $ECC_{MB} \leq 0.50$ . The exact values of ECC are evaluated using far-field patterns but due to resource limitations, the  $ECC_{MB}$  in the proposed work is evaluated by Equation (18) for two-port and Equation 19 for four-port. The  $ECC_{MB}$  (m,s) MIMO configuration with  $m=1,2,3,4$  and  $s=1,2,3,4$ ,  $ECC_{12}$ ,  $ECC_{13}$ ,  $ECC_{14}$ ,  $ECC_{23}$ ,  $ECC_{24}$ , and  $ECC_{34}$  are calculated for both simulated and measured as shown in Fig. 9(a)-(b). The simulated values of  $ECC_{MB}$  are below 0.40 and for measured these values fall below 0.30 satisfying well the ideal value condition  $ECC_{MB} \leq 0.50$ . Equation 18 & Equation 19 fulfills the requirement of evaluation of  $ECC_{MB}$  by assuming the antennas have higher efficiencies and no mutual-coupling losses, uniform multipath environment and the load termination of the antenna is 50Ω.

The Diversity-Gain<sub>MB</sub> ( $DG_{MB}$ ) is calculated to quantify the performance characteristics of the diversity scheme used (spatial, polarization, or radiation diversity). The  $DG_{MB}$  is related to  $ECC_{MB}$  by the following formula [41]

$$DG_{MB} = 10\sqrt{1 - |\rho_e|^2} \quad (20)$$

The  $DG_{MB}$  values as per the standard values are  $DG_{MB} \geq 9.95dB$ . Fig. 9(c)-(d) shows the calculated values of  $DG_{MB}$  for both simulated & measured s-parameters with values corresponding to more than 9.95dB and 9.9999dB respectively.

The radiation efficiencies and the operating bandwidth of the MIMO<sub>MB</sub> cannot be accurately calculated by S-parameters (reflection and transmission coefficients). Thus, the coupling and the random-signal combination are calculated by Total Active Reflection Coefficient<sub>MB</sub> ( $TARC_{MB}$ ) which gives a more meaningful sense to MIMO efficiency. The input signal to all the ports (4-ports in the proposed work) is the available power, transferred power is the radiation power and the difference between the two is the reflected power and thus, the square root ratio between the reflected power to that of the incident power is defined as  $TARC_{MB}$  which is given by following Equations [41]

$$\Gamma_a^t = \frac{\text{Available Power (AP)} - \text{Radiated Power (RP)}}{\text{Available Power (AP)}} \quad (21)$$

For a lossless MIMO system with an N-number of elements or N-port, the  $TARC_{MB}$  is given by

$$\Gamma_a^t = \frac{\sqrt{\sum_{i=1}^N |b_i|^2}}{\sqrt{\sum_{i=1}^N |a_i|^2}} \quad (22)$$

Where  $[b]=[s][a]$ ;  $a$  is the incident power with random phase and  $b$  is the reflected power.

In the propagation channel, the reflected signals are randomly phased because MIMO channels are assumed to be Gaussian and multipath spread propagation channels [41].

$$\begin{aligned} b_1 &= S_{11}a_1 + S_{12}a_2 = S_{11}a_0e^{j\theta_1} + S_{12}a_0e^{j\theta_2} = \\ a_1(S_{11} + S_{12}e^{j\theta}) \end{aligned} \quad (23)$$



$$b_2 = S_{21}a_1 + S_{22}a_2 = S_{21}a_0e^{j\theta_1} + S_{22}a_0e^{j\theta_2} = a_1(S_{21} + S_{22}e^{j\theta}) \quad (24)$$

Combining Equations,  $TARC_{MB}$

$$\Gamma_a^t = \frac{\sqrt{|S_{11}+S_{12}e^{j\theta}|^2 + |S_{21}+S_{22}e^{j\theta}|^2}}{\sqrt{2}} \quad (25)$$

Assuming  $\theta=0^\circ$ ,  $TARC_{MB}$  is evaluated from Equation 25 and is plotted in Fig. 9(e)-(f) for both simulated and measured S-parameters. The ideal values  $TARC \leq 0dB$  and for the simulated and measured  $TARC_{11}$ ,  $TARC_{12}$ ,  $TARC_{13}$ ,  $TARC_{23}$ ,  $TARC_{24}$ , and  $TARC_{34}$ , the values are below  $-4.0dB$ .

The maximum transmission rate of the information over the channel with maximum limit is defined by Channel-Capacity-Loss<sub>MB</sub> ( $CCL_{MB}$ ) with  $CCL_{MB} \leq 0.40b/s/Hz$ . The  $CCL_{MB}$  for a four-port MIMO antenna in generalized form is calculated by [41]

$$CCL_{60.0GHz} = -\log_2 \det(\alpha^s) \quad (26)$$

where

$$\rho_{mm} = 1 - \sum_{n=1}^4 |S_{mn}|^2 \quad (27)$$

$$\rho_{ms} = -(S_{mm}^* S_{ms} + S_{sm}^* S_{ms}) \quad (28)$$

Fig 9. (g)-(h) shows the calculation of  $CCL_{60.0GHz}$  for both simulated and measured reflection and transmission coefficients. The simulated averaged values  $CCL_{MB} \leq 0.1b/s/Hz$  and measured averaged values correspond to  $CCL_{MB} \leq 0.12b/s/Hz$ .

The Mean-Effective-Gain<sub>MB</sub> ( $MEG_{MB}$ ) is defined as the ratio of the received power to the power which will be received by an isotropic antenna when replaced [41].

In general, the  $MEG_{ms}$  are calculated by

$$MEG_{ms} = 0.5[1 - \sum_{s=1}^K |S_{ms}|^2] \quad (29)$$

$$MEG_m = 0.5 \times (1 - |S_{mm}|^2 - |S_{ms}|^2) \quad (30)$$

$$MEG_s = 0.5 \times (1 - |S_{ss}|^2 - |S_{sm}|^2) \quad (31)$$

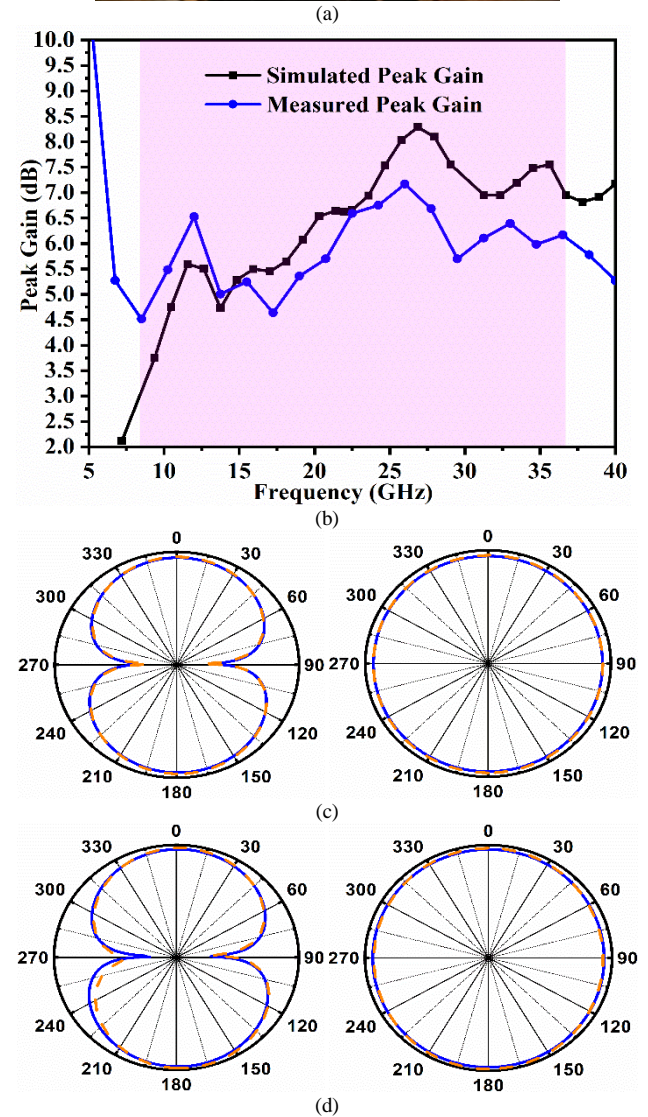
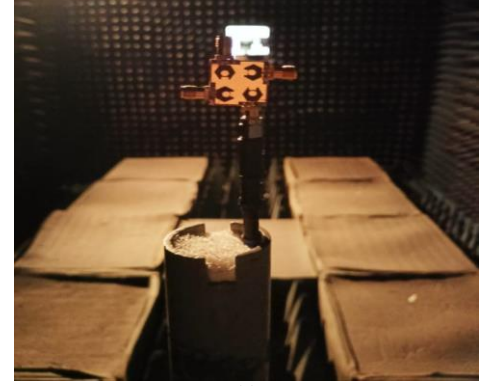
The ratio calculates the  $MEG_{60.0GHz}$  which is given by

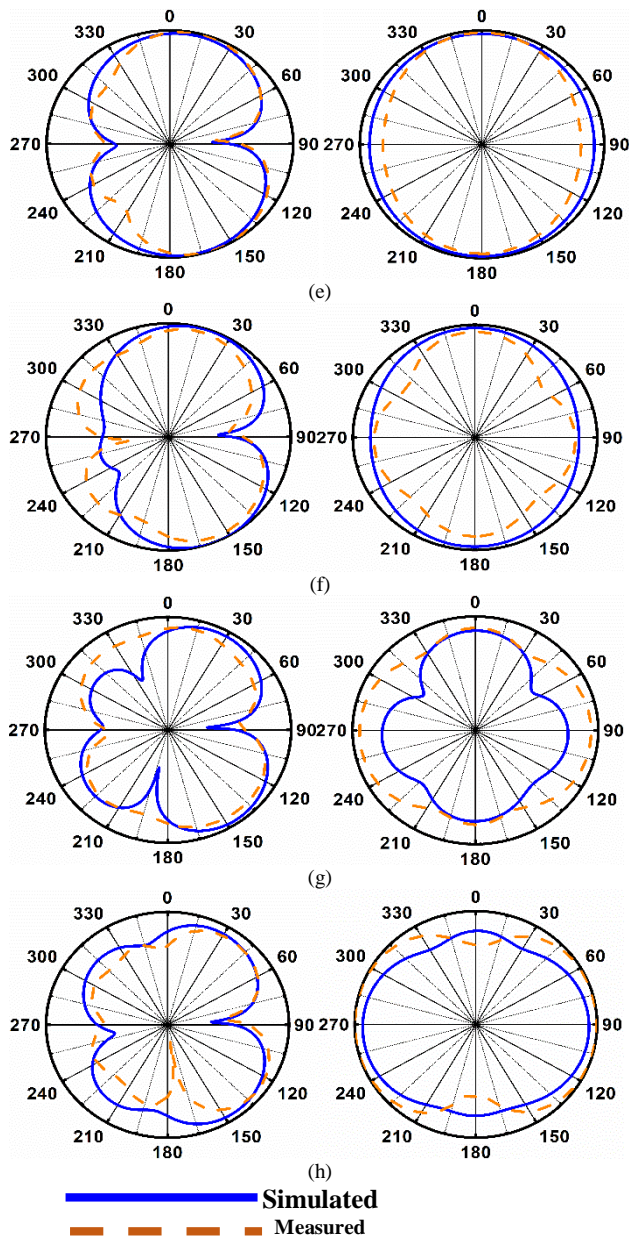
$$\frac{MEG_m}{MEG_n} = \frac{0.5 \times (1 - |S_{mm}|^2 - |S_{mn}|^2)}{0.5 \times (1 - |S_{nn}|^2 - |S_{nm}|^2)} \quad (32)$$

For the  $MEG_{port-m}$  and  $MEG_{port-n}$  are evaluated from Equations (30) and Equation (31) with Equation (32) being the standard equation for MEG. Mean-Effective-Gain<sub>60.0GHz</sub> ( $MEG_{MB}$ ) are evaluated between two-port as shown in Fig. 9(i)-(t). Fig. 9(i)-(k) corresponds to simulated  $MEG_{MB}$  for port1-port2, port1-port3 & port1-port4 and Fig. 9(l)-(n) corresponds to measured  $MEG_{MB}$  for port1-port2, port1-port3 & port1-port4. In both cases, the  $MEG_{MB}$  are nearly grazing  $-3.0dB$  values while the difference is nearly equal to  $0.0dB$ . Similarly,  $MEG_{MB}$  for port2-port3, port2-port4 & port3-port4 also corresponds to  $-3.0dB$  in both simulated- Fig. 9(o)-(q) and measured results shown in Fig. 9(r)-(t).

Fig. 10 shows the simulated and measured peak gain & 2D-radiation pattern in principal planes. The peak gain is plotted in Fig. 10(a) which compares simulated and measured values for the operating bandwidth of 7.92GHz-36.50GHz and varies between 3.0dBi-7.02dBi with the highest peak corresponding to 8.28dBi at 26.875GHz for simulation and 4.50dBi-6.16dBi in measured environment with highest peak gain of 7.17dBi at 26.0GHz. For lower frequency points at 11.0GHz, 15.0GHz, and 20.0GHz, the

2D-radiation pattern corresponds to the desired dipole-omnidirectional pattern in both simulation-measured environments. However, at higher frequency points corresponding to 24.0GHz, 28.0GHz, and 30.0GHz, the 2D-radiation pattern does deviate from the ideal desired patterns but, is capable of transmitting and receiving the signals effectively when deployed in practical applications with wider-radiation patterns.





**Fig. 10.** (a) Photograph of proposed MIMO antenna placed within ANECHOIC CHAMBER (b) Simulated & Measured peak gain; simulated and measured 2-D radiation pattern at (c) 11.0GHz (d) 15.0GHz (e) 20.0GHz (f) 24.0GHz (g) 28.0GHz (h) 30.0GHz.

## V. CONFORMAL CAPABILITY OF THE PROPOSED MIMO ANTENNA WITH SAR CALCULATION AND PRESENT STATE-OF-THE-ART MIMO ANTENNA COMPARISON

The proposed work utilizes a very thin substrate of thickness 0.254mm and hence, can be used in conformal applications. Due to the very compact size of the proposed MIMO antenna, it can be easily embedded within the devices designed for on-body applications with SAR within the specified limit.

The modeling of human tissue is utilized in calculating the Specific-Absorption-Rate (SAR) which signifies the amount of power absorbed by the human tissue when

subjected to the interference of electromagnetic waves. The phantom layered of the human tissue consists of skin, fat, and muscle which represent the outermost layer of the body. This modeled human tissue is utilized to interact with the electromagnetic waves which are transmitted from the proposed MIMO antenna. The electrical properties which are already calculated [33] are assigned to all the layers of tissue for SAR analysis. The proposed MIMO antenna is placed on the assembled tissue layer with a gap of 5.00mm. The electrical parameters are tabulated in Table 2 given below

**Table2:** Electrical property of phantom tissue

Tissue	Permittivity ( $\epsilon_r$ )	Conductivity (S/m)	Loss Tangent ( $\tan\delta$ )
Skin	38.8	1.18	0.30
Fat	5.30	0.07	0.14
Muscle	53.5	1.34	0.25

Fig. 11(a) shows the capability of the proposed four-port MIMO antenna with measured  $S_{11}$  parameter and corresponds to without and with conformal capability bent at  $45^\circ$ . From the observations, the  $S_{11}$  measures the bandwidth of 8.31GHz-36.14GHz without any bending and with bedding of  $45^\circ$  offers an impedance bandwidth of 9.05GHz-37.90GHz.

**Table 3:** SAR values at different frequency points

Frequency (GHz)	Maximum SAR Value (W/Kg)	Maximum SAR (W/Kg)	Operating Bands (GHz)
11.0	0.366	<1.60	X-band
15.0	0.313		Ku band
20.0	0.424		K-band
24.0	0.418		5G-mmWave FR2
28.0	0.377		K-band
30.0	0.309		5G-mmWave FR2
			Ka-band

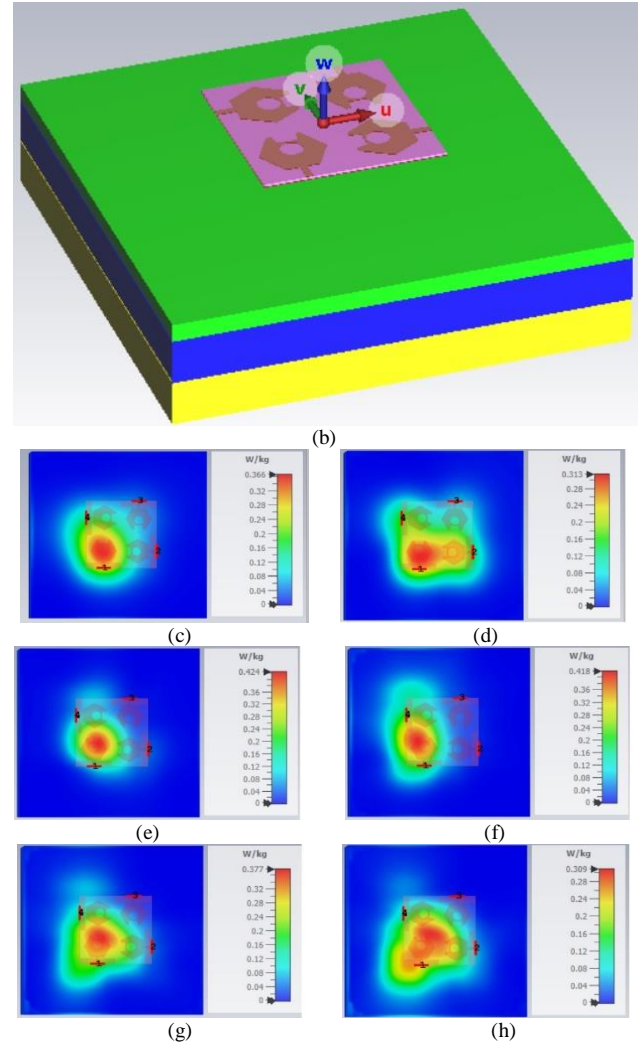
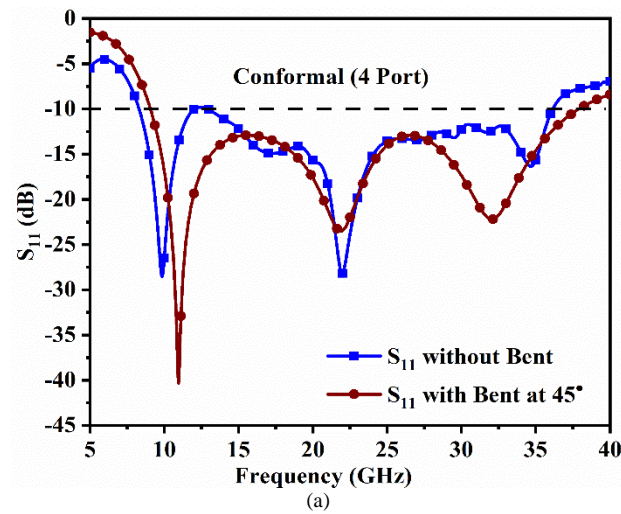
Fig. 11(b) shows the 3-D model of phantom tissue which is placed beneath the proposed MIMO configuration for SAR analysis and with a power input of 50 mW in CST-Microwave Studio EM-simulator. The three layers correspond to skin, fat, and muscle which forms the human tissue for the interaction of EM-wave radiated by the antenna to obtain SAR values for different frequency points specified within the operating bandwidth. The practical power absorbed by human tissue is defined by SAR when interacting with EM waves and is evaluated by [35]

$$SAR = \int \frac{\sigma(r)|E(r)|^2}{\rho(r)} dr \quad (33)$$

Where  $\sigma$ -tissur conductivity (S/m), E is the electric field intensity (V/m) and  $\rho$  corresponds to tissue mass density (Kg/m<sup>3</sup>). For 1g of the tissue, the average SAR value must

be less than 1.60 W/Kg [34]. Fig. 11(c)-(h) corresponds to the calculation of SAR for a wide range of frequencies and is given in Table 3

Table 4 shows a brief comparison of the proposed four-port MIMO antenna with the present state-of-the-art. It can be noted that the proposed MIMO antenna offers wider impedance bandwidth fabricated on a very thin substrate which is utilized for conformal applications. Also, for on-body applications, the antenna is analyzed for SAR applications which are calculated at different frequency points and include different band applications. All the above features outclass the other published work which are compared in Table 4.



**Fig. 11.** (a)  $S_{11}$  comparison of proposed MIMO antenna without bent and with bent at  $45^\circ$  (b) 3-D model of phantom tissue placed below radiating antenna; SAR at (c) 11.0GHz (d) 15.0GHz (e) 20.0GHz (f) 24.0GHz (g) 28.0GHz (h) 30.0GHz.

**Table 4:** State-of-the-art comparison

Ref.	Size (mm <sup>2</sup> )	Bandwidth (GHz)/%ge BW	Maximum Peak Gain (dBi)	No. of Rad. Elements	ECC	DG (dB)	TARC/ Isolation (dB)	CCL (b/s/H z)	MEG (dB)	SAR Value (W/Kg)	Conformal Capability	Potential Applications
[1]	$0.42\lambda_0 \times 0.71\lambda_0$	25.83-30.24 15.73%	4.00	1	NA	NA	NA	NA	NA	NA	No	mmWave
[2]	$2.14\lambda_0 \times 2.72\lambda_0$	26.52-29.50 16.22%	11.50	Array 1×4	NA	NA	NA	NA	NA	NA	No	mmWave
[4]	$3.00\lambda_0 \times 2.21\lambda_0$	45.0-62.5 28.0%	6.65	Array 2×2	NA	NA	NA	NA	NA	NA	No	mmWave
[6]	$0.15\lambda_0 \times 0.29\lambda_0$	2.26-2.73 13.20% 4.27-5.38 11.47%	7.18	1	NA	NA	NA	NA	NA	NA	No	ISM
[8]	$0.23\lambda_0 \times 0.39\lambda_0$	2.11-4.19 49.64% 4.98-6.81 26.87%	4.19	2	<0.004	>9.97	<-10.0 >21.0	<0.32	NA	NA	No	LTE Wi-Fi WLAN Bluetooth



												Wi-Max
[10]	$0.73\lambda_0 \times 1.21\lambda_0$	3.10-20.0 84.5%	NC	2	<0.15	>9.6 4	NC >20.0	NC	NC	NA	No	UWB X-band Ku-band
[17]	$0.47\lambda_0 \times 0.47\lambda_0$	2.97-15.48 74.15%	5.11	4	<0.20	>9.9 2	<-20.0 >16.0	<0.32	NC	NA	No	UWB X-band Ku-band
[18]	$1.20\lambda_0 \times 1.20\lambda_0$	3.18-11.50 72.35%	5.82	4	<0.00 5	>9.9 85	NC >16.0	<0.25	NC	NA	No	UWB X-band
[19]	$0.66\lambda_0 \times 0.66\lambda_0$	2.96-11.40 74.0%	4.85	4	<0.03	>9.9 82	<-12.0 >20.0	NC	NC	NA	No	UWB X-band
[20]	$0.21\lambda_0 \times 0.21\lambda_0$	1.15-40.0 97.13%	5.02	4	<0.00 5	NC	<-15.0 >25.0	NC	NC	NA	No	Bluetooth UWB X-band Ku-band K-band Ka-band
[21]	$3.89\lambda_0 \times 4.54\lambda_0$	25.5-29.6 13.85%	8.30	4	<0.02	>9.9 78	NC >30.0	<0.32	NC	NA	No	mmWave
[29]	$4.86\lambda_0 \times 6.11\lambda_0$	58.50	17.20	1	NA	NA	NA	NA	NA	NA	Yes	mmWave
[31]	$0.26\lambda_0 \times 0.29\lambda_0$	2.45 18.46%	0.50	2	<0.12	>9.8 5	NC >30.0	<0.08	NA	0.512	Yes	ISM
*P	$0.67\lambda_0 \times 0.67\lambda_0$	7.92-36.50 77.0%	7.17	4	<0.01	>9.9 96	<-4.0 >10.0	<0.38	$\cong 0.0$	0.366 at 11GHz 0.313 at 15GHz 0.424 at 20GHz 0.418 at 24GHz 0.377 at 28GHz 0.309 at 30GHz	Yes	X-band Ku-band K-band Ka-band FR2- mmWave

\*P – Proposed work; NA-Not Applicable, NC-Not Calculated

## V. CONCLUSION

A four-port MIMO antenna with conformal capability and SAR analysis is presented. The proposed MIMO antenna occupies wider impedance bandwidth which includes applications in Microwave-Millimeter wave bands including X-band, Ku-band, K-band, partial Ka-band, n257, n258, and n261 bands. The proposed MIMO antenna offers a measured averaged peak gain of 5.50dBi and 2D-dipole-omnidirectional patterns. The MIMO antenna is also evaluated with group delay and impulse response. The diversity parameters are well within the permissible values. The SAR analysis of the MIMO antenna is less than 1.60W/Kg in the frequency values within the operating bandwidth.

## References

- [1] P. Kumar *et al.*, "An Ultra-Compact 28 GHz Arc-Shaped Millimeter-Wave Antenna for 5G Application," *Micromachines (Basel)*, vol. 14, no. 1, Dec 20 2022.
- [2] S. H. Kiani *et al.*, "Square-Framed T Shape mmwave Antenna Array at 28 GHz for Future 5G Devices," *International Journal of Antennas and Propagation*, vol. 2021, pp. 1-9, 2021.
- [3] M. H. Sharaf, A. I. Zaki, R. K. Hamad, and M. M. M. Omar, "A Novel Dual-Band (38/60 GHz) Patch Antenna for 5G Mobile Handsets," *Sensors (Basel)*, vol. 20, no. 9, Apr 29 2020.
- [4] A. G. Alharbi *et al.*, "Design and Study of a Miniaturized Millimeter Wave Array Antenna for Wireless Body Area Network," *International Journal of Antennas and Propagation*, vol. 2022, pp. 1-25, 2022.
- [5] A. Dejen, J. Jayasinghe, M. Ridwan, and J. Anguera, "Synthesis of Quadband mm-Wave Microstrip Antenna Using Genetic Algorithm for Wireless Application," *Technologies*, vol. 11, no. 1, 2023.
- [6] A. Gupta, A. Kansal, and P. Chawla, "Design of a Compact Dual-Band Antenna for On-/Off Body Communication," *IETE Journal of Research*, vol. 69, no. 2, pp. 1013-1021, 2020.
- [7] M. Sharma, "Superwideband Triple Notch Monopole Antenna for Multiple Wireless Applications," *Wireless Personal Communications*, vol. 104, no. 1, pp. 459-470, 2018.
- [8] R. N. Tiwari, P. Singh, B. K. Kanaujia, S. Kumar, and S. K. Gupta, "A low profile dual band MIMO antenna for LTE/Bluetooth/Wi-Fi/WLAN applications," *Journal of Electromagnetic Waves and Applications*, vol. 34, no. 9, pp. 1239-1253, 2020.
- [9] H. Ullah, H. F. Abutarboush, A. Rashid, and F. A. Tahir, "A Compact Low-Profile Antenna for Millimeter-Wave 5G Mobile Phones," *Electronics*, vol. 11, no. 19, 2022.
- [10] B. T. Ahmed, P. S. Olivares, J. L. M. Campos, and F. M. Vázquez, "(3.1–20) GHz MIMO antennas," *AEU - International Journal of Electronics and Communications*, vol. 94, pp. 348-358, 2018.
- [11] J. Khan, S. Ullah, U. Ali, F. A. Tahir, I. Peter, and L. Matekovits, "Design of a Millimeter-Wave MIMO Antenna Array for 5G Communication Terminals," *Sensors (Basel)*, vol. 22, no. 7, Apr 4 2022.
- [12] H. Ekrami and S. Jam, "A compact triple-band dual-element IMO antenna with high port-to-port isolation for wireless applications," *AEU - International Journal of Electronics and Communications*, vol. 96, pp. 219-227, 2018.
- [13] Z. Li, C. Yin, and X. Zhu, "Compact UWB MIMO Vivaldi Antenna With Dual Band-Notched Characteristics," *IEEE Access*, vol. 7, pp. 38696-38701, 2019.
- [14] Z. Tang, J. Zhan, X. Wu, Z. Xi, L. Chen, and S. Hu, "Design of a compact UWB-MIMO antenna with high isolation and dual band-notched characteristics," *Journal of Electromagnetic Waves and Applications*, vol. 34, no. 4, pp. 500-513, 2020.
- [15] S. Tariq, S. I. Naqvi, N. Hussain, Y. Amin, "A Metasurface-Based MIMO Antenna for 5G Millimeter-Wave Applications," *IEEE Access*, vol. 9, pp. 51805-51817, 2021.
- [16] T. Addepalli, T. Vidyavathi, K. Neelima, M. Sharma, and D. Kumar, "Asymmetrical fed Calendula flower-shaped four-port 5G-NR band (n77, n78, and n79) MIMO antenna with high diversity



- performance," *International Journal of Microwave and Wireless Technologies*, pp. 1-15, 2022.
- [17] M. Sharma, V. Janghu, and N. Kumar, "Dual Notched Four-Port Multiband Reconfigurable MIMO Antenna with Novel Fork-Radiator and  $\Omega$ -shaped Ground," *IETE Journal of Research*, pp. 1-14, 2022.
- [18] M. N. Hasan, S. Chu, and S. Bashir, "A DGS monopole antenna loaded with U-shape stub for UWB MIMO applications," *Microwave and Optical Technology Letters*, vol. 61, no. 9, pp. 2141-2149, 2019.
- [19] A. McHbal, N. Amar Touhami, H. Elftouh, and A. Dkiouak, "Coupling reduction using a novel circular ripple-shaped decoupling mechanism in a four-element UWB MIMO antenna design," *Journal of Electromagnetic Waves and Applications*, vol. 34, no. 12, pp. 1647-1666, 2020.
- [20] D. Kumar Raheja, S. Kumar, B. Kumar Kanaujia, S. Kumar Palaniswamy, R. Rao Thipparaju, and M. Kanagasabai, "Truncated elliptical Self-Complementary antenna with Quad-Band notches for SWB MIMO systems," *AEU - International Journal of Electronics and Communications*, vol. 131, 2021.
- [21] M. Khalid *et al.*, "4-Port MIMO Antenna with Defected Ground Structure for 5G Millimeter Wave Applications," *Electronics*, vol. 9, no. 1, 2020.
- [22] K. Raheel *et al.*, "E-Shaped H-Slotted Dual Band mmWave Antenna for 5G Technology," *Electronics*, vol. 10, no. 9, 2021.
- [23] S. Juneja, R. Pratap, and R. Sharma, "Semiconductor technologies for 5G implementation at millimeter wave frequencies – Design challenges and current state of work," *Engineering Science and Technology, an International Journal*, vol. 24, no. 1, pp. 205-217, 2021.
- [24] E. Al Abbas, M. Ikram, A. T. Mobashsher, and A. Abbosh, "MIMO Antenna System for Multi-Band Millimeter-Wave 5G and Wideband 4G Mobile Communications," *IEEE Access*, vol. 7, pp. 181916-181923, 2019.
- [25] T. Addepalli, J. Babu Kamili, K. Kumar Bandi, A. Nella, and M. Sharma, "Lotus flower-shaped 4/8-element MIMO antenna for 5G n77 and n78 band applications," *Journal of Electromagnetic Waves and Applications*, vol. 36, no. 10, pp. 1404-1422, 2022.
- [26] T. Addepalli, M. Sharma, A. Nella, A. P. Ambalgi, and P. R. Kapula, "Experimental investigation of super-wideband 8-port Multiple-Input-Multiple-Output antenna with high isolation for future wireless applications including Internet of Things," *International Journal of Communication Systems*, 2022.
- [27] G. Kumar, and R. Kumar, "A survey on planar ultra-wideband antennas with band notch characteristics: Principle, design, and applications," *International Journal of Electronics and Communications*, vol. 109, pp. 76-98, 2019.
- [28] P. Gupta, L. Malviya, and S. V. Charhate, "5G multi-element/port antenna design for wireless applications: a review," *International Journal of Microwave and Wireless Technologies*, vol. 11, no. 9, pp. 918-938, 2019.
- [29] V. Semkin *et al.*, "Conformal antenna array for millimeter-wave communications: performance evaluation," *International Journal of Microwave and Wireless Technologies*, vol. 9, no. 1, pp. 241-247, 2015.
- [30] D. Negi, R. Khanna, and J. Kaur, "Design and performance analysis of a conformal CPW fed wideband antenna with Mu-Negative metamaterial for wearable applications," *International Journal of Microwave and Wireless Technologies*, vol. 11, no. 08, pp. 806-820, 2019.
- [31] A. Gupta, A. Kansal, and P. Chawla, "Design of a wearable MIMO antenna deployed with an inverted U-shaped ground stub for diversity performance enhancement," *International Journal of Microwave and Wireless Technologies*, vol. 13, no. 1, pp. 76-86, 2020.
- [32] A. Gupta, P. Chawla, A. Kansal, and K. Singh, "An Active and Low-cost Microwave Imaging System for Detection of Breast Cancer Using Back Scattered Signal," *Curr Med Imaging*, vol. 18, no. 5, pp. 460-475, 2022.
- [33] A. Gupta, A. Kansal, and P. Chawla, "Design of a patch antenna with square ring-shaped-coupled ground for on-/off body communication," *International Journal of Electronics*, vol. 106, no. 12, pp. 1814-1828, 2019.
- [34] A. Gupta, A. Kansal, and P. Chawla, "A survey and classification on applications of antenna in health care domain: data transmission, diagnosis and treatment," *Sādhanā*, vol. 46, no. 2, 2021.
- [35] V. Karthik and T. Rama Rao, "Investigations on SAR and Thermal Effects of a Body Wearable Microstrip Antenna," *Wireless Personal Communications*, vol. 96, no. 3, pp. 3385-3401, 2017.
- [36] A. Kumar, S. Sharma, N. Goyal, A. Singh, X. Cheng, and P. Singh, "Secure and energy-efficient smart building architecture with emerging technology IoT," *Computer Communications*, 176, pp. 207-217, 2021.
- [37] K. Liu *et al.*, "Investigation of Conformal MIMO Antenna for Implantable Devices Based on Theory of Characteristic Modes," in *IEEE Transactions on Antennas and Propagation*, vol. 70, no. 12, pp. 11324-11334, Dec. 2022, doi: 10.1109/TAP.2022.3209236.
- [38] G. Das, A. Sharma, R. K. Gangwar and M. S. Sharawi, "Performance Improvement of Multiband MIMO Dielectric Resonator Antenna System With a Partially Reflecting Surface," in *IEEE Antennas and Wireless Propagation Letters*, vol. 18, no. 10, pp. 2105-2109, Oct. 2019, doi: 10.1109/LAWP.2019.2938004.
- [39] G. Das, N. K. Sahu, A. Sharma, R. K. Gangwar and M. S. Sharawi, "FSS-Based Spatially Decoupled Back-to-Back Four-Port MIMO DRA With Multidirectional Pattern Diversity," in *IEEE Antennas and Wireless Propagation Letters*, vol. 18, no. 8, pp. 1552-1556, Aug. 2019, doi: 10.1109/LAWP.2019.2922276.
- [40] M. Sharma, Y.K. Awasthi, H. Singh, R. Kumar, S. Kumari, "Compact printed high rejection triple band-notch UWB antenna with multiple wireless applications," *Engineering Science and Technology, an International Journal*, vol. 19, pp. 1626-1634, 2016.
- [41] M. Sharma, P.C. Vashist, P.S. Ashtankar, and S.K. Mittal, "Compact  $2 \times 2/4 \times 4$  tapered microstrip feed MIMO antenna configurations for high-speed wireless applications with band stop filters," *International Journal of RF and Microwave Computer-Aided Engineering*, vol. 31, pp. 1-16, 2020.



**Dr. MANISH SHARMA** (M'19 – SM'23) received B.E. degree in Electronics and Communication Engineering from Mangalore University, Karnataka, India in 2000 and M.Tech degree from Visvesvaraya Technological University, Karnataka, India in 2007. He completed his Ph.D degree from the Department of Electronics Engineering, Banasthali University, Rajasthan, India in 2017. He is

currently working as Professor-Research in Chitkara University Research and Innovation Network (CURIN), Chitkara University, Punjab, India. His research interest includes computational electromagnetics, reconfigurable antennas, novel electromagnetic materials, dielectric resonator antennas, wideband/superwideband antennas, wideband/dual band/triple band microstrip antennas for wireless communication, smart and MIMO antennas systems, radio-frequency identification (RFID) antennas, antennas for healthcare, RF MEMS planar antenna on Si substrate, wireless networks, body area networks, meta surface based biosensors, Designing of Microstrip antennas using Machine Learning and Artificial network. He has published more than 100 research papers. Ganga Prasad Pandey, senior member of IEEE, received B. Tech degree in Electronics and Communication Engineering from KNIT Sultanpur, UP in year 2000 and M. E. from Delhi College of Engineering, Delhi India in 2004. He received PhD from Uttarakhand Technical University, Dehradun, UK, in 2015 in the field of Tunable microstrip antenna. With more than 21 years of teaching experience, currently he is heading the Department of Information and Communication Technology at Pandit Deendayal Energy University, Gandhinagar. He has published more than 40 paper in international Journals of repute and more than 20 papers in International conference. He has guided 4 PhD scholars in the domain of Antenna. His research interest include Energy harvesting, ME-dipole, active, Reconfigurable, frequency agile microstrip antennas and microwave/millimeter wave integrated circuits and devices. and granted with 8 patents. He is currently guiding 8 Ph.D students. He has also published 18 book chapters. He is also reviewer of IEEE Access, Journal of Electromagnetic Waves and Applications, AEU: International Journal of Electronics & Communication, International Journal of Communication Systems, International Journal of Microwave and Wireless Technologies, International Journal of RF & Microwave Computer-Aided Engineering.



**Dr. Prabhakara Rao Kapula** is a professor at the B V Raju Institute of Technology in Narsapur, Telangana, in the Department of Electronics and Communication Engineering. He has 24 years of teaching experience. In 2017, he graduated with his doctorate from Andhra University. In 1996, he graduated from Andhra University with a B.E. in Electronics and Communication Engineering, and

in 1999, he obtained an M.Tech. in Instrumentation and Control Systems from JNTU College of Engineering in Hyderabad. His research interests are in wireless communication, MIMO antennas, biosensors, and medical assistive technology. He is a Fellow of IETE and an active member of IEEE. In addition to being awarded one patent, he has more than 50 research publications published in various journals with SCI, SCOPUS and WoS indexes. He reviews articles for the International Journal of Wireless Information Networks and the IEEE Transactions on Signal Processing. He is now working on a DST-TIDE research project on a paraplegic patient's mobility assistance device.



**SHAILAJA ALAGRAMA** (Member, IEEE) received M.S., from Chicago State university in 2017. She worked with Accenture as Network Administrator. She also has more than 4 years of teaching experience at the level of an Assistant Professor in Indian University. Currently pursuing Ph.D. from University of the Cumberland's,

Kentucky USA. Her research area includes IOT, 5G, Wireless network, Body area Network and Machine Learning.



**Dr. KANHAIYA SHARMA**, member IEEE, received the M. Tech degree in Computer Science & Engineering(CSE) from the Jawaharlal Lal Technological University Hyderabad, India in 2011 and the Ph.D. degree in CSE from the Pandit Deendayal Energy University Gandhinagar Gujarat, India in 2021. From 2010 to 2017, he was Assistant Professor with Siddhant Group of Institution Sudumbare Pune Maharashtra, India.

From 2020 to 2021, he was an Assistant Professor with Computer Science and Engineering Department, Sandip University Nashik, Maharashtra, India. From August 2021 to July 2022 he was an Assistant Professor with Computer Science and Engineering Department, Anurag University Hyderabad, Telangana, India. He is currently an Assistant Professor with the Department of Computer Science & Engineering, Symbiosis Institute of Technology, Symbiosis International University Lavale Pune, Maharashtra, India. He has authored or co-authored several journals and IEEE proceeding publications. His current research interests include next-generation of mm-wave and development of low-cost solutions for wireless applications, Machine Learning, Wireless Communications, Microstrip Antenna design, Filter design, and Artificial Intelligence. He is an active reviewer for many reputed IEEE journals and letters.



**Dr. Ganga Prasad Pandey**, senior member of IEEE, received B. Tech degree in Electronics and Communication Engineering from KNIT Sultanpur, UP in year 2000 and M. E. from Delhi College of Engineering, Delhi India in 2004. He received PhD from Uttarakhand Technical University, Dehradun, UK, in 2015 in the field of Tunable microstrip antenna. With more than 21 years of teaching experience, currently he is

heading the Department of Information and Communication Technology at Pandit Deendayal Energy University, Gandhinagar. He has published more than 40 paper in international Journals of repute and more than 20 papers in International conference. He has guided 4 PhD scholars in the domain of Antenna. His research interest include Energy harvesting, ME-dipole, active, Reconfigurable, frequency agile microstrip antennas and microwave/millimeter wave integrated circuits and devices.



**Dr. Dinesh Kumar Singh** received B.E. in Electronics and Communication Engineering from Kumaon Engineering College, Dwarahat, Almora in 2003. He has done M. Tech in Digital Communication from RGPV University, Bhopal, India. He has done Ph.D. from Indian Institute of Technology (ISM), Dhanbad, Jharkhand, India.

His area of interest is microwave Engineering. He is currently working as Professor in Electronics and Communication Engineering Deptt., G L Bajaj Institute of Technology and Management, Greater Noida, UP, India. His research interest, include designing of high gain, compact, re-configurable, fractal-shape, circularly polarized micro strip antennas, substrate integrated wave-guide (SIW) and Magneto-Electric (ME) dipole antenna for modern communication system. He has been credited to publish more than 20 papers with various reputed international journals and conferences. He is also reviewer of the AEU-International Journal of Electronics and Communication, Electronics Letter, etc.



Milind Mahajan obtained his BE (electronics) degree in 1991 from Marathwada University and MTech degree in Microwave Engineering from I.I.T., BHU in 1993. He received PhD degree from D.D. University, Nadiad in 2015. He started his carrier in Spacecraft Payload Group of Space Applications Centre, Ahmedabad in 1993. He is currently working as Group Director, Antenna Systems Group. He worked as a guest Scientist at

German Aerospace Centre (DLR) in 2001. His current areas of interest are contoured beam reflector and digital beamforming antennas. He has led the team to develop the antenna systems for navigation, radar imaging,

communication satellites (like INSAT-4A/4B/4C, GSAT-7/7A, GSAT-11), and Chandrayaan-2 mission. He is recipient of Space Gold Medal of Astronautical Society of India in 2005, ISRO's team excellence awards in 2007, 2008, 2015, and 2017. He has more than 50 publications in national/international journals and conferences and three patents to his credit. Email: mb\_mahajan@sac.isro.gov.in.




**Dr. Anupma Gupta** received B.E. degree in Electronics and Communication Engineering from Guru Jambheshwar University, Haryana, India in 2006 and M.Tech degree from Maharishi Markandeshwar University, Haryana, India in 2010. She completed his Ph.D degree from the Department of Electronics Engineering, Thapar Institute of Engineering and Technology, Punjab,

India in 2021. She is currently working as Assistant Professor-Chitkara University, Punjab, India. Her research interest includes computational Electromagnetics, reconfigurable antennas, novel electromagnetic materials, dielectric resonator antennas, wideband/superwide band antennas, wideband/dual band/triple band microstrip antennas for wireless communication, smart and MIMO antennas systems, antennas for healthcare, RF MEMS planar antenna. She has published more than 25 research papers and granted with 4 patents. He is also reviewer various reputed journals.

RESEARCH ARTICLE | SEPTEMBER 05 2023

## Mixing enhancement in vortex serpentine micromixer having two and four non-aligned inlets

Deepak Kumar ; Abhyuday Singh Latwal; Mohammad Zunaid; Samsher



AIP Conf. Proc. 2863, 020016 (2023)

<https://doi.org/10.1063/5.0172205>



CrossMark

### Articles You May Be Interested In

Numerical investigation of multiscale lateral microstructures enhancing passive micromixing efficiency via secondary vortex flow

*Physics of Fluids* (September 2022)

Chaotic mixing in a planar, curved channel using periodic slip

*Physics of Fluids* (March 2015)

Integrated microfluidic chip for rapid DNA digestion and time-resolved capillary electrophoresis analysis

*Biomechanics* (March 2012)

500 kHz or 8.5 GHz?  
And all the ranges in between.

Lock-in Amplifiers for your periodic signal measurements



Find out more



# Mixing Enhancement in Vortex Serpentine Micromixer Having Two and Four Non-Aligned Inlets

Deepak Kumar <sup>a)</sup>, Abhyuday Singh Latwal <sup>b)</sup>, Mohammad Zunaid <sup>c)</sup>, Samsher <sup>d)</sup>

*Department of Mechanical Engineering, Delhi Technological University, Delhi, India*

<sup>a)</sup> Corresponding author: deepak750369@gmail.com

<sup>b)</sup> abhyuday258@gmail.com

<sup>c)</sup> mzunaid3k@gmail.com

<sup>d)</sup> samsher@dce.ac.in

**Abstract.** Micromixers are important devices based on mechanical microparts used to mix fluids, which work in chemical, pharmaceutical, analytical and biochemical analysis. Fluid mixing at the microscale is a critical stage in microfluidic systems. The flow and mixing behavior of fluid changes significantly when the length size is scaled down to the order of microns because of the dominance of the viscous forces over the inertial forces at the macroscale level. Mixing at the microscale is based on the diffusion mass transport phenomenon, which can take a long time and require an extensive microchannel length to achieve the outcome. The current study investigates flow characteristics and mixing behavior of a standard serpentine, vortex serpentine micromixer having two and four non-aligned inlets with Newtonian fluid water. In micromixer, inlet passage is tangentially oriented to the main microchannel at the one end. Some mathematical equations listed under mathematical modeling are used to calculate mixing performance in terms of the mixing index. The results were found to be related to a simple serpentine mixer. Vortex flow is generated by a vortex serpentine mixer having two and four inlet channels to improve mixing. The vortex mixer having two non-aligned inlet passages has the great blending performance of the three cases due to the highest mixing index.

**Keywords:** Serpentine Micromixer; vortex flow; nonaligned inlets; micromixer; Dean flows; mixing Index

## INTRODUCTION

Microfluidic devices are used in various industries varying from analysis and reaction to bio-engineering and their applications have increased significantly in recent years [1]. The popularity of advantage of using these devices is that it offers several benefits over conventionally sized systems. It allows analysis and use of less volume of samples, chemicals, and reagents. Micromixer technologies are used in every industry like chemical applications including polymerization and extraction and biological applications including DNA analysis and many more [2]. Multiprocessing capability has resulted in various applications [3]. Methodologies include soft-lithography, 3D printing, laser-assisted chemical engraving, and even paper-based materials, number of scholars have shown interest in them and have begun to investigate these. Important operations that these mixer devices must perform to function is mixing. Efficient mixing in a microfluidic device is a big challenge and to solve this a group of researchers focused on conceptual, execution, and production aspects of microscale stir. Lee et al. [4], Nguyen et al. [5], and Wu et al. [6] have given an in-detail analysis of the various plan used in enhancing microfluidic platforms. Two types of micromixers active and passive are available. An active micromixer is a microfluidic [7] device that improves species mixing by adding outer energy [8] interruption and it is caused by moving components within the micromixer, such as magnetic stirrers or by introducing external force such as pressure, electrohydrodynamic, thermal, and so on. One of the microfluidic devices is a passive micromixer. Except for the appliance it is used to steer the fluid flow at a constant rate and it requires no input energy, because



the laminar flow has an advantage at the microscale, mixing in passive micromixer is based on chaotic advection noticed by exploiting laminar flow in molecular diffusion [9]. Although active micromixers have high blending productivity and control over a wide span of Reynolds numbers.

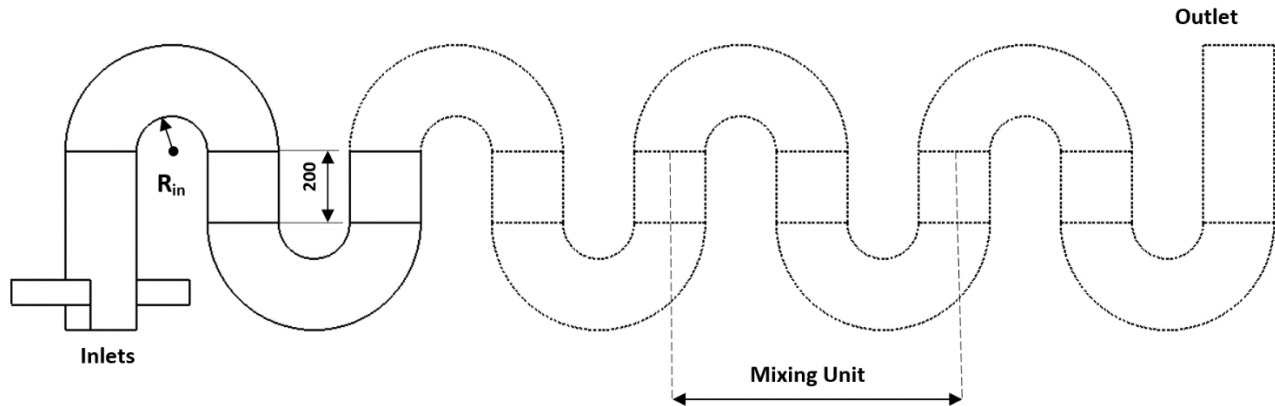
It becomes more difficult to produce and combine within microfluidic components and it needs external power sources [10]. Using 3D structures with complex geometrical structures results in fabrication problems in some cases. Passive micromixers have no impact from the reason mentioned above about micromixers. Because they require only pressure-driven flow and do not need an external energy source [11], and are easy to merge into complex microfluidic systems using quality fabrication techniques. Furthermore, the lack of advanced multi-physics [12] interactivity to account for makes them far more compliant to conceptual or computational modeling and it enables a further simple and logical optimization of the various geometrical [13] and motion parameters required to enhance blending in designs [14]. In passive micromixers, frequent curved sections and turns are popular designs to attain cross-sectional flows [15]. These systems take advantage of the radial forces accomplished by the fluid as it moves through a curved trajectory because it is guided by geometry. Dean [17] analyzed the flow field that develops. As the evolution of transversal whirlwind occurs within a system it is also called Dean flows. These use serpentine or spiral-shaped designs to promote advective transport in microchannels in a geometrically simple way. Primary advantage builds them appealing for biological applications requiring the safe pick up for large biomolecules [18].

Offset or non-aligned micromixers have been studied deeply by many researchers but more work needs to be done regarding micromixers with four non-aligned inlets and their effect on mixing phenomena. Therefore the purpose of this paper is to investigate the mixing performance of standard serpentine mixers and vortex serpentine mixers having two and four non-aligned inlets based on mixing index and various contours.

## METHODOLOGY

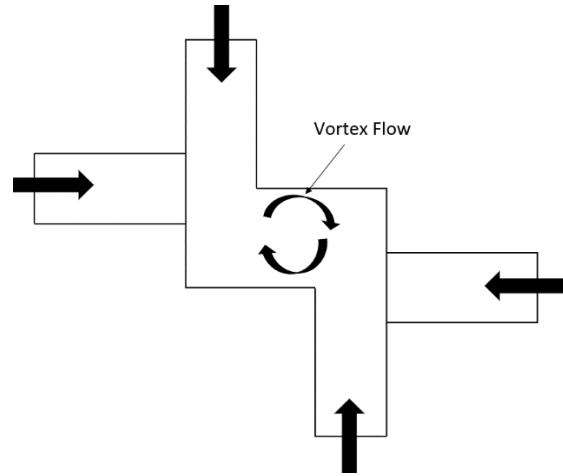
### Geometrical Design of the Micromixer

Three micromixers are chosen for analysis. The base case is a standard serpentine micromixer with the main microchannel with depth(H) and width(W) of 100  $\mu\text{m}$  and 200  $\mu\text{m}$ , respectively, which is compared to vortex serpentine micromixers with two and four nonaligned inlet channels. The square inlet is designed to keep the mass flow rate the same in all cases. In this case, the Reynolds number represents the dimensions of the main channel and water as the working fluid. Each mixing unit is made up of two semi-circular sections joined together by the same line. The section which is straight and its length which connects successive turns is the width(W). The fluid flow and mixing performance were examined in given cases with the inner turn radius  $R_{in}$  as 0.5  $W=100 \mu\text{m}$ . Inlet passage is only positioned tangent to the mixing channel with offset in the case of two nonaligned inlet micromixers, resulting in whirlwind movement in the microchannel. Further drawing is expanded to a micromixer having four non-aligned microchannels. The dimensions listed above were used to design and model all three types of serpentine micromixers in Solid Works.

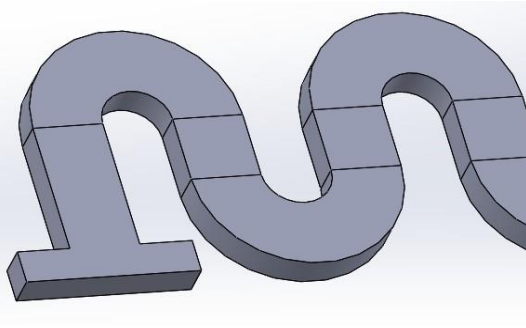


**FIGURE 1.** Top view of serpentine micromixer having four non-aligned inlets

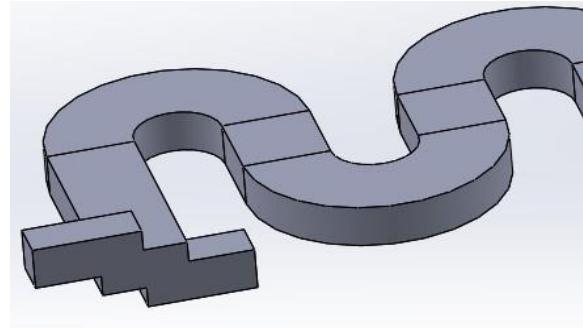




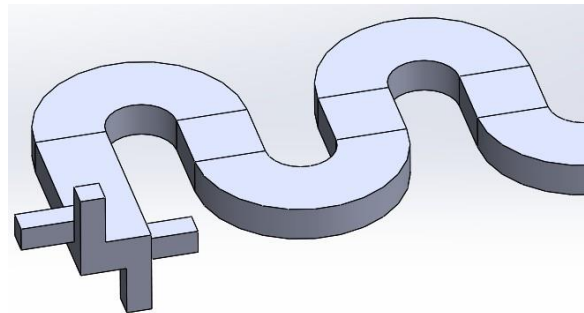
**FIGURE 2.** The design concept of vortex serpentine micromixer



**FIGURE 3.** Standard serpentine micromixer  
Fig. created by authors using data from ref. [16]



**FIGURE 4.** Vortex serpentine micromixer having two non-aligned inlets



**FIGURE 5.** Vortex serpentine micromixer having four non-aligned inlets

## Mathematical Modelling

The nonlinear partial differential equations governing momentum, continuity, and species transport are discretized using the finite volume method and numerically solved using the CFD solver of ANSYS fluent. Fluid is incompressible and has a steady flow. To achieve velocity & concentration fields in all the cases [10] various equations are used and these are given below:

### Continuity Equation

$$\nabla \cdot \vec{u} = 0 \quad (1)$$

### Navier-Stokes Equation

$$\rho(\vec{u} \cdot \nabla) \vec{u} = -\nabla p + \nabla \cdot \vec{T} \quad (2)$$

In this  $\nabla$  represents Del Operator,

$$\nabla = \frac{\partial}{\partial x} \hat{i} + \frac{\partial}{\partial y} \hat{j} + \frac{\partial}{\partial z} \hat{k} \quad (3)$$

And  $\rho$  denotes fluid density,  $p$  is static pressure

$\vec{T}$  represents stress tensor and it is formulated below:

$$\vec{T} = \eta(\nabla \vec{u} + \nabla \vec{u}^t) \quad (4)$$

Here  $t$  is transpose operation, whereas  $\eta$  represents apparent fluid viscosity.

### Species Transport Equation

$$(\vec{u} \cdot \nabla) C_A = D_{AB} \nabla^2 C_A \quad (5)$$

Here  $\vec{u}$  is the velocity vector,  $C_A$  denotes concentration of species A and  $D_{AB}$  is the molecular diffusivity. The density, dynamic viscosity, and diffusivity are considered as 1000 Kg/m<sup>3</sup>, 0.001 Pa-s, and  $1 \times 10^{-9}$  m<sup>2</sup>/s, respectively for water which is a Newtonian fluid

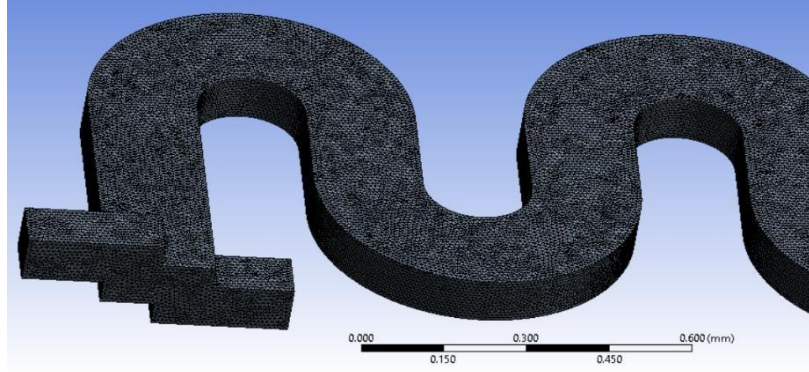
The formula for the Mixing index is written below:

$$Mi = \left( 1 - \sqrt{\frac{\sigma^2}{\sigma_{\max}^2}} \right) \quad (6)$$

$\sigma^2$  is the real variance in the mass fraction of the liquid constituent &  $\sigma_{\max}^2$  denotes the highest variance.  $c_i$  is mass fraction at the given position  $i$ , and Avg is the average value of accumulation of the liquid at a cross-section under consideration. In mixing with two identical flows of liquid, the value of  $\sigma_{\max}^2$  is 0.25. In this case,  $n$  represents the gross number of points considered on a given level, the large standard of  $n$  enhances the accuracy of the mixture therefore  $n$  is 900 in the current study. The range of mixing index (Mi) is 0 to 1, with 0 representing no mixing and 1 representing proper mixing of fluids.

### Numerical Methodology

As the complexity of flow pattern in Serpentine micromixer, the geometry of micromixers is modeled using Solid Works software and ANSYS and meshing module create a grid with tetrahedral elements, illustrated in the picture below.

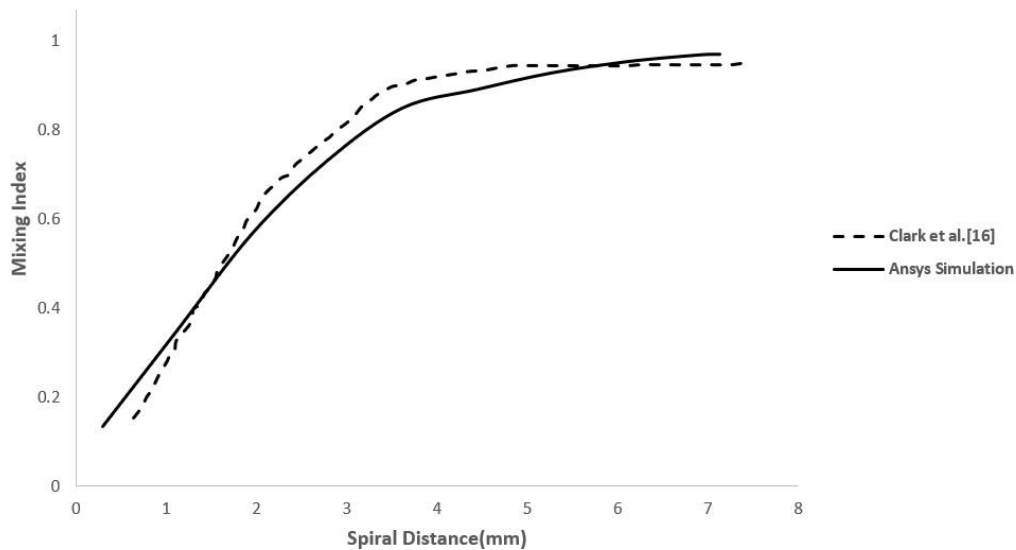


**FIGURE 6.** Tetrahedral elements mesh for standard serpentine micromixer

To analyze the movement of fluid and mixing in the micromixers and for simulation, ANSYS is used and also to check the mixing behavior. Some algorithms are used like SIMPLE for pressure velocity integration and the residual value is  $10^{-8}$  for its numerical solution.

In blending, water goes in inlet with species concentration 'one' and water-dye mixture enters in the shaft with species concentration set out as 'Zero'. Velocity in all the shafts is established using Reynolds number. Zero wall flux conditions are set to the fence of the micromixer. It is set to 'Zero' diffusive flux at walls, and the outlet of the mixer has been given zero specific pressure.

Validation was carried out to verify the exactness of this model by comparing the results from Clark et al.[16] for a standard serpentine micromixer with Reynolds number 100.

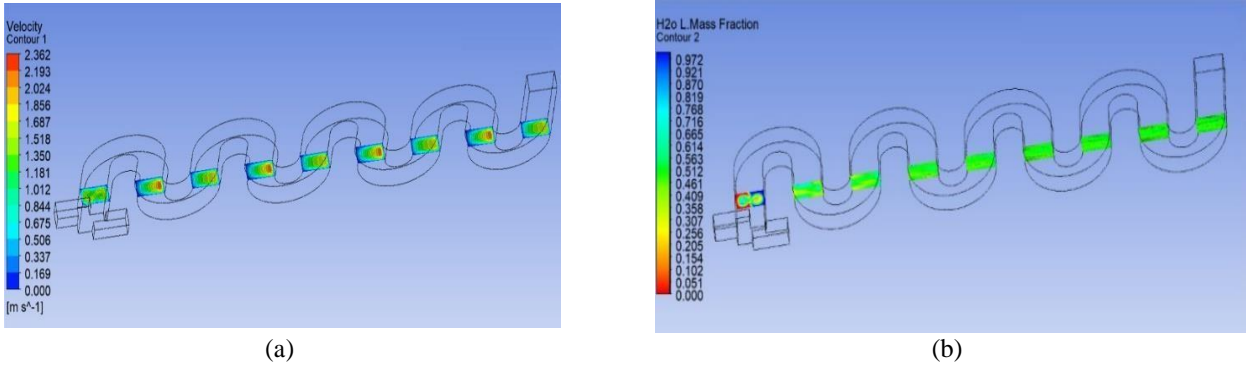


**FIGURE 7.** Validation of the computational method of the present study with existing literature by comparing Mixing Index

## RESULTS AND DISCUSSION

The mixing index or mixing efficiency calculated at a particular cross-section is a parameter to quantify the blending phenomena occurring between the flowing liquid in the microchannel. Concentration contour and velocity plots were compared for all the 3 cases corresponding to flow with Reynolds number 100 along the micromixer for different planes. Mixing Index is also determined and compared for all the three cases considered in this study

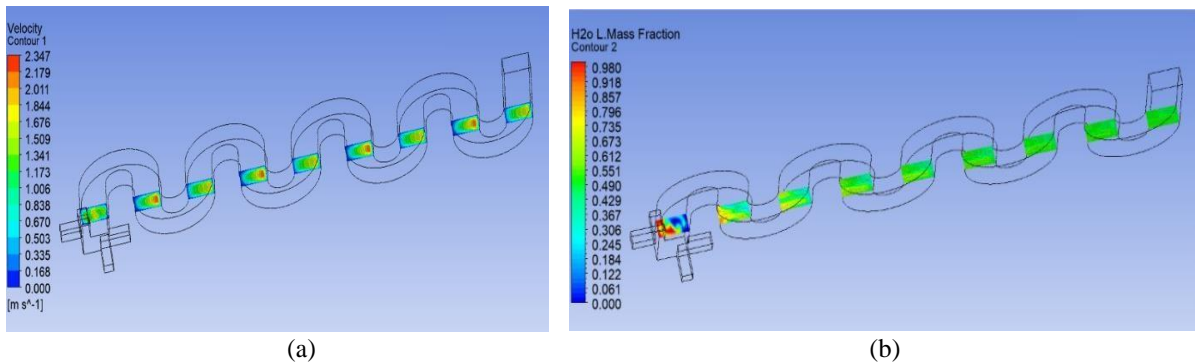
### Vortex Serpentine Micromixer Having Two Non-Aligned Inlets



**FIGURE 8.** (a) Velocity contours for vortex serpentine micromixer having two non-aligned inlets (b) Concentration contours for vortex serpentine micromixer having two non-aligned inlets

From the above figure in two inlets, vortex flow is created due to the large interface area of the fluid as it reaches its offset site thus improving blending performance. Complete mixing of the two liquids is obtained in a short distance as compared to other cases in concentration contour. Hence proving the effectiveness of this device. From velocity contour, it can be inferred that the highest magnitude is obtained out of all three cases.

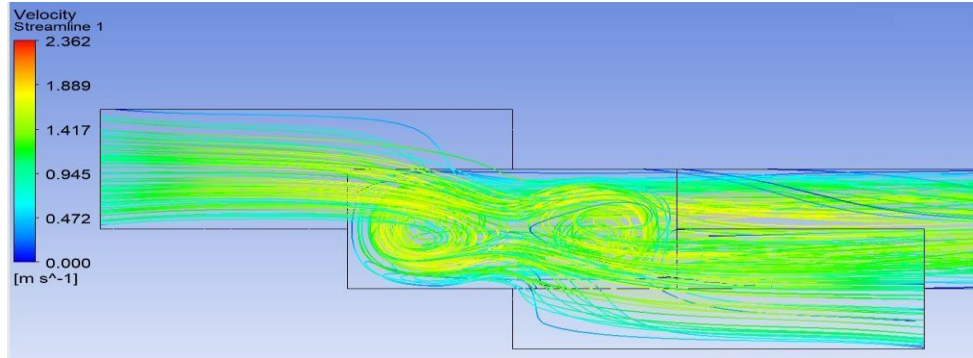
### Vortex Serpentine Micromixer Having Four Non-Aligned Inlets



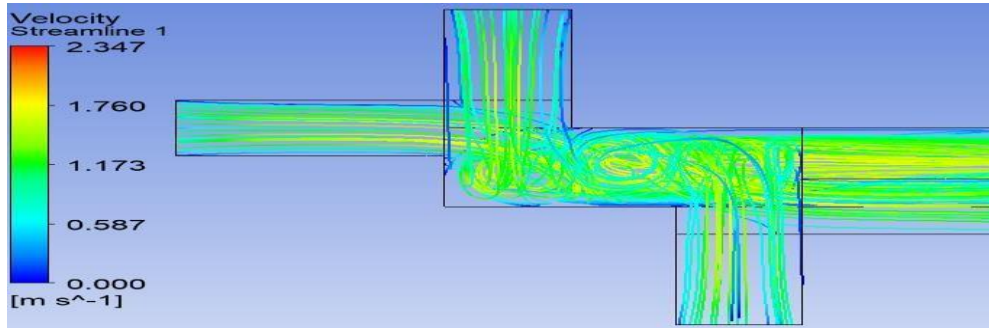
**FIGURE 9.** (a) Velocity contours for vortex serpentine micromixer with having four non-aligned inlets (b) Concentration contours for vortex serpentine micromixer having four non-aligned inlets

It is expected that there is a stronger whirlwind flow and better mixing performances. But results were against our expectation, as it is not able to create strong flow in four inlets because the interfacial area of the fluid streams is less than that of a non-aligned two inlet micromixer but more than a simple serpentine micromixer. Near the junction, a whirlwind is formed and it exists at a certain distance due to the effect with viscous force its intensity decreases for both the cases of vortex serpentine micromixer with non-aligned inlets.

On comparison of velocity streamlines between nonaligned two inlets and four inlets, serpentine micromixer it can be seen that vortex mixer creates vortex flow and better mixing is obtained in two non-aligned inlets due to higher magnitude of velocity. Such flows are desirable because they aid in the increased and improved mixing of fluids.

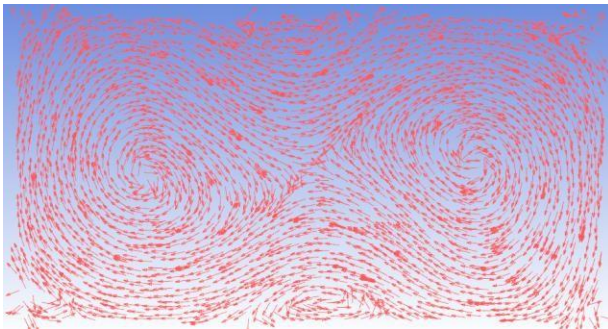


**FIGURE 10.** Streamlines for vortex serpentine micromixer having two non-aligned inlets

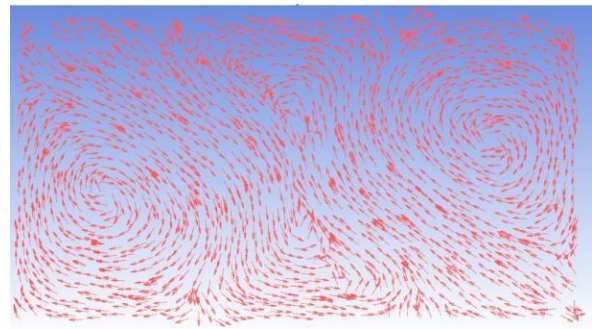


**FIGURE 11.** Streamlines for vortex serpentine micromixer having four non-aligned inlets

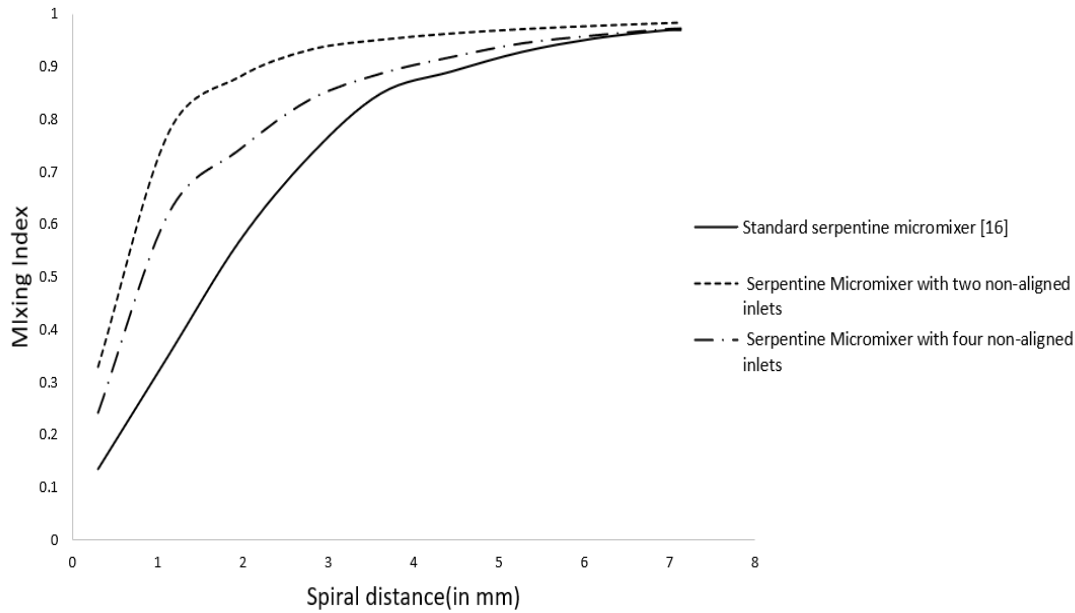
To understand the passive mixing in the microchannel, it is important to refer to the flow pattern for both vortex mixers separately. Figures 12 and 13 depict the velocity vector on a plane located at the junction of inlets of the micromixer. It can be inferred from the above-mentioned figures that the formation of vortices begins at the junction due to the high-velocity collision in the case of Newtonian fluid water which enhances the mass diffusion between the interface. Velocity vector plots based on velocity fields in vortex micromixer having non-aligned inlets shows that having orientation along with the bend results in the formation of counter-rotating transversal flows.



**FIGURE 12.** Velocity vector plot for vortex serpentine micromixer having two non-aligned inlets



**FIGURE 13.** Velocity vector plot for vortex serpentine micromixer having four non-aligned inlets



**FIGURE 14.** Comparison of mixing index for all the three cases

Simple serpentine micromixer, as expected, has the poorest mixing performance among the three designs along the length of the micromixer. Among three design mixers having two non-aligned inlets generates the most whirlwind in the channel at  $Re = 100$ , implying that it is very effective at mixing. At lower Reynolds numbers, a whirlwind mixer having four non-aligned inlet passages is less efficient of producing whirlwind motion than two non-aligned inlets.

## CONCLUSIONS

Three different micromixer designs are investigated in this study. A standard serpentine mixer, as well as mixers having two and four non-aligned inlet passages, are among those designs. Based on Dean flows and doing various quantitative studies of mixing quality that can be taken from the micromixers shows that using serpentine micromixers in conjunction with vortex two and four nonaligned inlets results in greater mixing performance, due to the simplicity of the proposed modification, it is simple to incorporate these serpentine designs used in microfluidic devices for efficient mixing. Computational fluid dynamics show that the cross-sectional flow structures present in these types of channels reveal that the formation of secondary transversal vortexes is responsible for their improved performance. Due to its presence and large quantity transfer of mass in the microchannel stretch the interface between the component and able to combine it promoting chaotic advection. At the other end of the microchannel, vortex mixers join at tangential positions. After entering in microchannel stream generate whirlwind flow. Out of the given cases, a mixer having two non-aligned inlet channels is best for mixing due to its higher mixing index.

## ACKNOWLEDGMENT

The authors would like to express their gratitude to Delhi Technological University for its assistance.

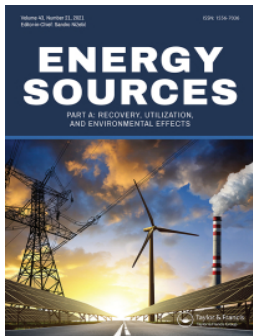
## REFERENCES

1. Sackmann, E.K.; Fulton, A.L.; Beebe, D.J. The present and future role of microfluidics in biomedical research. *Nature* 507, (*Arabian Journal for Science and Engineering*, Washington dc USA, 2014), pp. 181–189.
2. Capretto, L.; Carugo, D.; Mazzitelli, S.; Nastruzzi, C.; Xunli, Z. Microfluidic and lab-on-a-chip



preparation routes for organic nanoparticles and vesicular systems for nanomedicine applications. *Adv. Drug Deliv. Rev.* 65, (Advance Drug delivery, London United Kingdom 2013), pp. 1496–1543.

3. Chiu, D.T.; deMello, A.J.; Di Carlo, D.; Doyle, P.S.; Hansen, C.; Maceiczky, R.M.; Wootton, R.C.R. Small but perfectly formed? Successes, challenges, and opportunities for microfluidics in the chemical and biological sciences. (Journal of Micromechanics and microengineering Salt Lake City, United States of America 2017), pp.201–223.
4. Lee, C.Y.; Chang, C.L.; Wang, Y.N.; Fu, L.M. Microfluidic mixing: A review. *Int. J. Mol. Sci.*, (International Journal of Molecular Sciences, Pingtung, Taiwan 2011), pp.3263–3287
5. Nguyen, N.-T. *Micromixers: Fundamentals, Design and Fabrication*, 2nd ed. (Elsevier: Oxford, UK; ISBN 2012), pp. 301– 332.
6. Cai, G.; Xue, L.; Zhang, H.; Lin, J. A review of micromixers. *Micromachines*, 8 (Beijing, China 2018).
7. Bamford, R.A.; Smith, A.; Metz, J.; Glover, G.; Titball, R.W.; Pagliara, S. Investigating the physiology of viable but non-culturable bacteria by microfluidics and time-lapse microscopy. (ASM Journal South West England 2017), pp. 23-44.
8. Zilionis, R.; Nainys, J.; Veres, A.; Savova, V.; Zemmour, D.; Klein, A.M.; Mazutis, L. Single-cell barcoding, and sequencing using droplet microfluidics. (Nature Protocols Boston, Massachusetts, USA 2017) pp .44–73.
9. R. Prakash, M. Zunaid and Samsher, Simulation analysis of mixing quality in T-junction micromixer with bend mixing channel, (Materials Today: Proceedings, India 2021).
10. A. Sinha and M. Zunaid, Numerical study of passive mixing in a 3-dimensional helical micromixer with two inlets at the offset, (Materials Today: Proceedings, India 2021).
11. S. Tokas, M. Zunaid, Mubashshir Ahmad Ansari, Numerical investigation of the performance of 3D-helical passive micromixer with Newtonian fluid and non-Newtonian fluid blood, (Asia-Pac. Journal, India 2020).
12. Scherr, T.; Quitadamo, C.; Tesvich, P.; Park, D.S.; Tiersch, T.; Hayes, D.; Choi, J.W.; Nandakumar, K.; Monroe, W.T. A planar microfluidic mixer based on logarithmic spirals. (J. Micromech. Euclid Avenue, USA 2012).
13. Shamloo, A.; Madadelahi, M.; Akbari, A. Numerical simulation of centrifugal serpentine micromixers and analyzing mixing quality parameters. (Elsevier, Azadi Ave. Tehran, 2016), pp. 243–252.
14. Minakov AV, Rudyak VY, Gavrilov AA, Dekterev AA. Mixing in a T-shaped micromixer at moderate Reynolds numbers. (International Journal of Heat and flow, Novosibirsk, Russia, 2012), pp. 385-395.
15. Designer, G.; I'm, S.; Ha, B.H.; Jung, J.H.; Ansari, M.A.; Sung, H.J. Adjustable, rapidly switching microfluidic gradient generation using focused traveling surface acoustic wave (Seodaemun-gu, Seoul, 2014).
16. Clark, J., Kaufman, M., & Fodor, P. S. Mixing Enhancement in Serpentine Micromixers with a Non-Rectangular Cross-Section. *Micromachines* (Euclid Avenue, Cleveland, USA, 2018).
17. Patel, M.V.; Tovar, A.R.; Lee, A.P. Lateral cavity acoustic transducer as an on-chip cell/particle microfluidic switch. *Lab Chip* 12 (Saarbrücken, Germany, 2012), pp. 139–145.
18. Abbas, Y.; Miwa, J.; Zengerle, R.; von Stetten, F. Active continuous-flow micromixer using an external braille pin actuator array. *Micromachines* (ISSN, Freiburg, Germany, 2013), pp. 80–89.



## Modeling and analysis for enhanced hydrogen production in process simulation of methanol reforming

Neeraj Budhraja, Amit Pal & R. S. Mishra

**To cite this article:** Neeraj Budhraja, Amit Pal & R. S. Mishra (2023) Modeling and analysis for enhanced hydrogen production in process simulation of methanol reforming, Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 45:4, 11553-11565, DOI: [10.1080/15567036.2023.2262414](https://doi.org/10.1080/15567036.2023.2262414)

**To link to this article:** <https://doi.org/10.1080/15567036.2023.2262414>



Published online: 25 Sep 2023.



Submit your article to this journal [↗](#)



Article views: 14



View related articles [↗](#)



View Crossmark data [↗](#)



# Modeling and analysis for enhanced hydrogen production in process simulation of methanol reforming

Neeraj Budhraj<sup>ID</sup>, Amit Pal<sup>ID</sup>, and R. S. Mishra<sup>ID</sup>

Department of Mechanical Engineering, Delhi Technological University, Delhi, India

## ABSTRACT

Hydrogen has emerged as the most suitable fuel for a nation's greener and sustainable development. In contrast, the feedstock and hydrogen production methods remain a concern for environmental pollution. This study uses methanol as the feedstock for hydrogen production via a low-temperature methanol-reforming process. A simulation model was developed in Aspen Hysys, where an equilibrium reactor is used in the reforming process, and examined the effects of parameters like temperature, pressure, and Methanol-to-Water (M-to-W) molar ratio. Hydrogen mole fraction and selectivity increase by roughly 18.5% and 10.5% when the reaction temperature increases from 100°C to 400°C. At the same time, the methanol conversion rate reaches 95% at 400°C. Reactor pressure shows inverse effects where pressure rises from 1 atm. to 7 atm. that reduces hydrogen mole fraction and selectivity by about 10% and 6%, and a similar reduction of 5% is noticed in the methanol conversion rate. M-to-W molar ratio plays a crucial role in the reaction pathway and the M-to-W ratio between 0.5 and 1.5 at 400°C and 1 atm. reactor pressure showed the highest hydrogen mole fraction (>0.57) and a maximum methanol conversion rate (>90%). Therefore, the present simulation model successfully determines the impacts of various parameters to help design a commercial plant for large-scale hydrogen production via the reforming process.

## ARTICLE HISTORY

Received 31 May 2023  
Revised 15 August 2023  
Accepted 15 September 2023

## KEYWORDS

Reforming; hydrogen; methanol conversion; Aspen hysys; Peng-robinson

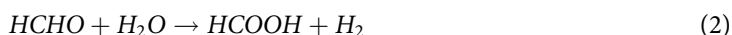
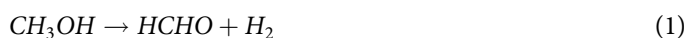
## Introduction

The global energy demand is rising with the development of the human race. After technological advancement and environmental concerns, new and renewable energy sources have been developed (Qureshi et al. 2022). But each renewable energy source has its disadvantages- solar energy is dynamic in nature and available during sunshine hours, wind energy is recommended for seashore areas where sufficient wind speed is available, hydropower affects flora and fauna of the region, and biomass energy requires large land masses for cultivation and storage. The transportation of energy to the place of utilization during demand time is also a big challenge for most renewable energy sources (Garcia et al. 2021). Therefore, a lot of work has been performed in the last few years to develop an energy source that can overcome these drawbacks without causing damage to the environment. And Hydrogen is one such alternative.

Hydrogen is an energy carrier that can store energy and be transported to any place and time when demanded (Budhraj, Pal, and Mishra 2023b). In contrast, greener hydrogen production is still a challenge to meet the demand of industries. Therefore, the need is met using conventional methods and fossil fuels to generate hydrogen, contributing to about 98% of the total hydrogen production worldwide (Ranjekar and Yadav 2021). Fossil fuels like coal, oil, and natural gas, and conventional methods like steam reforming, partial oxidation, and auto-thermal reforming directly or indirectly add

to environmental pollution (Du, Mo, and Li 2015). However, with the agreement of the 21<sup>st</sup> Conference of Parties (COP25) held in Paris, where the focus is put on reducing greenhouse gas (GHG) emissions, each member country is responsible for reducing the GHG emissions to limit the global temperature rise to less than 2°C (Singh et al. 2022). Hydrogen is one of the contenders, and renewable sources like biogas, water, and alcohol can replace fossil fuel feed.

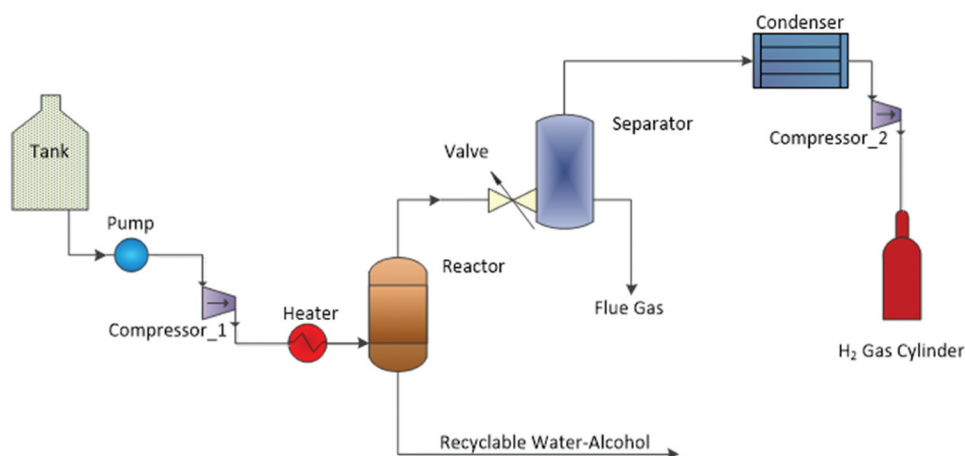
Methanol has higher hydrogen content (12.5%). It is less toxic, liquid at room temperature, and can generate hydrogen at a comparatively lower temperature (AlNouss, McKay, and Al-Ansari 2020). These advantages developed an interest in methanol as a greener source of hydrogen production. Shen et al. (2017) successfully tested an enzyme mimic, [Cp\*IrCl(phen)]Cl, that can continuously generate hydrogen from methanol at near-room temperature. The process involves the conversion of methanol into formaldehyde and then later into hydrogen (Shen et al. 2017). Zhang et al. (2018) examined the catalytic activity of Au-Ti-Ce/Na-ABen for hydrogen production from methanol via steam reforming. The results showed a 72% methanol conversion with 99% hydrogen selectivity at 350°C (Zhang et al. 2018). The methanol-reforming process involves reactions (1), (2), and (3) as stated below:



In some studies, the researchers have used temperatures slightly above 100°C and a particular catalyst to produce hydrogen at lower temperatures. Awasthi et al. (2021) developed efficient ruthenium (Ru) catalyst for hydrogen production from methanol at a low-temperature range of 110–130°C. The study revealed a 186 L hydrogen production from 1 g Ru and a 1.43 mol hydrogen per mol of methanol, which is an outstanding achievement (Awasthi et al. 2021). The process does not produce CO<sub>2</sub>, while formic acid is generated as the by-product. Few researchers have simulated and modelled different hydrogen feeds using Aspen software.

Simulation of a chemical plant requires developing models for different processes during plant operation. Only a few works are available for methanol reforming; however, other sources for hydrogen production, like glycerol, biomass, etc., have similar working and influencing parameters. Unlu and Hilmioglu (2020) investigated glycerol steam reforming in an Aspen Plus simulator for hydrogen production. The simulation showed that the reaction temperature directly impacts the hydrogen concentration, while the reactor pressure has adverse effects (Unlu and Hilmioglu 2020). The study defined a 9:1 glycerol ratio and 1 atm. pressure at 500°C as the optimum condition for hydrogen production. Mohammadidoust and Omidvar (2020) simulated the model developed for wheat straw biomass gasification at supercritical conditions. The model attained a 32.7 kg/h hydrogen production rate at 700°C with 2000 kg/h and 1500 kg/h of water and biomass flow rate (Mohammadidoust and Omidvar 2020). Tavares et al. (2020) determined the influence of gasification temperature and steam-to-biomass ratio in hydrogen production using Aspen Plus. The simulation showed a high hydrogen content in syngas at higher temperature ranges (Tavares et al. 2020). Ye et al. (2009) also developed a model for simulating the influencing parameters in the hydrogen production process. Hydrogen yield increases with the rise in operating temperature and steam-to-carbon ratio (Ye et al. 2009). Therefore, the operating temperature and feed ratio are the two most influencing parameters in hydrogen production.

In the current study, a simulation model of methanol-reforming is developed in the Aspen Hysys simulator. An equilibrium reactor is used while designing the methanol-reforming reactor, while the Peng-Robinson thermodynamic model simulates the reaction sets. The model validation is performed by considering data from various published works, and the results are compared on a percentage difference basis. The study aims to understand the influence of the input parameters on hydrogen selectivity and methanol-water conversion rate



**Figure 1.** Process flow diagram of the methanol-reforming process.

**Table 1.** Simulating parameters and their values.

Parameter	Unit	Values
Temperature	°C	100, 150, 200, 250, 300, 350, 400
Reactor pressure	atm.	1, 2, 3, 4, 5, 6, 7
Methanol-to-Water (M-to-W) molar ratio	—	0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 2.0, 2.5, 3.0, 3.5, 4.0

through the model simulation. The parameters chosen for simulation include reaction temperature (from 100°C to 400°C), reactor pressure (from 1 atm. to 7 atm.), and methanol-to-water molar ratio (from 0.3 to 4.0). A purification unit (or separator) is also attached to the model to generate pure hydrogen (99.99% pure) that can be used in fuel cells or other industrial applications. It will aid in designing large-scale hydrogen production for commercial applications with optimized parameters.

## Simulation modeling

The simulation model is developed in Aspen Hysys for a methanol-reforming plant (see Figure 1), which uses mathematical models to simulate chemical processes in chemical plants and refineries. The high purity of methanol and water at a 1.0 molar ratio is considered for simulation because, while simulation and from published works, it is observed that the best methanol-to-water ratio lies between 0.5 and 1.5; therefore, the middle value is chosen. Though for determining the effect of the molar ratio, the molar ratio varied from 0.3 to 4 methanol-to-water ratios. The methanol and water are mixed in a mixer unit, and the mixture is then passed to a heat exchanger (part of the reformer) to adjust the temperature from 100°C to 400°C. Reactor pressure is another parameter considered for the simulation process, and its range is maintained between 1 atm. and 7 atm. Similar parameter ranges (temperature from 50°C to 500°C and pressure from 1 atm. to 5 atm.) were also considered for glycerol steam reforming (Unlu and Hilmioglu 2020).

The parameter ranges are taken from previously published work (Chen et al. 2019). Table 1 describes the various parameters and their range in the simulation process.

A hydrogen purification unit is also added to the model that generates 99.99% pure hydrogen, which is helpful in various applications like fuel cells and industries. The hydrogen production rate, hydrogen selectivity, and feed conversion rate are calculated using the simulation outputs.

### Selection of reactor

Aspen Hysys software has different types of reactors for various applications and reaction types (Shamsi et al. 2022). Therefore, choosing an appropriate reactor type becomes essential to simulate the model correctly and obtain the desired results. The reactor selection is based on the reaction set and input-output parameters. The equilibrium reactor is observed to be suitable for the available reaction set; hence, it is selected to simulate the methanol-reforming process (Giwa, Giwa, and Giwa 2013). The Eqs. (4) and (5) involved in the methanol-reforming process are equilibrium reactions; therefore, an equilibrium reactor is chosen for the modeling and simulation process (Zhao et al. 2020).



However, the water gas shift reaction neutralizes carbon monoxide and produces extra hydrogen (Kanatlı and Ayas 2021), as shown in Eq. (6),



The following assumptions are made during the modeling and simulation of the methanol-reforming process:

- (1) the system works in steady-state
- (2) methanol and water are fed at constant temperature and pressure
- (3) methanol used is 99.9% pure
- (4) the stream is adiabatic in nature
- (5) the formation of free carbon is not considered in the model/system
- (6) flow rate is constant
- (7) in the system, gases behave like ideal gases
- (8) the methanol-water mixture remains constant throughout the simulation
- (9) hydrogen coming out is at a constant temperature (40°C)

### Thermodynamic model

Aspen Hysys has many choices of thermodynamic models. The earlier studies showed Peng-Robinson model is suitable for simulating steam reforming processes (Wang et al. 2022). The Peng-Robinson thermodynamic model is proposed for oil, gas, and petrochemical applications. The degree of efficiency and reliability are the main advantages of solving single, two, and three-phase systems, and the model can be used for a wide range of applications. Therefore, the Peng-Robinson thermodynamic model is chosen for the simulation process in this work. Figure 2 is the simulation model used in the methanol-reforming process. The model consists of a mixing unit “Mixture” that mixes water and methanol feed, and an appropriate methanol-to-water (M-to-W) ratio is obtained. Before the methanol-water mixture is fed to the reactor, a “heater” is placed to attain the desired temperature for the reactor feed. An equilibrium reactor as a “reformer” is selected due to equilibrium reaction sets and suitable input-output parameters. The downstream from the reformer consists of Syngas containing hydrogen and other gases, and the unreacted methanol-water mixture is recycled into the reformer. To get 99.99% pure hydrogen, a hydrogen purification unit “separator” is fed with syngas, as shown in Figure 9.



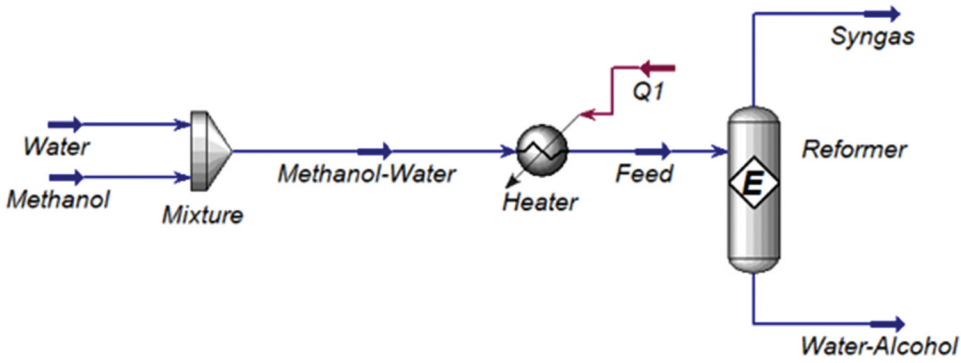


Figure 2. Methanol-reforming simulation model flowsheet.

### Selectivity and methanol conversion

Hydrogen mole fraction (Eq. 7) and methanol mole fraction (eq. 8) were calculated using hydrogen selectivity and methanol conversion percentage from the system (Budhreja, Pal, and Mishra 2023a).

$$x_{H_2} = \frac{W_{H_2}}{W_{total}} \quad (7)$$

$$x_{MeOH} = \frac{W_{MeOH}}{W_{total}} \quad (8)$$

where  $x_{H_2}$  and  $x_{MeOH}$  represent hydrogen mole fraction and methanol mole fraction,  $W_{H_2}$  represents hydrogen production rate (mol/h),  $W_{MeOH}$  represents the methanol flow rate (mol/h), and  $W_{total}$  represents the total feed flow rate in mol/h. In contrast, the methanol conversion percentage and hydrogen selectivity is calculated using Eq. 9 and eq. 10. Here,  $W_{MeOH_{in}}$  and  $W_{MeOH_{out}}$  represent methanol feed rates (mol/h) and unreacted methanol flow rates (mol/h).

$$Conversion\ percentage_{(MeOH)} = \frac{W_{MeOH_{in}} - W_{MeOH_{out}}}{W_{MeOH_{in}}} \times 100 \quad (9)$$

$$Selectivity_{H_2}(\%age) = \frac{W_{H_2}}{3(W_{MeOH_{in}} - W_{MeOH_{out}})} \times 100 \quad (10)$$

## Results and discussions

### Process modeling and validation

The developed Aspen Hysys model is validated against the four different experimental setups under similar conditions, and the feedstock selected is methanol. The aim is to verify the replication of the developed model at different process parameters and conditions. Following are the case studies:

- Case 1: Lee and Kim used a tubular quartz reactor with annular-shaped electrodes for methanol conversion into hydrogen through electric discharge, in which CuO/ZnO/Al<sub>2</sub>O<sub>3</sub> was used as a catalyst. Results showed about 57.6% methanol conversion at about 220°C (Lee and Kim 2013).

Likewise, the Aspen Hysys model simulation achieved a closer 59.3% methanol conversion result at 220°C. This shows a variation of 2.95% compared to the experimental results.

- Case 2: Wang and Wang performed catalytic methanol-reforming, where CuO/ZnO/Al<sub>2</sub>O<sub>3</sub> catalyst was filled in the micro-reactor designed for hydrogen production. An 87.1% methanol conversion was obtained at about 270°C inlet temperature (Wang and Wang 2016). However, the Aspen Hysys model with similar experimental conditions and parameters achieved a methanol conversion of 89%. It shows a 2.18% deviation from the experimental results.
- Case 3: Kim *et al.* performed methanol-reforming in a prototype reactor filled with CuO/ZnO/Al<sub>2</sub>O<sub>3</sub> catalyst. Methanol conversion was calculated at different temperatures, and the highest (96%) methanol conversion was obtained at 290°C (Kim *et al.* 2019). When similar experimental conditions are created in the Aspen Hysys model, the methanol conversion shows a 1.25% reduction from the experimental results. In contrast, the highest methanol conversion obtained was 94.8% in simulation as against 96% in experimental results at 290°C.
- Case 4: Liu *et al.* used 1Pt/3In<sub>2</sub>O<sub>3</sub>/CeO<sub>2</sub> catalyst for hydrogen production from methanol in a process called methanol steam reforming. Experimentally, a very high methanol conversion of 98.7% was achieved at 325°C (Liu *et al.* 2017). When the experimental conditions are put into the Aspen Hysys model, a similar methanol conversion of 98.2% is obtained, which is just 0.51% lower than the experimental results.

Therefore, it concludes that the Aspen Hysys model has minimal variations (below 5%) from the experimental results, as observed in Cases 1, 2, 3, and 4. Hence, the developed model is applicable for methanol reforming and determining the effects of various process parameters in the methanol conversion and hydrogen production process with percentage differences (as shown in Table 2).

### Effect of reaction temperature

Figure 3 depicts the variation of components' mole fraction versus temperature curves, while Figure 4 describes the methanol and water conversion rate into hydrogen at elevated temperatures. The methanol-reforming is performed at atmospheric pressure and a methanol-to-water ratio of 1.0, respectively.

A significant rise in hydrogen mole fraction is observed with the increase in reaction temperature from 100°C to 400°C (as shown in Figure 3). In contrast, the methanol mole fraction reduces uniformly. The water mole fraction curve is almost flat. It is due to the higher methanol conversion rate than water, as represented in Figure 4. It also shows a sharp increase in hydrogen selectivity with the rise in the reaction temperature. At 400°C, the hydrogen selectivity and mole fraction obtained are 67.28% and 0.64, respectively. A similar surge in hydrogen mole fraction and the feed conversion rate was observed in the findings of Zaccara *et al.* (2020) (Zaccara *et al.* 2020) and Pashchenko (2021) (Pashchenko 2021), where the higher temperature gives a higher hydrogen mole fractions.

**Table 2.** Case study-based validation from earlier published data.

	Feedstock	Operating Temperature	Catalyst Type	Experimental Methanol Conversion (%)	Simulated Methanol Conversion (%)	Difference (%)
Case 1	Methanol	220°C	CuO/ZnO/Al <sub>2</sub> O <sub>3</sub>	57.6	59.3	+2.95
Case 2	Methanol	270°C	CuO/ZnO/Al <sub>2</sub> O <sub>3</sub>	87.1	89.0	+2.18
Case 3	Methanol	290°C	CuO/ZnO/Al <sub>2</sub> O <sub>3</sub>	96.0	94.8	-1.25
Case 4	Methanol	325°C	1Pt/3In <sub>2</sub> O <sub>3</sub> /CeO <sub>2</sub>	98.7	98.2	-0.51

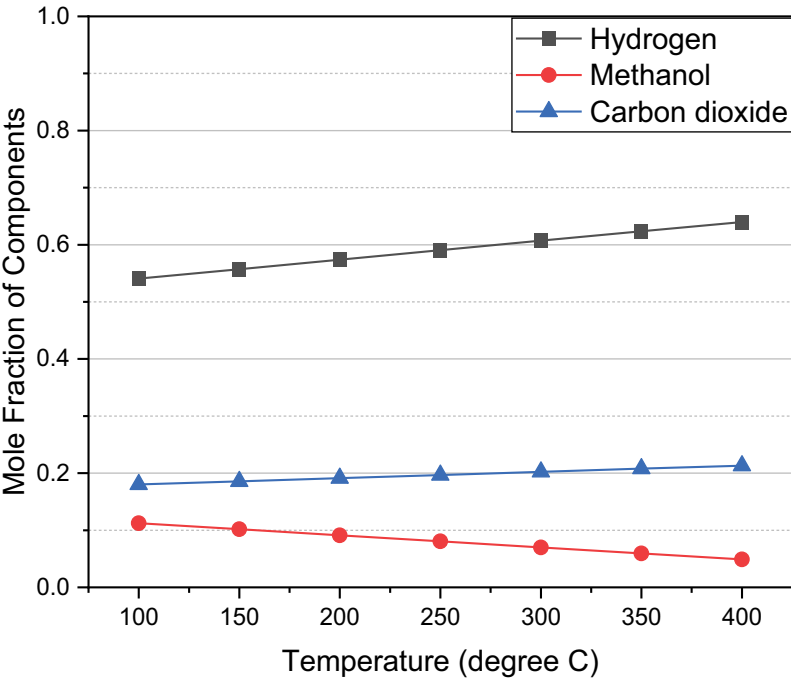


Figure 3. Mole fraction of components at elevated temperature.

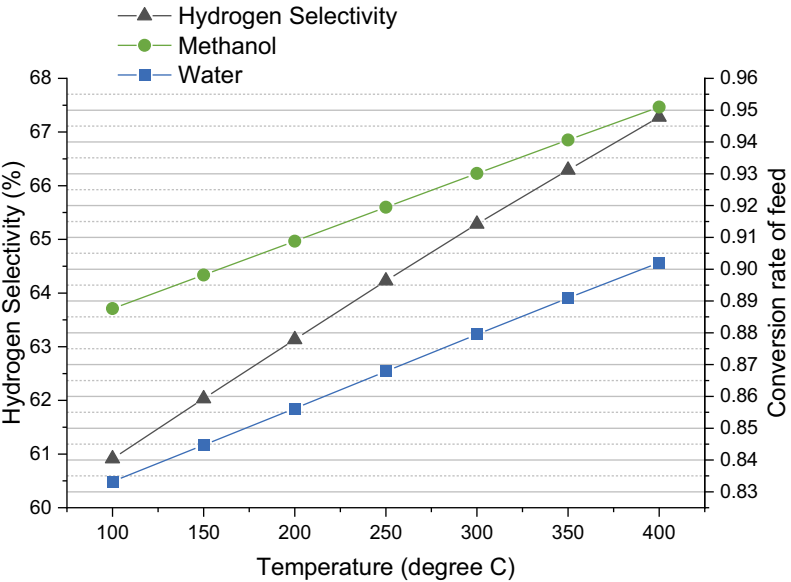


Figure 4. Hydrogen selectivity and feed conversion rate curves at elevated temperatures.

**Effect of reactor pressure**

The reactor pressure plays a crucial role in the reaction pathway. The impact of reactor pressure on the mole fraction of various components is presented in Figure 5. At the same time, the variation in hydrogen selectivity, methanol, and water conversion rate is shown in Figure 6. The

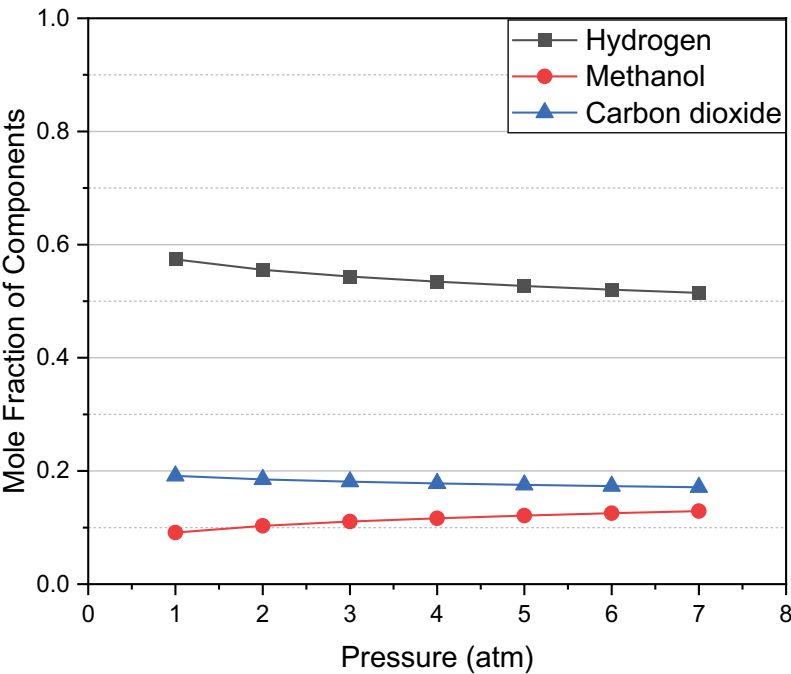


Figure 5. Mole fraction of components curves versus reactor pressure.

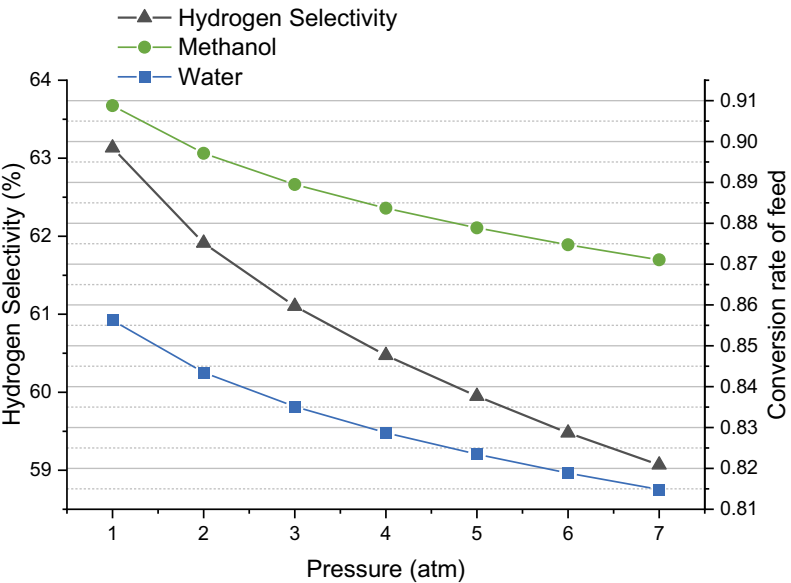


Figure 6. Hydrogen selectivity and feed conversion rate curves versus reactor pressure.

methanol-reforming is performed at 400°C, and a methanol-to-water ratio of 1.2 is chosen for simulation.

The curves in Figure 5 show a drastic reduction in hydrogen and water mole fractions, whereas the methanol mole fraction rises when the reactor pressure goes from 1 atm. to 7 atm. The variation indicates that methanol contributes more pressure to produce hydrogen.

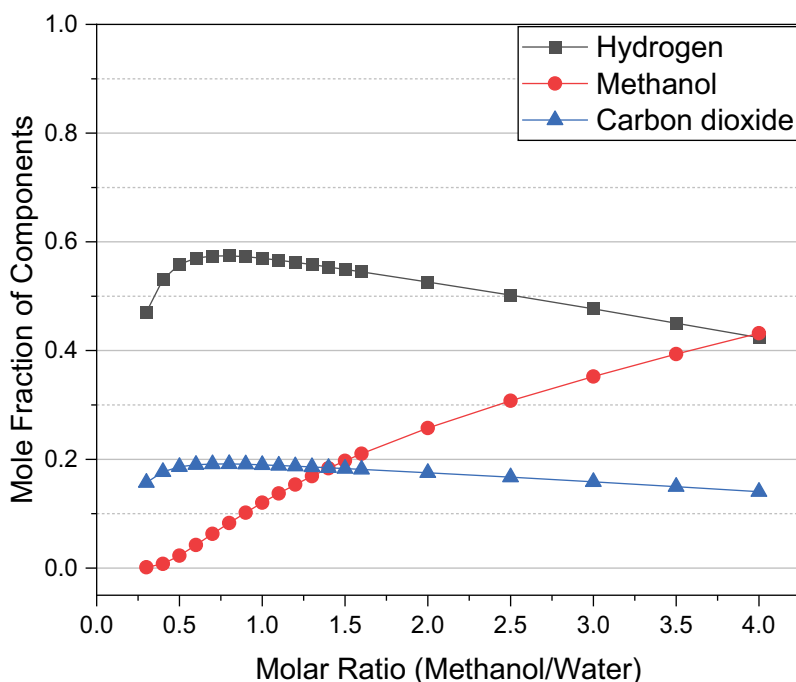


Figure 7. Mole fraction of components versus methanol-to-water molar ratio.

Similarly, in Figure 6, a decline is observed in hydrogen selectivity from 63.13% to 59.07% when reactor pressure is raised from 1 atm. to 7 atm. This decline can also be observed in methanol and water conversion rates, as both curves show a fall with a pressure rise. Similar results were followed by Hakandai, Sidik Pramono, and Aziz (2022), where a higher reactor pressure sharply reduces the hydrogen mole fraction in the syngas (Hakandai, Sidik Pramono, and Aziz 2022).

### Effect of methanol-to-water molar ratio

The molar ratio is another factor that is very important in increasing the concentrations of desired products. An optimum molar ratio generates the best results desirable in a reaction. In contrast, a higher or lower product concentration can shift the reaction pathway, thus, reducing the concentration of the desired product. Figures 7 and figure 8 present the influence of the molar ratio (methanol-to-water ratio, from 0.3 to 4.0) on the mole fraction of products, hydrogen selectivity, and feed conversion rates. The reactor pressure and temperature maintained for the simulation process are 1 atm. and 400°C, respectively.

From Figures 7 and 8, it is observed that a methanol-to-water (M-to-W) molar ratio between 0.5 and 1.5 favors hydrogen production at a much higher rate. The hydrogen mole fraction initially increases from 0.47 (at 0.3 M-to-W) to 0.574 (at 0.9 M-to-W) and then reduces to 0.42 at 4.0 M-to-W molar ratio, respectively. In contrast, the methanol conversion rate shows a sharp decline and a relative reduction is observed in hydrogen selectivity. However, the water conversion rate rises in this case because the water concentration reduces with the increase in the M-to-W molar ratio. The behavior of reduction in hydrogen production due to molar ratio was also observed by Unlu and Hilmioglu (2020), where a higher molar ratio significantly reduced hydrogen production (Unlu and Hilmioglu 2020).

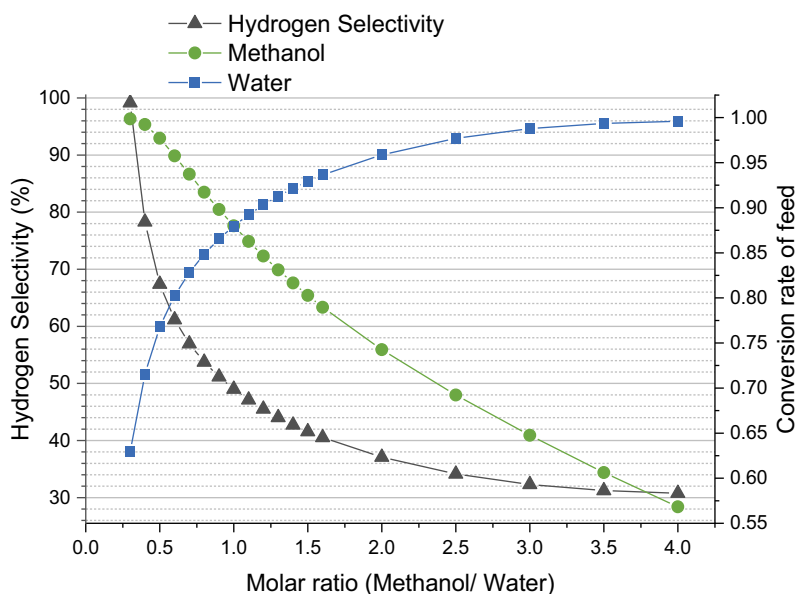


Figure 8. Hydrogen selectivity and feed conversion rate curves versus methanol-to-water molar ratio.

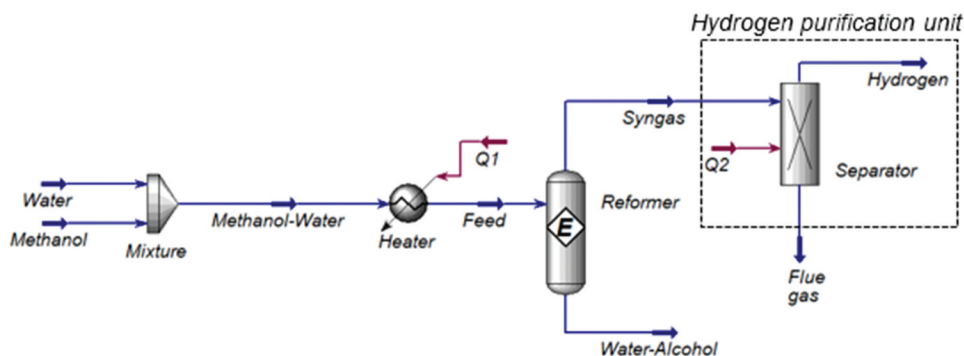


Figure 9. Hydrogen purification unit attached to the methanol-reforming simulation model.

Table 3. Upstream and downstream composition of various components at different reaction points.

	Feed	Syngas	Flue gas	Hydrogen
Methanol	0.5000	0.0680	0.1981	0.0000
Carbon monoxide	0.0000	0.0001	0.0002	0.0000
Hydrogen	0.0000	0.6566	0.0000	9.9999
Water	0.5000	0.0565	0.1644	0.0000
Carbon dioxide	0.0000	0.2188	0.6373	0.0001

### Hydrogen purification

Hydrogen purity is essential for its applications. Fuel cells require hydrogen at a purity level of 99.9%. Therefore, a hydrogen purification unit (as shown in Figure 9) is attached to the methanol-reforming simulation model.

Hydrogen is present in the syngas, one of the products coming out of the outlet of the reformer; therefore, a purification unit is attached to the system. The upstream and



**Table 4.** Stream table of the methanol-reforming model (at a reactor temperature of 400°C, reactor pressure of 1 atm. And the methanol-to-water ratio of 1.0).

	Unit	Feed	Hydrogen	Carbon dioxide	Unreacted Feed
Molecular Weight		25.03	2.016	44.01	24.37
Vapor Fraction		0	1	1	0.1594
Temperature	°C	40	40	45	83
Pressure	atm.	1	1	1	1
Mole Flow	kg-mole/h	120	134.1	44.7	30.6
Mass Flow	kg/h	3003	270.4	1967	765.6
Volume Flow	m <sup>3</sup> /h	3.565	3448	1162	141
Molar Enthalpy	kJ/kg-mole	−26100	426.6	−393100	−253500
Mass Enthalpy	kJ/kg	−10550	211.6	−8931	−10130
Enthalpy Flow	J/h	−31690	57.22	−17570	−7755
Molar Entropy	kJ/kg-mole-°C	42.44	124.4	174.9	72.85
Mass Entropy	kJ/kg-°C	1.696	61.71	3.974	2.911
Molar Density	kg-mole/m <sup>3</sup>	33.66	0.0389	0.0385	33.29
Mass Density	kg/m <sup>3</sup>	842.5	0.0784	1.693	811.2

downstream compositions of various components are mentioned in Table 3, respectively. The other gases in the reformer are carbon dioxide, water vapor, methanol, and carbon monoxide. Hydrogen is separated in the hydrogen purification unit operated at 20 atm. pressure and 40°C. The hydrogen gas from the separator is 99.99% pure and can be used for energy generation in fuel cells.

The stream table (Table 4) of methanol-reforming for hydrogen selectivity and methanol conversion is obtained at a reactor temperature of 400°C, with a reactor pressure of 1 atm. and methanol-to-water molar ratio of 1.0, respectively. The stream table describes the energy and flow rate at different input and output positions; it helps calculate the necessary yield.

During the simulation process, it is observed that the temperature directly influences hydrogen production, while the reactor pressure has adverse impacts. However, the result of the molar ratio mainly depends on the hydrogen content available for conversion in the feedstock.

## Conclusions

The model is developed in Aspen Hysys to simulate methanol-reforming for hydrogen production. The equilibrium reactor with Peng-Robinson thermodynamic model is used to develop a simulation model. The model also contains a separator unit that generates hydrogen gas at a purity level of 99.99%. The simulation is carried out to determine the impact of the reaction temperature, reactor pressure, and methanol-to-water molar ratio (M-to-W) on the hydrogen mole fraction in syngas, hydrogen selectivity, and feed conversion rate (for both methanol and water). The simulation results are as follows:

- A drastic increase in hydrogen mole fraction from 0.54 to 0.64 is observed with a temperature rise from 100°C to 400°C, showing an 18.5% rise in hydrogen mole fraction.
- Similarly, a sharp rise in hydrogen selectivity is obtained with the temperature rise. At the same time, the methanol conversion rate is more than the water conversion rate, which shows more influence of methanol than water in hydrogen production.
- A rise in reactor pressure adversely affects hydrogen concentration, represented by lower hydrogen mole fraction and selectivity at higher pressures.
- Methanol-to-Water (M-to-W) molar ratio plays a crucial role in hydrogen production. The M-to-W ratio between 0.5 and 1.5 showed a higher hydrogen mole fraction.
- The hydrogen selectivity curve also showed the importance of the M-to-W molar ratio. The higher molar ratio results in a lower methanol conversion rate and, thus, a reduction in hydrogen selectivity is observed.

Therefore, the methanol-reforming simulation model efficiently determines the influence of lower temperature range, varied reactor pressure, and different M-to-W molar ratios.

The model can be used to simulate the reforming process to generate hydrogen from other alcohols and lower hydrocarbon compounds. Other parameters that can also be considered for simulation are reactor length and diameter, type of feed, and feed flow rate. These parameters can be used to determine the optimized hydrogen yield in the reactor. It will help design a commercial plant for large-scale hydrogen production through the reforming process.

## Acknowledgements

The authors acknowledge the support of the Department of Mechanical Engineering, Delhi Technological University, for conducting this research.

## Disclosure statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## ORCID

Neeraj Budhraj  <http://orcid.org/0000-0002-9349-4892>

Amit Pal  <http://orcid.org/0000-0003-2979-8127>

R. S. Mishra  <http://orcid.org/0000-0002-2326-0417>

## Data availability statement

All the data used in the manuscript have been appropriately cited with the corresponding reference.

## References

- AlNouss, A., G. McKay, and T. Al-Ansari. 2020. Enhancing waste to hydrogen production through biomass feedstock blending: A techno-economic-environmental evaluation. *Applied Energy* 266 (March):114885. doi:10.1016/j.apenergy.2020.114885.
- Awasthi, M. K., R. K. Rai, S. Behrens, and S. K. Singh. 2021. Low-temperature hydrogen production from methanol over a ruthenium catalyst in water. *Catalysis Science and Technology* 11 (1):136–42. doi:10.1039/d0cy01470b.
- Budhraj, N., A. Pal, and R. S. Mishra. 2023a. Optimizing methanol reforming parameters for enhanced hydrogen selectivity in an Aspen Hysys simulator using response surface methodology. *Energy Technology*. 11 (7):2300203. doi:10.1002/ente.202300203.
- Budhraj, N., A. Pal, and R. S. Mishra. 2023b. Plasma reforming for hydrogen production: Pathways, reactors and storage. *International Journal of Hydrogen Energy*. 48 (7):2467–82. Elsevier Ltd. doi:10.1016/j.ijhydene.2022.10.143.
- Chen, L. N., K. P. Hou, Y. S. Liu, Z. Y. Qi, Q. Zheng, Y. H. Lu, J. Y. Chen, J. L. Chen, C. W. Pao, S. B. Wang, et al. 2019. Efficient hydrogen production from methanol using a single-site Pt1/CeO2 catalyst. *Journal of the American Chemical Society* 141 (45):17995–99. doi:10.1021/jacs.9b09431.
- Du, C., J. Mo, and H. Li. 2015. Renewable hydrogen production by alcohols reforming using Plasma and Plasma-catalytic Technologies: Challenges and opportunities. *Chemical Reviews* 115 (3):1503–42. doi:10.1021/cr5003744.
- Garcia, G., E. Arriola, W. H. Chen, and M. D. de Luna. 2021. A comprehensive review of hydrogen production from methanol thermochemical conversion for sustainability. In *Energy*, Vol. 217. Elsevier Ltd. doi:10.1016/j.energy.2020.119384.
- Giwa, S. O., A. Giwa, and O. Giwa. 2013. Application of Aspen Plus to hydrogen production from alcohols by steam reforming: Effects of reactor temperature. *International Journal of Engineering Research & Technology* 2 (8):648–57. [www.ijert.org](http://www.ijert.org).
- Hakandai, C., H. Sidik Pramono, and M. Aziz. 2022. Conversion of municipal solid waste to hydrogen and its storage to methanol. *Sustainable Energy Technologies and Assessments* 51:51. doi:10.1016/j.seta.2022.101968.

- Kanath, T. K., and N. Ayas. 2021. Simulating the steam reforming of sunflower meal in Aspen Plus. *International Journal of Hydrogen Energy* 46 (57):29076–87. doi:10.1016/j.ijhydene.2020.12.195.
- Kim, D. H., J. H. Kim, Y. S. Jang, and J. C. Kim. 2019. Hydrogen production by oxidative steam reforming of methanol over anodic aluminum oxide-supported Cu-Zn catalyst. *International Journal of Hydrogen Energy* 44 (20):9873–82. doi:10.1016/j.ijhydene.2018.11.009.
- Lee, D. H., and T. Kim. 2013. Plasma-catalyst hybrid methanol-steam reforming for hydrogen production. *International Journal of Hydrogen Energy* 38 (14):6039–43. doi:10.1016/j.ijhydene.2012.12.132.
- Liu, X., Y. Men, J. Wang, R. He, and Y. Wang. 2017. Remarkable support effect on the reactivity of Pt/In<sub>2</sub>O<sub>3</sub>/MO<sub>x</sub> catalysts for methanol steam reforming. *Journal of Power Sources* 364:341–50. doi:10.1016/j.jpowsour.2017.08.043.
- Mohammadidoust, A., and M. R. Omidvar. 2020. Simulation and modeling of hydrogen production and power from wheat straw biomass at supercritical condition through Aspen Plus and ANN approaches. *Biomass Conversion and Biorefinery* 12 (9):3857–73. doi:10.1007/s13399-020-00933-5.
- Pashchenko, D. 2021. Thermochemical waste-heat recuperation as on-board hydrogen production technology. *International Journal of Hydrogen Energy* 46 (57):28961–68. doi:10.1016/j.ijhydene.2020.11.108.
- Qureshi, F., M. Yusuf, H. Kamyab, S. Zaidi, M. Junaid Khalil, M. Arham Khan, M. Azad Alam, F. Masood, L. Bazli, S. Chelliapan, et al. 2022. Current trends in hydrogen production, storage and applications in India: A review. *Sustainable Energy Technologies and Assessments* 53:102677. doi:10.1016/j.seta.2022.102677.
- Ranjekar, A. M., and G. D. Yadav. 2021. Steam reforming of methanol for hydrogen production: A critical analysis of catalysis, processes, and scope. *Industrial and Engineering Chemistry Research* 60 (1):89–113. doi:10.1021/acs.iecr.0c05041.
- Shamsi, M., A. A. Obaid, S. Farokhi, and A. Bayat. 2022. A novel process simulation model for hydrogen production via reforming of biomass gasification tar. *International Journal of Hydrogen Energy* 47 (2):772–81. doi:10.1016/j.ijhydene.2021.10.055.
- Shen, Y., Y. Zhan, S. Li, F. Ning, Y. Du, Y. Huang, T. He, and X. Zhou. 2017. Hydrogen generation from methanol at near-room temperature. *Chemical Science* 8 (11):7498–504. doi:10.1039/c7sc01778b.
- Singh, M., S. A. Salaudeen, B. H. Gilroyed, and A. Dutta. 2022. Simulation of biomass-plastic co-gasification in a fluidized bed reactor using Aspen plus. *Fuel* 319 (March):123708. doi:10.1016/j.fuel.2022.123708.
- Tavares, R., E. Monteiro, F. Tabet, and A. Rouboa. 2020. Numerical investigation of optimum operating conditions for syngas and hydrogen production from biomass gasification using Aspen Plus. *Renewable Energy* 146:1309–14. doi:10.1016/j.renene.2019.07.051.
- Unlu, D., and N. D. Hilmioglu. 2020. Application of aspen plus to renewable hydrogen production from glycerol by steam reforming. *International Journal of Hydrogen Energy* 45 (5):3509–15. doi:10.1016/j.ijhydene.2019.02.106.
- Wang, H., R. Ren, B. Liu, and C. You. 2022. Hydrogen production with an auto-thermal MSW steam gasification and direct melting system: A process modeling. *International Journal of Hydrogen Energy* 47 (10):6508–18. doi:10.1016/j.ijhydene.2021.12.009.
- Wang, F., and G. Wang. 2016. Performance and cold spot effect of methanol steam reforming for hydrogen production in micro-reactor. *International Journal of Hydrogen Energy* 41 (38):16835–41. doi:10.1016/j.ijhydene.2016.07.083.
- Ye, G., D. Xie, W. Qiao, J. R. Grace, and C. J. Lim. 2009. Modeling of fluidized bed membrane reactors for hydrogen production from steam methane reforming with Aspen Plus. *International Journal of Hydrogen Energy* 34 (11):4755–62. doi:10.1016/j.ijhydene.2009.03.047.
- Zaccara, A., A. Petrucciani, I. Martino, T. A. Branca, S. Dettori, V. Iannino, V. Colla, M. Bampaou, and K. Panopoulos. 2020. Renewable hydrogen production processes for the off-gas valorization in integrated steelworks through hydrogen intensified methane and methanol syntheses. *Metals* 10 (11):1–24. doi:10.3390/met10111535.
- Zhang, R., C. Huang, L. Zong, K. Lu, X. Wang, and J. Cai. 2018. Hydrogen production from methanol steam reforming over TiO<sub>2</sub> and CeO<sub>2</sub> pillared clay supported Au catalysts. *Applied Sciences (Switzerland)* 8 (2). doi: 10.3390/app8020176.
- Zhao, J., R. Shi, Z. Li, C. Zhou, and T. Zhang. 2020. How to make use of methanol in green catalytic hydrogen production? *Nano Select* 1 (1):12–29. doi:10.1002/nano.202000010.

## Modeling of water surface profile in non-prismatic compound channels

Vijay Kaushik<sup>a,\*</sup>, Munendra Kumar<sup>a</sup>, Bandita Naik<sup>b</sup> and Abbas Parsaie<sup>c</sup>

<sup>a</sup>Department of Civil Engineering, Delhi Technological University, Delhi 110042, India

<sup>b</sup>Department of Civil Engineering, Methodist College of Engineering, Hyderabad 500001, India

<sup>c</sup>Department of Water Structures, Faculty of Water and Environmental Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran

\*Corresponding author. E-mail: vijaykaushik\_2k20phdce01@dtu.ac.in

 VK, 0000-0002-7270-4520

### ABSTRACT

Estimating the water surface elevation of river systems is one of the most complicated tasks in formulating hydraulic models for flood control and floodplain management. Consequently, utilizing simulation models to calibrate and validate the experimental data is crucial. HEC-RAS is used to calibrate and verify the water surface profiles for various converging compound channels in this investigation. Based on experimental data for converging channels ( $\theta = 5^\circ$ ,  $9^\circ$ , and  $12.38^\circ$ ), two distinct flow regimes were evaluated for validation. The predicted water surface profiles for two relative depths ( $\beta = 0.25$  and  $0.30$ ) follow the same variational pattern as the experimental findings and are slightly lower than the observed values. The MAPE for the simulated and experimental results is less than 3%, indicating the predicted HEC-RAS value performance and accuracy. Therefore, our findings imply that in the case of non-prismatic rivers, the proposed HEC-RAS models are reliable for predicting water surface profiles with a high generalization capacity and do not exhibit overtraining. However, the results demonstrated that numerous variables impacting the water surface profile should be carefully considered since this would increase the disparities between HEC-RAS and experimental data.

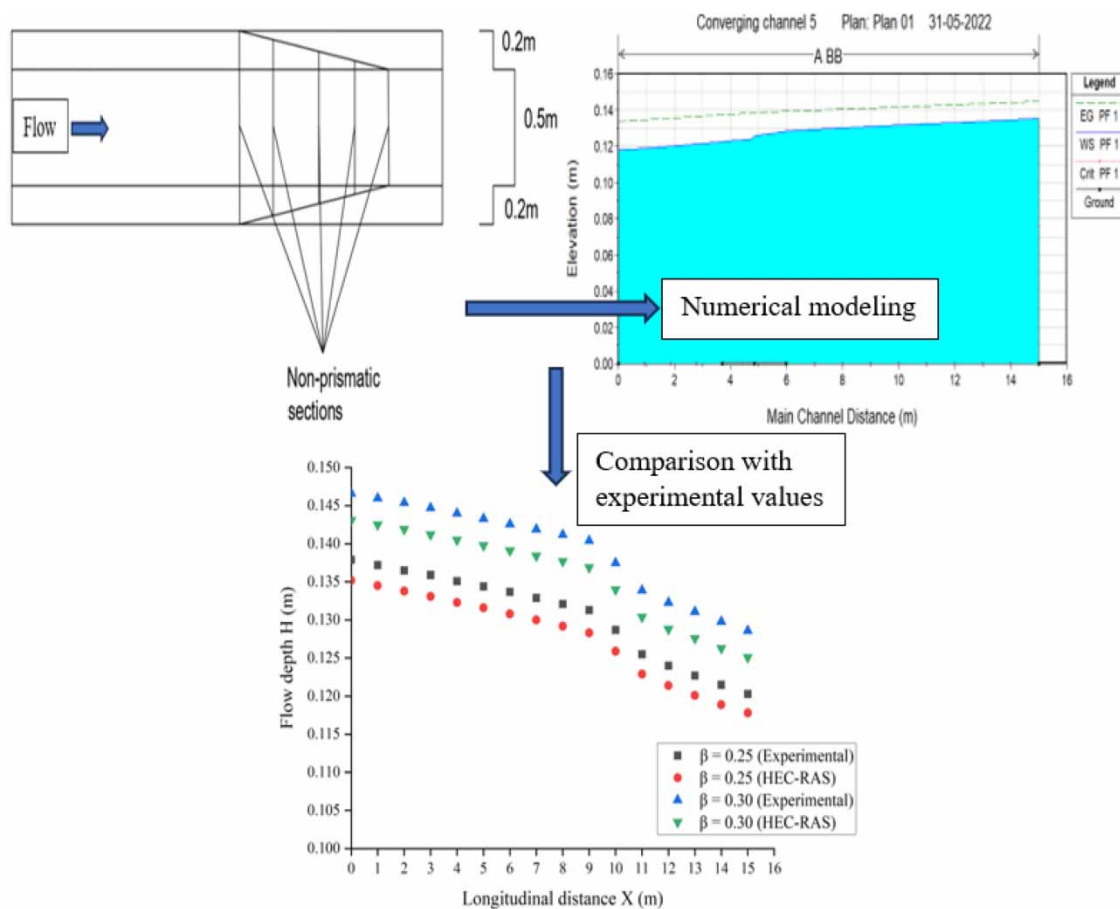
**Key words:** compound channel, converging floodplains, HEC-RAS modeling, water surface profile

### HIGHLIGHTS

- In this article, research was conducted for the non-prismatic compound channel with converging floodplains, utilizing the HEC-RAS software.
- The findings depict the HEC-RAS models are accurate for forecasting the water surface profile of non-prismatic rivers, have a high capacity for generalization, and do not display any signs of overexertion.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## GRAPHICAL ABSTRACT



## NOTATIONS

- $B$  total width of the compound channel  
 $b$  total width of the main channel  
 $H$  flow depth  
 $h$  height of the main channel  
 $L$  converging length  
 $S$  longitudinal bed slope  
 $\beta$  relative flow depth  $[(H - h)/H]$   
 $\theta$  converging angle

## ACRONYMS

- |         |   |
|---------|---|
| 2D      | two dimensional   |
| ADV     | acoustic Doppler velocimeter                            |
| ANFIS   | adaptive neuro-fuzzy inference system                   |
| ANN     | artificial neural network                               |
| DCM     | divided channel method                                  |
| DEM     | digital elevation model                                 |
| GEP     | gene expression programming                             |
| GLM     | generalized linear model                                |
| GMDH    | group method of data handling                           |
| HEC-RAS | Hydrologic Engineering Centre's – River Analysis System |
| ISM     | independent subsection method                           |
| LDM     | lateral distribution technique                          |
| MAE     | mean absolute error                                     |

MAPE	mean absolute percentage error
MLM	machine learning model
MLPNN	multi-layer perceptron neural network
NF-GMDH	neuro-fuzzy group method of data handling
RF	random forest
RMSE	root mean squared error
SCM	single channel method
SVM	support vector machine

## INTRODUCTION

Increased human settlements, buildings, and activities along river floodplains have resulted in severe repercussions during natural river floods due to the global population rise. River floods cause massive human casualties as well as economic damage. Flood catastrophes account for a third of all-natural disaster damages worldwide; flooding accounts for half of all fatalities, with trend analysis revealing that these percentages have dramatically grown (Berz 2000). Flood protection needs to predict the conveyance capacity of natural streams precisely. When the amount of water running through a channel exceeds the waterway's capacity, it results in flooding. Consequently, the requirement for precise flow parameter prediction during flood conditions to limit damage and save lives and property has piqued the interest of academics and engineers in recent years. Various methodologies and procedures have been used to aid precise measurement and forecast of river discharge, velocity distribution, shear stress distribution, and water surface level during overbank flows. Compound channels are the most common river feature during overbank flow. During the course of a river's flow, the geometry of the floodplain changes, resulting in a compound channel that is either converging or diverging. It is more challenging to replicate flow in a non-prismatic compound channel because more momentum is carried from the main channel to the floodplains. Sellin (1964), Myers & Elsayy (1975), Knight *et al.* (2010), and Khatua *et al.* (2012) have explored the flow models of straight and meandering prismatic two-stage channels, but little is known about non-prismatic compound channels. A converging channel shape causes the flow on floodplains to rise, while the flow on floodplains expanding is reduced (James & Brown 1977). Compound channels with symmetrically declining floodplains were studied by Bousmar & Zech (2002), Bousmar *et al.* (2004), Rezaei (2006), and Rezaei & Knight (2009) and found the extra loss of head and transfer of momentum from the main channel to floodplains. Asymmetric geometry with a greater convergence rate was examined by Proust *et al.* (2006). A greater convergence angle (22°) results in increased mass transfer and head loss. Chlebek *et al.* (2010) studied the flow behavior of skewed, two-stage converging, and diverging channels. A new experiment on converging compound channels was done by Rezaei & Knight (2011), Yonesi *et al.* (2013), and Naik & Khatua (2016) that yielded significantly more precise results than previously accessible. In their study, Das *et al.* (2018) sought to enhance the conventional independent subsection method (ISM) for the estimation of flow magnitudes and velocities in the upper and lower main channels. The calculated results demonstrate the method's ability to accurately forecast the discharge distributions in both the floodplain and main channel. Das & Khatua (2018a) constructed a multivariable regression model that accounts geometric and hydraulic characteristics in order to estimate the Manning's roughness coefficient for non-prismatic compound channels. In their study, Das & Khatua (2018b) explored a numerical approach for estimating water surface elevations in compound channels with converging floodplains, using the momentum balancing concept. The findings derived from the simulation exhibit a strong concurrence with the empirical datasets. Das *et al.* (2020) used artificial neural network (ANN) and adaptive neuro-fuzzy inference system (ANFIS) methodologies to forecast the discharge in compound channels with converging and diverging geometries. The discharge is affected by many key input factors, including the friction factor ratio, hydraulic radius ratio, relative flow depth, and bed slope. The ANFIS model has superior performance in comparison to the ANN model. In their study, Das *et al.* (2022) proposed a non-linear multivariable regression model for estimating discharge distribution in diverging compound channels. This model utilizes geometric non-dimensional factors. The model that has been built demonstrates improved results in terms of statistical analysis when compared to earlier methodologies. Naik *et al.* (2022) proposed a novel equation by GEP using the non-dimensional variables to predict the water surface profile in converging compound channels. Kaushik & Kumar (2023a) used machine learning methodologies to predict the water surface profile of a compound channel with converging floodplains, using a blend of geometry and flow characteristics. Additionally, the researchers Kaushik & Kumar (2023b) have used gene expression programming (GEP) as a methodology to develop an innovative equation for compound channels with converging



floodplains. This equation serves to quantify the boundary shear force transmitted by floodplains. In their study, [Kaushik & Kumar \(2023c\)](#) used the support vector machine (SVM) method to estimate the water surface profile of compound channels with shrinking floodplains. This was achieved by using non-dimensional geometric characteristics. The outcomes of this study suggest that the water surface profile created by the SVM has a significant level of agreement with both the observed data and the results obtained from prior investigations. In their study, [Bijanvand \*et al.\* \(2023\)](#) employed various soft computing models, namely the multi-layer perceptron neural network (MLPNN), group method of data handling (GMDH), neuro-fuzzy group method of data handling (NF-GMDH), and SVM, to make predictions on the surface elevation of water in compound channels with converging and diverging floodplains. The findings indicated that all of the used models exhibited satisfactory performance. Nevertheless, the SVM model had the most favorable performance, as shown by its strong statistical indicators. The influence of channel shape and flow characteristics on the water surface profile in non-prismatic compound channels has received little attention. As a result, exact water surface profile modeling is necessary to detect flooded regions, enhancing flood mitigation and risk management studies.

Over the past several decades, much work has gone into using 2D and 3D modeling to enhance the estimation of water levels and velocities in rivers. Still, minimal work was done on non-prismatic streams. Calculation techniques like single channel method (SCM) and divided channel method (DCM) are incorporated in software like HEC-RAS and MIKE 11. For the whole segment, the SCM uses the same velocity. The DCM divides the cross-section into zones with varied flow characteristics, such as the main channel and floodplains. According to [Wormleaton \*et al.\* \(1982\)](#), the SCM underestimates conveyance capacity, whereas the DCM overestimates compound channel capacity. [Wormleaton & Merrett \(1990\)](#) offered a simple change to enhance DCM estimation, while [Ackers \(1992\)](#) experimentally corrected the DCM.

The lateral distribution technique (LDM) proposed by [Wark \*et al.\* \(1990\)](#) and the approach proposed by [Shiono & Knight \(1991\)](#) were created as alternate and more sophisticated methods. Like a quasi-2D model, these two techniques are based on the same equations and determine the lateral velocity distribution in the cross-section. In natural and artificial channels, HEC-RAS, a widely used hydraulic model developed by the U.S. Army Corps of Engineers, calculates water surface elevation and other flow characteristics in 1D/2D dimensions with progressively altering dimensions for steady and turbulent flow ([Brunner 2016](#)). HEC-RAS enables sediment transport/mobile bed calculations and water temperature modeling ([Arcement & Schneider 1989](#); [Brunner 2016](#)). The stability of the HEC-RAS modeling was assessed by the use of model verification and validation techniques, which included comparing the model's predictions with experimental findings or actual field data. Stability of a model is determined when the numerical outputs closely align with the experimental findings or actual field data, exhibiting a consistent pattern of fluctuation. River hydraulics and other river-related phenomena have been substantially enhanced by using computer programs in recent years. [Leandro \*et al.\* \(2009\)](#) give extensive information on the most often used hydraulic models and their advantages and disadvantages for open channel modeling. Globally, computer hydraulic models are being used for flood defence planning in vulnerable locations to help better understand flood size and frequency trends and help prepare for future flood scenarios ([Liu & Merwade 2018](#)). The HEC-RAS model was used in various studies to estimate flow characteristics in the main channel and floodplain under different climatic circumstances. [Ramesh \*et al.\* \(2000\)](#) estimated roughness for open channel flow using an optimization technique with boundary conditions as constraints. The HEC-RAS model was calibrated using Manning's  $n$  roughness coefficient, as reported by [Hicks & Peacock \(2005\)](#) and [Kuriqi & Ardiclioglu \(2018\)](#) when applied to river analysis. [Timbadiya \*et al.\* \(2011\)](#) developed an integrated hydrodynamic model with MIKE11 to calibrate Manning's  $n$  roughness in assessing the sensitivity of flow resistance for the Tapi River in India. [Mowinckel \(2011\)](#) used the HEC-RAS to increase the flood conveyance capacity of an artificial San Jose Creek in Goleta, California. This assessment allowed us to recommend a revised channel design to accommodate a 100-year flood better while reducing harm to the surrounding region. [Parhi \*et al.\* \(2012\)](#) calibrated the channel roughness coefficient along the Mahanadi River in Odisha using the HEC-RAS. [Boulomytis \*et al.\* \(2017\)](#) discovered that using Manning's  $n$  roughness coefficients for various hydraulic models causes inaccuracies in inflow predictions for the Bashar River. Rivers must be studied since they are often used for agriculture or hydropower generation. An accurate estimation of the water surface elevation is necessary to construct and deploy the appropriate flood control structures and produce proper flow behavior ([Kuriqi \*et al.\* 2019](#)). In order to lessen the dependence on arbitrary static friction coefficients, [Klipalo \*et al.\* \(2022\)](#) conducted research by measuring and presenting actual data collected via quantitative testing. Full-scale field testing was conducted as part of this research to measure the frictional resistance produced between filled polypropylene bulk bags and seven typical bedding

surfaces. Coefficients of static friction are used to convey the results of testing each interaction scenario. Three machine learning models (MLMs), including random forest (RF), ANN, and generalized linear model (GLM), were used by [Avand \*et al.\* \(2022\)](#) to investigate the impact of the spatial resolution of the DEMs 12.5 m (ALOS PALSAR) and 30 m (ASTER) on the precision of flood probability prediction. The findings show that, regardless of the employed MLM and irrespective of the statistical model used to measure the performance accuracy, resolving the DEM alone cannot substantially impact the accuracy of flood probability prediction. In contrast, the elements that affect floods in this area the most include height, precipitation, and distance from the river. The alterations in the water surface profile and flow velocity brought on by the bridge structural arrangement were studied by [Ardiclioglu \*et al.\* \(2022\)](#). For this reason, five flow discharges, four distinct bridge spans' water surface profiles, and flow velocities above and downstream of the bridge were examined. The HEC-RAS model was used to conduct the study both numerically and experimentally. At the bridge's upstream section, the average velocities calculated by HEC-RAS were vastly exaggerated. The average downstream and upstream measured velocities in the various apertures showed linear connections.

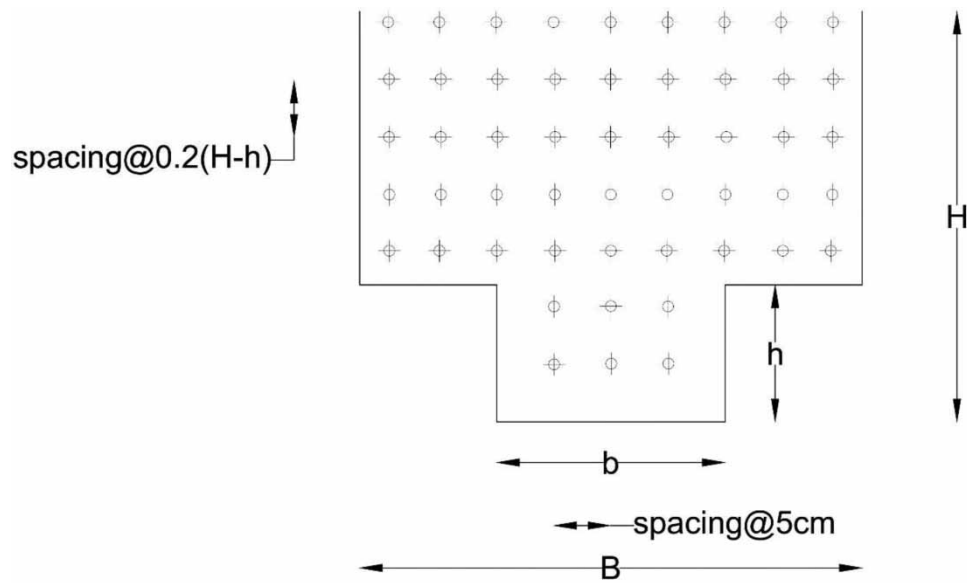
The aim of the present study is to validate the experimental results of the water surface profile of a two-stage channel with narrowing floodplains using one-dimensional numerical models. The approach proposed in this article uses the HEC-RAS to enhance numerical modeling. The dataset used in this study effort to complete the simulation effectively was gathered from the work of [Naik & Khatua \(2016\)](#), which was done on a variety of converging compound channels and provided the basis for this work. In order to compare and validate the experimental results, the same boundary conditions, cross-section data, and flow parameters were used. Finally, the simulated water surface level results were analyzed and compared to existing experimental data to evaluate and validate the findings.

## MATERIALS AND METHODS

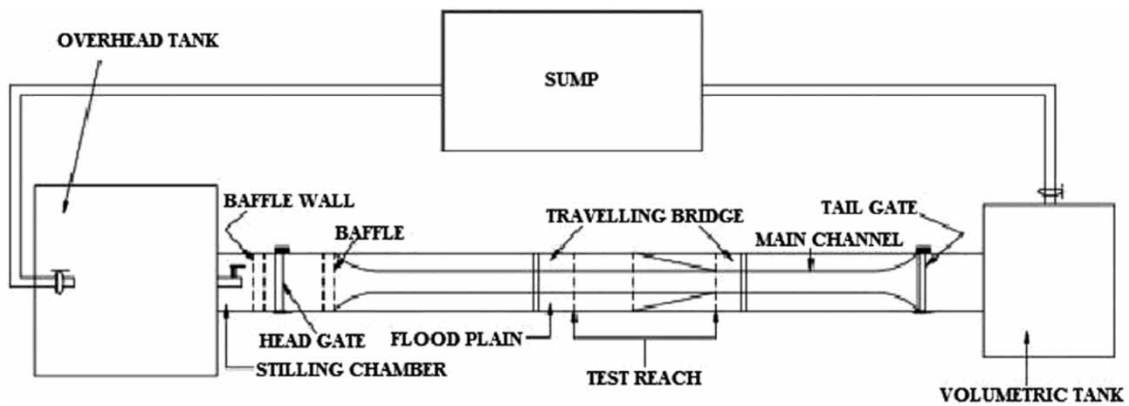
### Physical modeling

A series of experiments were conducted in a concrete flume 15 m long, 0.9 m wide, and 0.5 m deep with three different converging compound channels. The converging portion of the channel was constructed with the help of the Perspex sheet. The converging angles of the channel were 12.38°, 9°, and 5°, respectively, keeping the geometry constant. The non-prismatic compound channel has converging lengths of 0.84, 1.26, and 2.28 m, respectively. The subcritical flow regime was attained in several conditions of the two-stage channel with a longitudinal bed slope of 0.0011. The main channel and the floodplain subsections of these compound channels exhibit uniform roughness. Manning's  $n$  value of 0.01 was selected for smooth main channel and floodplain surfaces with trowel finishes ([Subramanya 2015](#)). Based on data collected from in-bank and overbank flows in the floodplains and main channel, Manning's  $n$  value variation was estimated in the converging section of the channel. This system recirculates the water supply by pumping it from an underground sump to a reservoir in the experimental channel. The rectangular notch has been surrounded by adjustable vertical gates and flow strengtheners. The removable flume tailgates help maintain a consistent flow over the test reach. A volumetric tank at the end of the channel is fitted with v-notch which has been used to measure the flow rate from the channel. The water collected in the volumetric tank goes back to the underground sump. The experimental channel has a limit of discharge that cannot increase beyond 0.055 m<sup>3</sup>/s. The geometric characteristics:  $B$  is the total width of the compound channel,  $b$  is the width of the main channel,  $h$  is the main channel depth,  $H$  is the flow depth at any discharge, and a cross-section of a two-stage channel are described in [Figure 1](#).

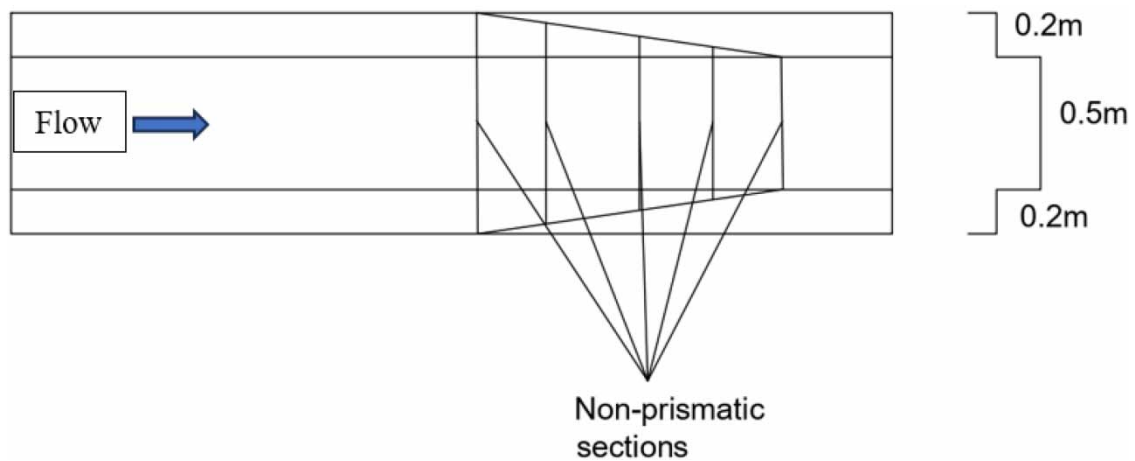
[Figure 2](#) illustrates the experimental setup from the top. The plan view of non-prismatic cross-sections of [Naik & Khatua \(2016\)](#) channels is shown in [Figure 3](#). Each point on the channel's plan could be accessed for measuring as part of the compound channel design. A moveable bridge could be used to collect the measurements. The research relies heavily on the channel's width ratio and aspect ratio. The flow velocity at the grid locations (shown in [Figure 1](#)) was measured using a pitot-static tube with a diameter of 4.77 mm. The order of maximum velocity for a given flow path was determined using a flow detector with a minimum least count of 0.1°. Use a circular scale and pointer configuration on the flow direction sensor to measure the pitot tube leg angle concerning the channel longitudinal direction. When combining the longitudinal velocity plot with the volumetric tank collection, the total discharge computed was within  $\pm 3\%$  of the actual data. This study used velocity data and a semi-log plot to predict channel bed and wall shear stresses. The boundary shear stresses were calculated using



**Figure 1** | Cross-section of a compound channel.



**Figure 2** | Experimental setup.



**Figure 3** | Plan view of non-prismatic sections.

Patel (1965) relationship and manometer measurements of Preston tube head differences. Shear values were corrected by comparing them to the equivalent values computed using the energy gradient technique.

Thus, the results were always within 3% of the actual value. According to laboratory data analysis, the pitot tube calculated tractive stresses are more accurate than ADV. For one thing, measuring velocity at the boundary with ADV is never trustworthy. In addition, ADV has specific limits for measuring the velocity near the bed and top surface. It can penetrate up to 5 cm below the top edge. Consequently, the micro-ADV down probe could not reach a distance of 5 cm from the free surface. It cannot measure the velocity beyond 2 m/s. In order to measure the transient decrease, a pitot tube was used near the bed and top surface. The U-tube manometer fitted along with the pitot tube measures the pressure difference values up to certain values. Verification of the validity of this approach was carried out using the energy gradient methodology (Naik & Khatua 2016).

### Numerical modeling

To simulate a constantly changing flow, researchers opted for the HEC-RAS model, which works by computing the Saint-Venant equations (Equations (1) and (2)):

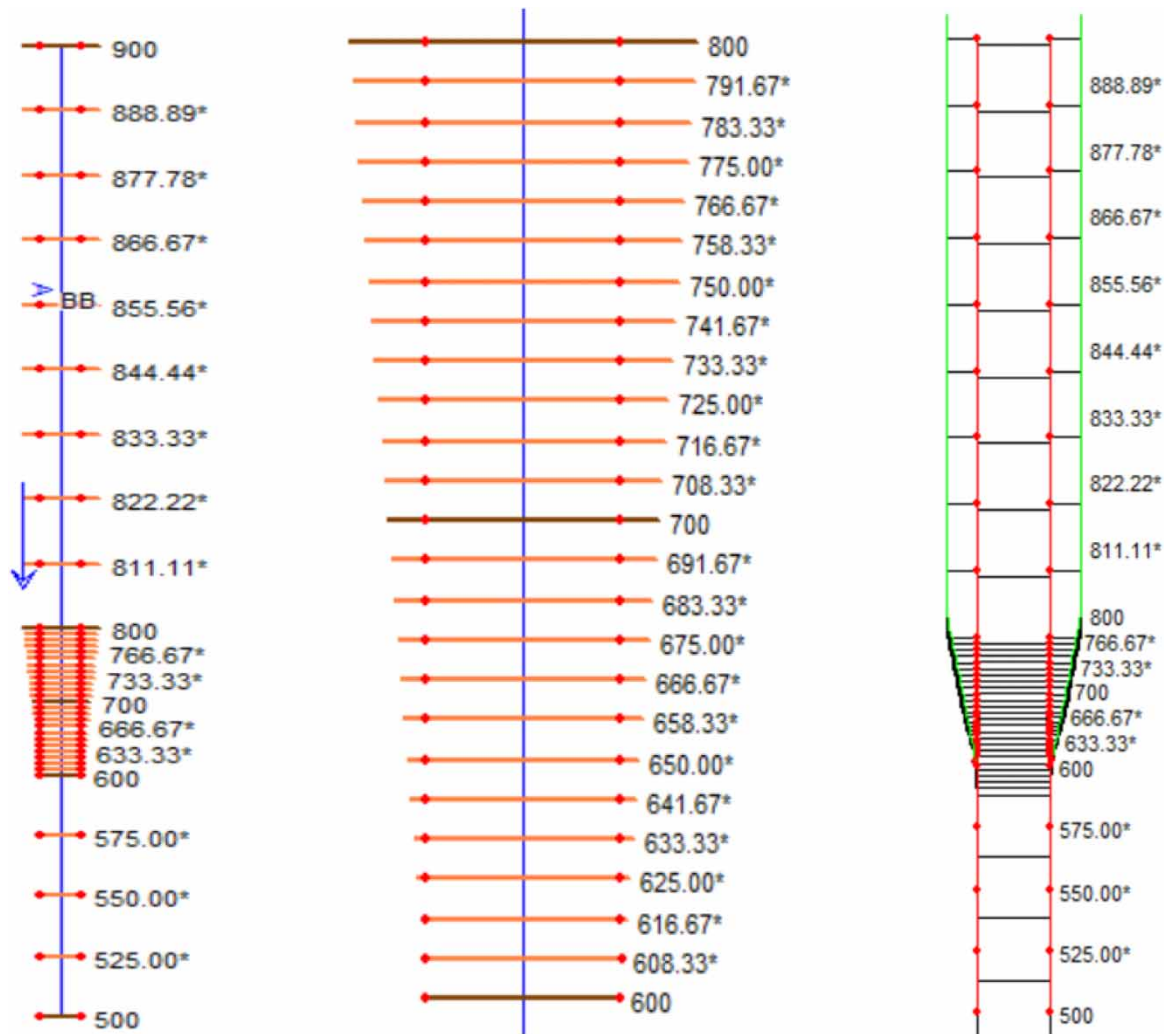
$$\frac{\partial A}{\partial t} + \frac{\partial Q}{\partial x} = q_l \quad (1)$$

$$\frac{\partial Q}{\partial t} + \frac{\partial \left( \frac{Q^2}{A} \right)}{\partial x} + gA \frac{\partial H}{\partial x} + gA(S_o - S_f) = 0 \quad (2)$$

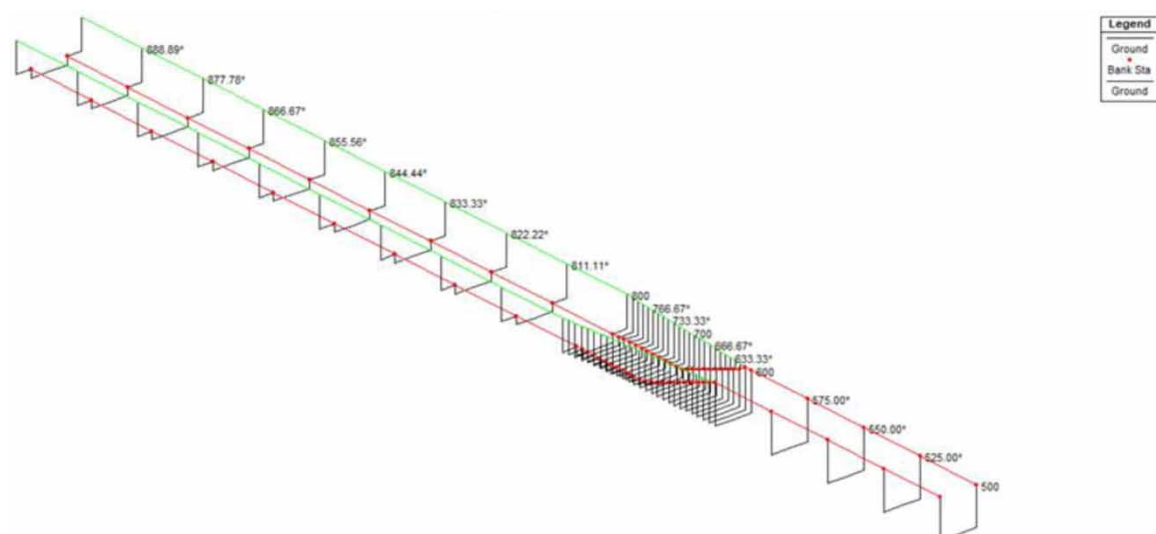
Equations (1) and (2) represent continuity and momentum equations, respectively, in which  $H$  denotes the elevation of water level,  $t$  and  $x$  indicate temporal and longitudinal coordinates,  $q_l$  represents lateral inflow, which in our instance is 0;  $S_o$  is the bed slope of channel, and  $S_f$  denotes the slope of total energy line or friction slope. These equations were computed with the help of the finite-difference approach based on a four-point implicit box (Brunner 2016). Even though the finite-difference principle has limitations in transitioning between the subcritical and supercritical flow regimes, this procedure necessitates a new solution strategy for each flow state. However, the previous constraint may be avoided by employing the HEC-RAS model's mixed-flow regime option; by doing so, HEC-RAS can propose patching solutions in the river reach's sub-zones (Hicks & Peacock 2005; Timbadiya *et al.* 2011; Brunner 2016). The HEC-RAS model was calibrated by keeping the same experimental data, such as channel dimensions, boundary conditions, longitudinal bed slope, discharge data, and roughness coefficient values used in the experimental procedure. Manning's  $n$  value of 0.011 was inserted at the simulation stage for the main channel and floodplains based on the smooth trowel finish used in the physical modeling. Due to the almost continuous and uniform flow conditions at the study river reach, the downstream boundary condition was reset to its upstream condition. Figures 4 and 5 depict the plan, and 3D views of converging compound channel geometry created in the HEC-RAS with the 1 m resolution DEM was calibrated using measured experimental values within the channel reach. Figure 4 displays stations 900 and 500, which represent the upstream and downstream of the waterway, respectively. In contrast, stations 800, 700, and 600 correspond to the beginning, intermediate, and final segments of the converging section, respectively. Stations 900 and 500 are considered to be stationary; however, stations 800, 700, and 600 are not stationary, for all three converging angles of 5°, 9°, and 12.38°. The variation in distance between stations 800 and 600 is seen to be 2.28, 1.26, and 0.84 m, corresponding to converging angles of 5°, 9°, and 12.38°, respectively. The use of the same nomenclature for stations was implemented in order to mitigate any misunderstanding during the simulation procedure. The remaining stations located in both the straight and converging parts of the two-stage channel are considered interpolated stations. The HEC-RAS estimated water depth was then compared to the experimental water depth to validate the developed non-prismatic model of the compound channel.

### RESULTS AND DISCUSSION

Figure 6 illustrates the stage–discharge correlation for a relative depth of  $\beta = 0.25$  at several locations along the converging section of different channels. These locations include the upstream of the channel, the start, middle, and end of the converging portion. The converging channels have varying degrees of convergence, denoted by  $\theta = 5^\circ$ ,  $9^\circ$ , and  $12.38^\circ$ . The increase in discharge leads to a corresponding rise in flow depth. Nevertheless, a little decrease in the rate of increase occurs beyond the point when the river reaches its maximum capacity,

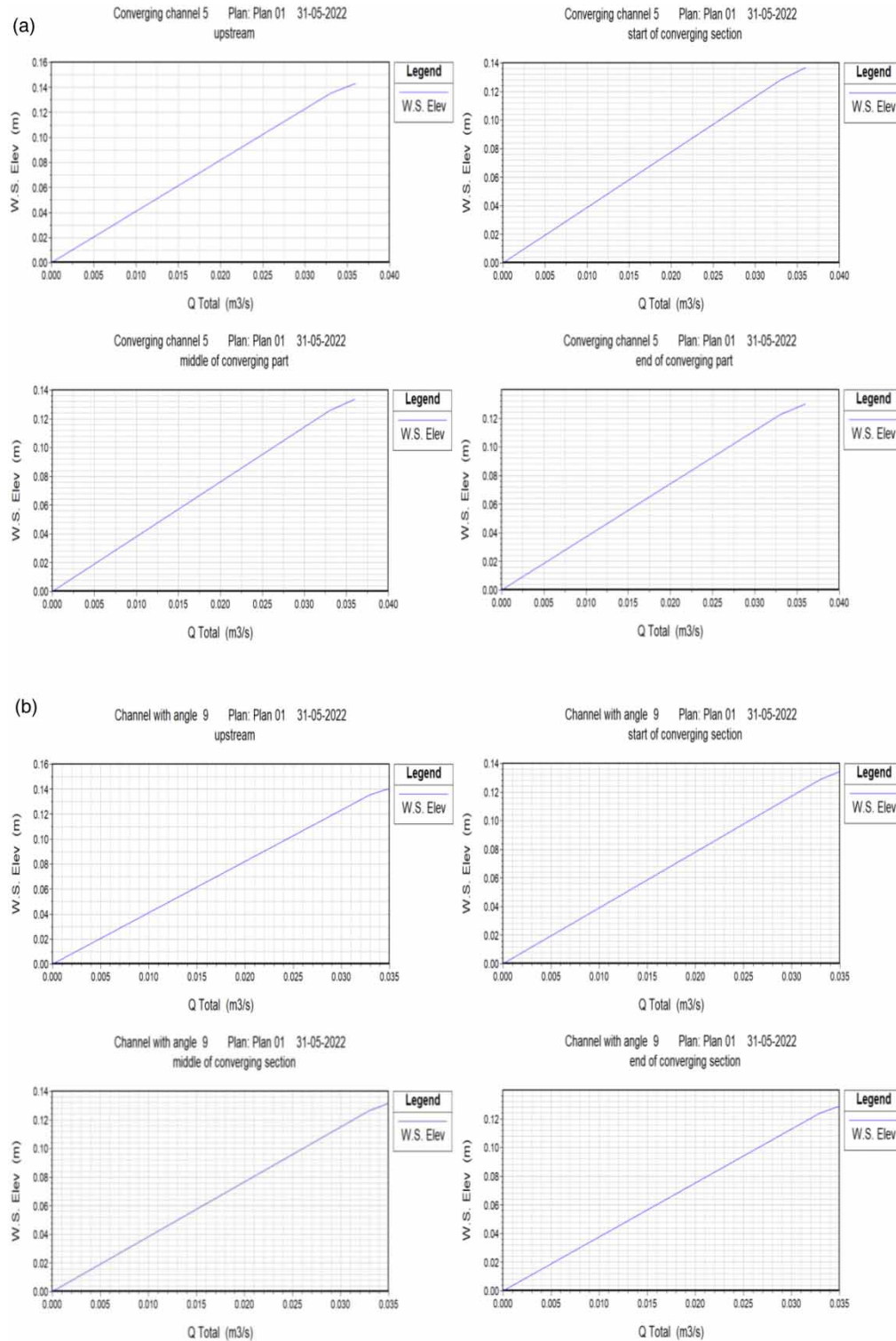


**Figure 4** | Plan view of a converging compound channel in HEC-RAS.



**Figure 5** | 3D view of the converging compound channel in HEC-RAS.





**Figure 6** | Stage discharge relationship for various converging channels: (a)  $\theta = 5^\circ$ , (b)  $\theta = 9^\circ$ , and (c)  $\theta = 12.38^\circ$ . (continued.).

mostly as a result of the interaction and subsequent transfer of momentum between the primary channel and the adjacent floodplains. The reduction in flow depth occurs in the converging section due to the convergence of channel geometry, while maintaining the same discharge. Furthermore, it has been shown that an increase in the angle of floodplain convergence at a given stage is accompanied by a corresponding rise in the flow rate. The empirical evidence obtained from stage–discharge correlations supports the notion that power functions exhibit consistency when applied to datasets containing big values.



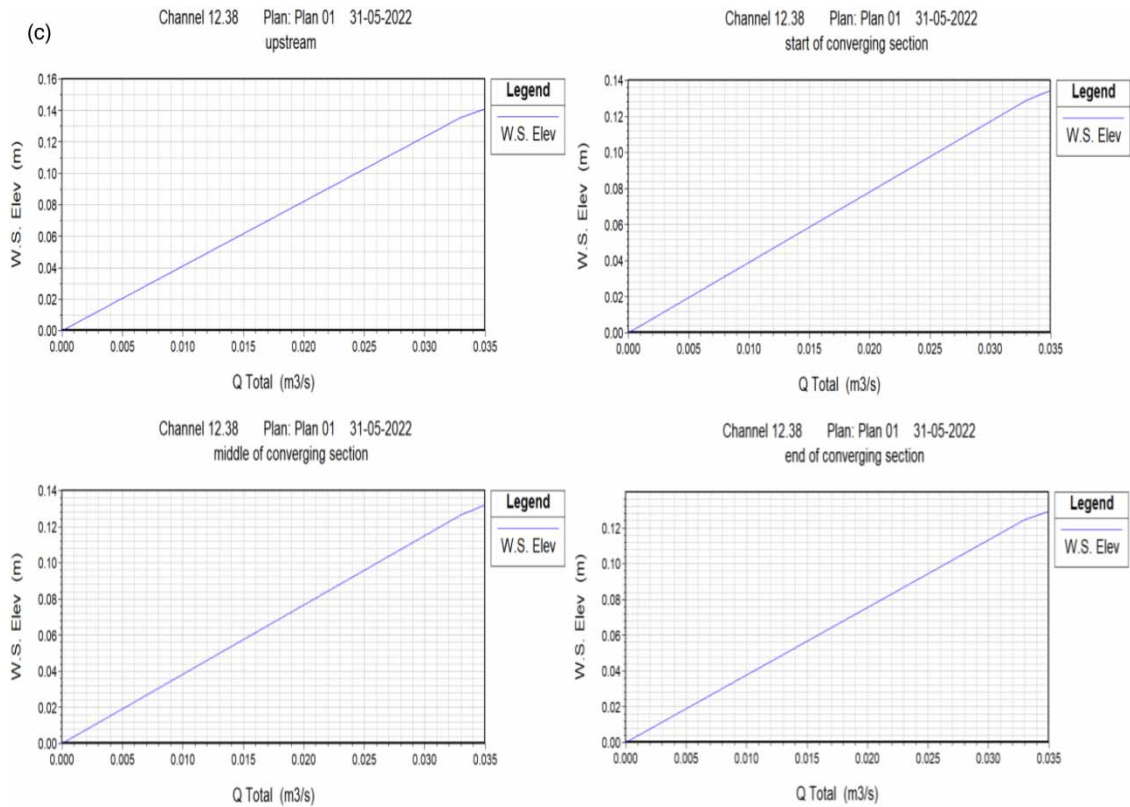


Figure 6 | Continued.

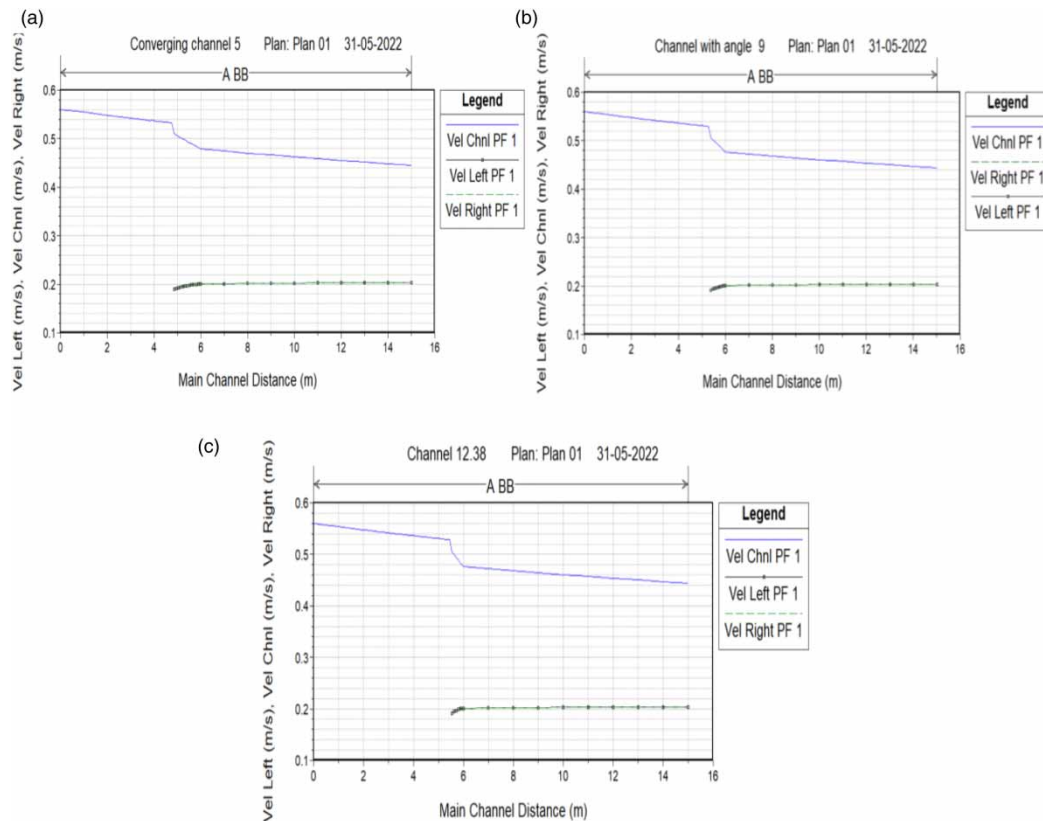
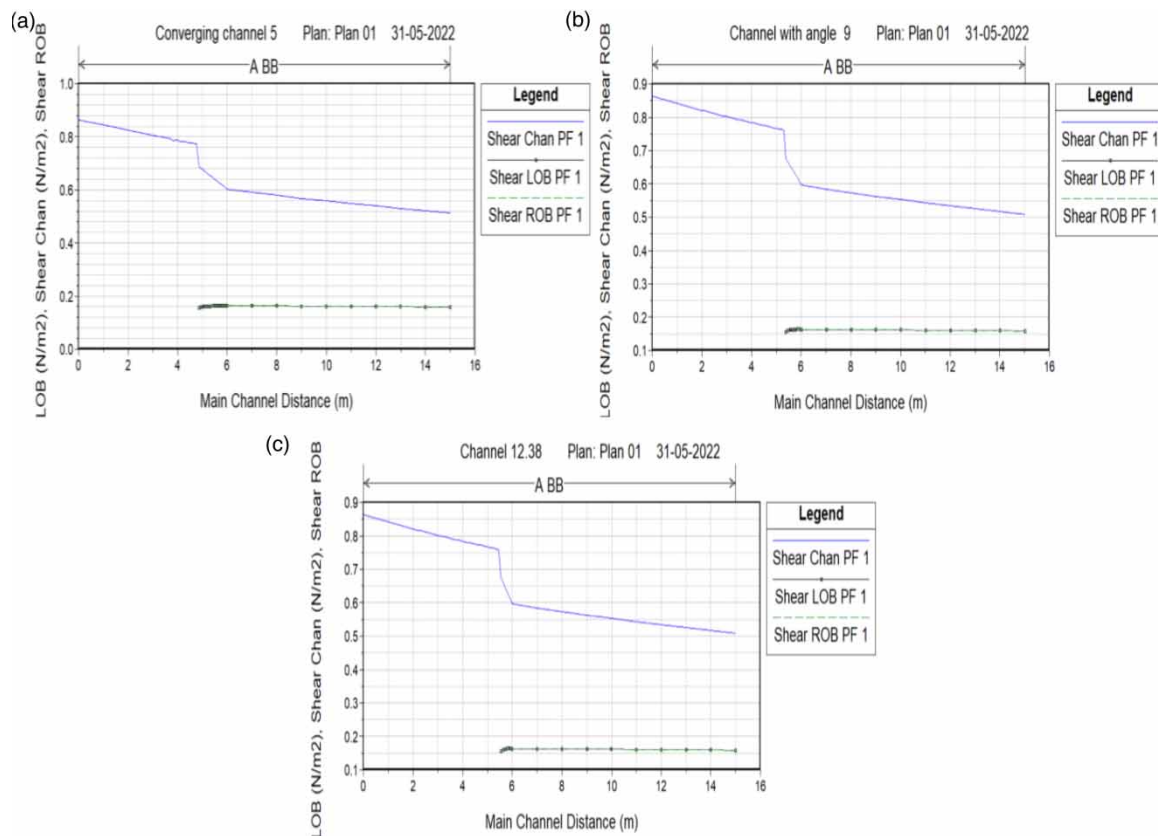
Figure 7 | Variation of velocity with longitudinal distance for relative depth  $\beta = 0.25$  for various converging channels: (a)  $\theta = 5^\circ$ , (b)  $\theta = 9^\circ$ , and (c)  $\theta = 12.38^\circ$ .

Figure 7 demonstrates the changes in velocity as a function of longitudinal distance for different converging channels, specifically at a relative depth  $\beta$  of 0.25. The terms ‘Vel chnl PF’, ‘Vel left PF’, and ‘Vel right PF’ refer to the mean velocities of the whole channel, the floodplain on the right side, and the floodplain on the left side, respectively. The variable ‘PF’ represents the flow depth at which the simulation has been done. The study determined that the average velocity of the channel was greater than the average velocity observed in both the right and left floodplains. The velocities are seen to begin at a uniform channel distance of 15 m, indicating the upstream portion of the compound channel. On the contrary, a numerical value of zero signifies the downstream portion of the compound channel. The commencement of the building of the converging segment of the canal took place at a distance of 6 m. After the start of the converging section, the velocities inside the floodplain zones undergo a decrease due to the convergence of the channel morphology, which enables the transfer of momentum from the floodplains to the main channel. An incremental rise in the velocity of the channel is noted in the prismatic section. However, the velocity experiences a significant spike in the converging section as a result of the abrupt constriction in the channel’s shape. An increase in the convergence angle of the non-prismatic compound channel leads to a significant rise in velocity in the converging section.

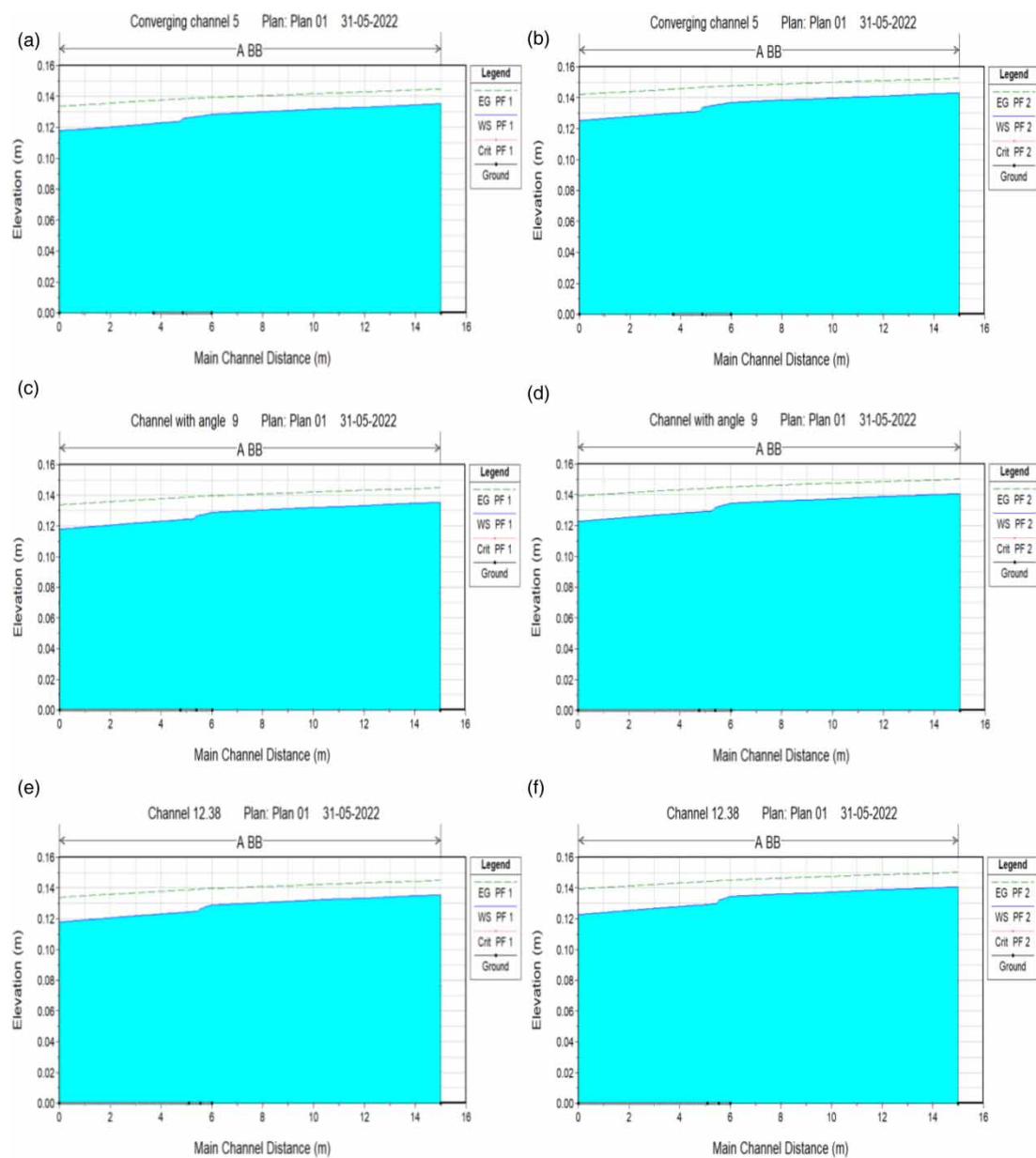
Figure 8 exhibits the fluctuation of shear stress in relation to the longitudinal distance for converging channels with a relative depth  $\beta$  of 0.25. The expressions ‘Shear Chan PF1’, ‘Shear LOB PF1’, and ‘Shear ROB PF1’ are used to refer to the shear stress experienced by the whole channel, the left floodplain (left bank), and the right floodplain (right bank), respectively. The variable ‘PF1’ represents the flow depth, denoted by  $\beta = 0.25$ , at which the simulation was performed. The study determined that the shear stress inside the channel exhibited a larger magnitude compared to the shear stress seen on the right and left floodplains. The initiation of shear stresses is seen at a uniform channel distance of 15 m, indicating the upstream portion of the compound channel. On the other hand, a numerical value of zero signifies the downstream portion of the compound channel. The commencement of the converging section of the canal occurred at a distance of 6 m. After the start of the converging section, the velocities inside the floodplain zones undergo a decrease due to the convergence of the channel morphology. This convergence enables the transfer of momentum, specifically in terms of shear, from



**Figure 8** | Variation of shear stress with longitudinal distance for relative depth  $\beta = 0.25$  for various converging channels: (a)  $\theta = 5^\circ$ , (b)  $\theta = 9^\circ$ , and (c)  $\theta = 12.38^\circ$ .

the floodplains to the main channel. The magnitude of the channel shear stress exhibits an increasing trend as it progresses throughout the length of the flow. The convergence angle of the non-prismatic half of the channel has a direct impact on the shear stress increment seen in all three converging channels. As the convergence angle rises, the shear stress increment becomes more pronounced, mostly owing to increased resistance from the channel boundaries.

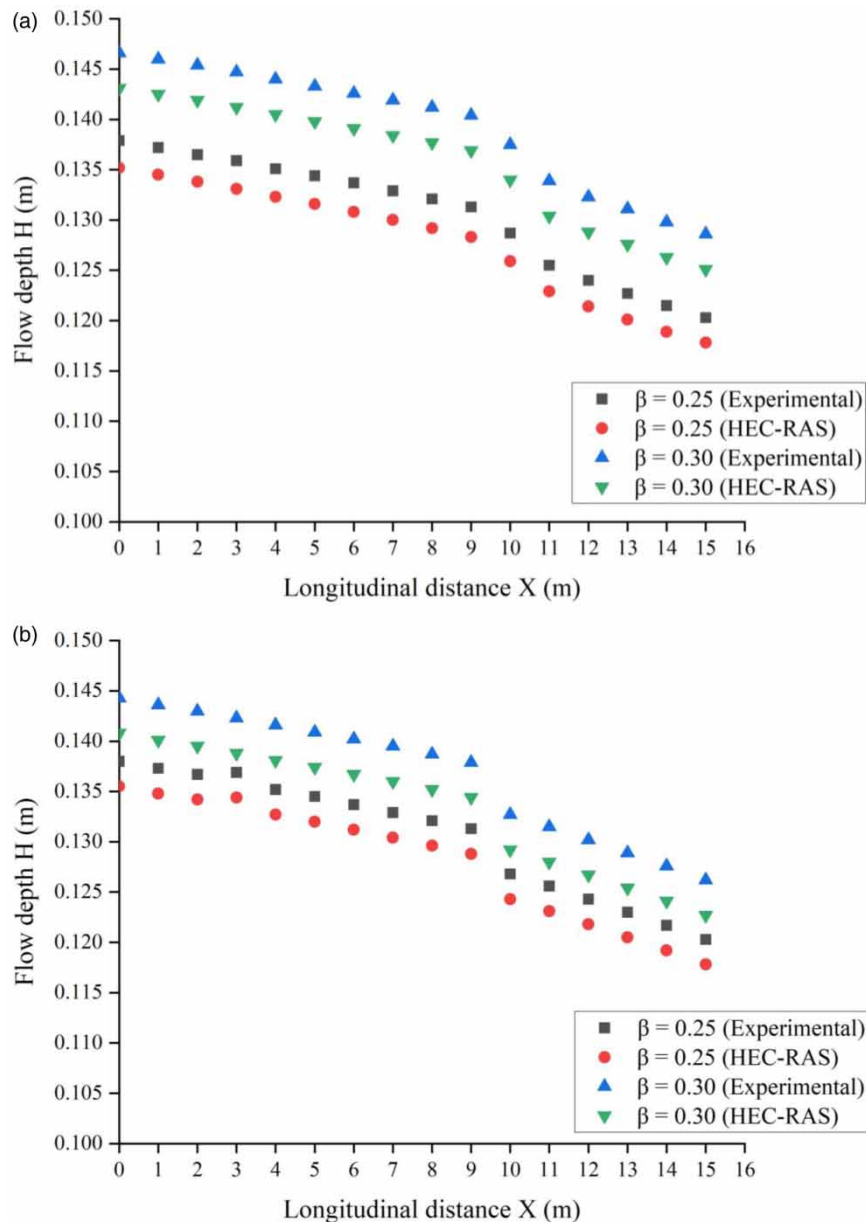
Figure 9 displays the longitudinal water surface profile for different relative depths ( $\beta = 0.25, 0.30$ ) at various locations along the converging part of channels with different convergence angles ( $\theta = 5^\circ, 9^\circ, 12.38^\circ$ ). The locations include the upstream of the channel, the start, middle, and end of the converging segment. In the prismatic section of the flume, the water surface profile remains constant. However, it should be noted that throughout the converging section of the flume, there is a noticeable decline in the water level. This decline may be attributed to the acceleration of the flow, particularly in the latter half of the transition. In the lower section of the flume, the flow exhibits a mostly consistent pattern, but with occasional fluctuations. The magnitude of flow depth diminishes as the relative distance increases, and this decrease is more noticeable at larger converging



**Figure 9** | Longitudinal water surface profile for different relative depths and converging angles: (a)  $\beta = 0.25$ ,  $\theta = 5^\circ$ ; (b)  $\beta = 0.30$ ,  $\theta = 5^\circ$ ; (c)  $\beta = 0.25$ ,  $\theta = 9^\circ$ ; (d)  $\beta = 0.30$ ,  $\theta = 9^\circ$ ; (e)  $\beta = 0.25$ ,  $\theta = 12.38^\circ$ ; and (f)  $\beta = 0.30$ ,  $\theta = 12.38^\circ$ .

angles. The reason for this phenomenon may be attributed to the convergence of the channel geometry inside the non-prismatic part of the compound channel.

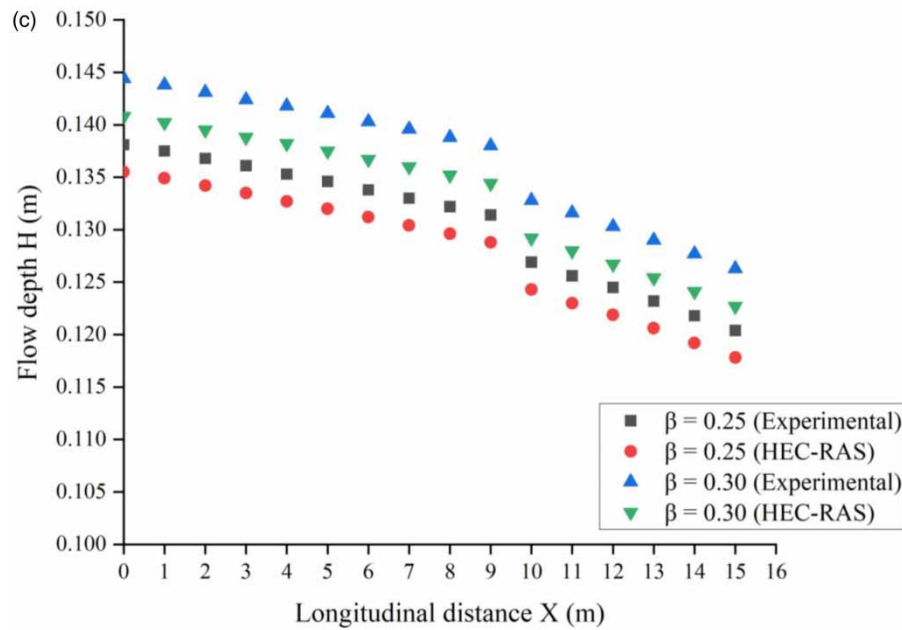
In Figure 10, a comparison is made between the empirically obtained flow depth and the flow depth calculated using HEC-RAS for non-prismatic compound channels with converging floodplain angles of  $\theta = 5^\circ$ ,  $9^\circ$ , and  $12.38^\circ$ . The simulated values exhibit a similar pattern of variance to that found in the experimental data. The observed flow depths in the experiment exhibit a modest elevation compared to the values predicted by the



**Figure 10** | Comparison of experimental and HEC-RAS simulated values of flow depth for different converging channels: (a)  $\theta = 5^\circ$ , (b)  $\theta = 9^\circ$ , and (c)  $\theta = 12.38^\circ$ . (continued.).

HEC-RAS model. This disparity becomes more pronounced as the relative depths increase. The variability in flow depth measurements exhibits an upward trend as the degree of floodplain convergence increases.

Different forms of error assessment, including the coefficient of determination ( $R^2$ ), root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE), are examined using established equations to conduct further assessments on the precision of the simulated flow depths produced by HEC-RAS. Tables 1–3 provide a comprehensive investigation of statistical errors pertaining to flow depths in



**Figure 10** | Continued.

several converging compound channels. The findings indicate that the values of  $R^2$  for all three converging compound channels are more than 0.90, while the values of RMSE are less than 0.20. The MAPE for both the simulated and experimental findings is below 3%, suggesting a high level of performance and accuracy in the anticipated HEC-RAS results. Consequently, the HEC-RAS models that have been provided demonstrate a

**Table 1** | Statistical error analysis for converging channel,  $\theta = 5^\circ$

Parameters	Converging channel, $\theta = 5^\circ$			
	Relative depth, $\beta = 0.25$		Relative depth, $\beta = 0.30$	
	Experimental flow depth, $H$	HEC-RAS flow depth, $H$	Experimental flow depth, $H$	HEC-RAS flow depth, $H$
Range	0.1379–0.1203	0.1352–0.1178	0.1466–0.1286	0.1431–0.1251
$R^2$	0.946	0.954	0.932	0.932
RMSE	0.183	0.183	0.195	0.195
MSE	0.0335	0.0335	0.038	0.038
MAE	0.131	0.128	0.139	0.136
MAPE	2.10	2.10	2.52	2.52

**Table 2** | Statistical error analysis for converging channel,  $\theta = 9^\circ$

Parameters	Converging channel, $\theta = 9^\circ$			
	Relative depth, $\beta = 0.25$		Relative depth, $\beta = 0.30$	
	Experimental flow depth, $H$	HEC-RAS flow depth, $H$	Experimental flow depth, $H$	HEC-RAS flow depth, $H$
Range	0.1380–0.1203	0.1355–0.1178	0.1443–0.1262	0.1408–0.1227
$R^2$	0.952	0.952	0.940	0.940
RMSE	0.183	0.183	0.191	0.191
MSE	0.0335	0.0335	0.0365	0.0365
MAE	0.131	0.128	0.137	0.133
MAPE	1.92	1.92	2.56	2.56



**Table 3** | Statistical error analysis for converging channel,  $\theta = 12.38^\circ$ 

Parameters	Converging channel, $\theta = 12.38^\circ$			
	Relative depth, $\beta = 0.25$		Relative depth, $\beta = 0.30$	
	Experimental flow depth, $H$	HEC-RAS flow depth, $H$	Experimental flow depth, $H$	HEC-RAS flow depth, $H$
Range	0.1381–0.1204	0.1355–0.1178	0.1444–0.1263	0.1408–0.1227
$R^2$	0.953	0.953	0.940	0.940
RMSE	0.183	0.183	0.191	0.191
MSE	0.0335	0.0335	0.0365	0.0365
MAE	0.131	0.128	0.137	0.133
MAPE	1.99	1.99	2.63	2.63

dependable methodology for forecasting the water surface profile in compound channels with converging floodplains. These models possess a notable capacity for adaptation and do not display any signs of excessive exertion.

## CONCLUSIONS

In the present study, one-dimensional models have been made to simulate the water surface profile of a compound channel with converging floodplains using the HEC-RAS. Two relative depths ( $\beta = 0.25$  and  $0.30$ ) and three converging angles ( $\theta = 5^\circ$ ,  $9^\circ$ , and  $12.38^\circ$ ) were investigated. Flow depth rises as discharge increases up to bankfull depth, but beyond bankfull depth, a modest decrease in depth was seen at all converging angles owing to interaction and momentum transfer between the main channel and floodplains. Due to the convergence of the channel geometry, the flow depth decreases with the length of the channel, and the same tendency has been seen for greater relative depths and varied floodplain convergence angles. The velocity and boundary shear stress followed the same trend of variation and observed a sharp rise in the converging portion of the compound channel. The flow regime is subcritical for both prismatic and non-prismatic reaches of a compound channel. The HEC-RAS projected water surface profile is slightly lower than the experimental values but follows the same trend as the observed water surface profile. The MAPE for flow depth computed experimentally, and HEC-RAS simulated is less than 3% for all three converging channels, showing the model's high performance and accuracy. It was observed that the estimated results are affected by bed slope, velocity distribution, flow resistance, secondary currents, and shear stress distribution. Localized variations in channel shape and 2D effects due to the curvature of a channel may also affect the water surface profile. The models developed in the study can have a practical application to non-prismatic rivers such as the River Main in Northern Ireland, the Brahmaputra River in India, and other similar rivers. The findings of the study will be useful in the design of flood control and diversion structures and thereby reducing economic as well as human losses. The present study was focused on non-prismatic compound channels with smooth floodplains. In terms of future study, it would be interesting to investigate the water surface profile under overbank flow circumstances with rough floodplains to improve the comprehensive flood defence plans.

## ACKNOWLEDGEMENTS

The authors would like to express their most profound appreciation to the anonymous reviewers for their time spent on this paper.

## DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

## CONFLICT OF INTEREST

The authors declare there is no conflict.



## REFERENCES

- Ackers, P. 1992 *Hydraulic design of two-stage channels*. *Proceedings of the Institution of Civil Engineers – Water, Maritime and Energy* **96**(4), 247–257.
- Arcement, G. J. & Schneider, V. R. 1989 *Guide for Selecting Manning's Roughness Coefficients for Natural Channels and Flood Plains*, Vol. I. United States Government Printing Office, Washington, DC. <https://doi.org/10.3133/wsp2339>.
- Ardiclioglu, M., Hadi, A. M. W. M., Periku, E. & Kuriqi, A. 2022 *Experimental and numerical investigation of bridge configuration effect on hydraulic regime*. *International Journal of Civil Engineering* **20**(8), 981–991. <https://doi.org/10.1007/s40999-022-00715-2>.
- Avand, M., Kuriqi, A., Khazaei, M. & Ghorbanzadeh, O. 2022 *DEM resolution effects on machine learning performance for flood probability mapping*. *Journal of Hydro-Environment Research* **40**, 1–16. <https://doi.org/10.1016/j.jher.2021.10.002>.
- Berz, G. 2000 *Flood disasters: lessons from the past – worries for the future*. *Proceedings of the Institution of Civil Engineers – Water, Maritime and Energy* **142**(1), 3–8.
- Bijanvand, S., Mohammadi, M. & Parsaie, A. 2023 *Estimation of water's surface elevation in compound channels with converging and diverging floodplains using soft computing techniques*. *Water Supply* **23**(4), 1684–1699. <https://doi.org/10.2166/ws.2023.079>.
- Boulomytis, V. T. G., Zuffo, A. C., Dalfré, F. J. G. & Imteaz, M. A. 2017 *Estimation and calibration of Manning's roughness coefficients for ungauged watersheds on coastal floodplains*. *International Journal of River Basin Management* **15**(2), 199–206. <https://doi.org/10.1080/15715124.2017.1298605>.
- Bousmar, D. & Zech, Y. 2002 Periodical turbulent structures in compound channels. In *River Flow International Conference on Fluvial Hydraulics*, Louvain-la-Neuve, Belgium, pp. 177–185.
- Bousmar, D., Wilkin, N., Jacquemart, J. H. & Zech, Y. 2004 *Overbank flow in symmetrically narrowing floodplains*. *Journal of Hydraulic Engineering* **130**(4), 305–312.
- Brunner, G. W. 2016 *HEC-RAS River Analysis System (Trans: Center HE)*, 5.0 edn. U.S. Army Corps of Engineers, Davis, CA. Available from: <https://www.hec.usace.army.mil/software/hecras/documentation.aspx>.
- Chlebek, J., Bousmar, D., Knight, D. W. & Sterling, M. A. 2010 Comparison of overbank flow conditions in skewed and converging/diverging channels. In *River Flows International Conference*, pp. 503–511.
- Das, B. S. & Khatua, K. K. 2018a *Flow resistance in a compound channel with diverging and converging floodplains*. *Journal of Hydraulic Engineering* **144**(8), 04018051. [https://doi.org/10.1061/\(ASCE\)HY.1943-7900.0001496](https://doi.org/10.1061/(ASCE)HY.1943-7900.0001496).
- Das, B. S. & Khatua, K. K. 2018b *Numerical method to compute water surface profile for converging compound channel*. *Arabian Journal for Science and Engineering* **43**(10), 5349–5364. <https://doi.org/10.1007/s13369-018-3161-y>.
- Das, B. S., Devi, K., Proust, S. & Khatua, K. K. 2018 *Flow distribution in diverging compound channels using improved independent subsection method*. In *River Flow 9th International Conference on Fluvial Hydraulics*. Vol. **40**, No. 05068, p. 8. <https://doi.org/10.1051/e3sconf/20184005068>.
- Das, B. S., Devi, K., Khuntia, J. R. & Khatua, K. K. 2020 *Discharge estimation in converging and diverging compound open channels by using adaptive neuro-fuzzy inference system*. *Canadian Journal of Civil Engineering* **47**(12), 1327–1344. <https://doi.org/10.1139/cjce-2018-0038>.
- Das, B. S., Devi, K., Khuntia, J. R. & Khatua, K. K. 2022 *Flow distributions in a compound channel with diverging floodplains*. *River Hydraulics: Hydraulics, Water Resources, and Coastal Engineering* **2**, 113–125.
- Doncker, L. D., Troch, P., Verhoeven, R., Bal, K., Meire, P. & Quintelier, J. 2009 *Determination of the Manning roughness coefficient influenced by vegetation in the river Aa and Biebrza river*. *Environmental Fluid Mechanics* **9**(5), 549–567.
- Hicks, F. E. & Peacock, T. 2005 *Suitability of HEC RAS for flood forecasting*. *Canadian Water Resources Journal* **30**(2), 159–174. <https://doi.org/10.4296/cwrj3.002159>.
- James, M. & Brown, R. J. 1977 *Geometric parameters that influence floodplain flow*. U.S. Army Engineer Waterways Experimental Station. Vicksburg Miss, June, Research report H-77.
- Kaushik, V. & Kumar, M. 2023a *Assessment of water surface profile in nonprismatic compound channels using machine learning techniques*. *Water Supply* **23**(1), 356–378. <https://doi.org/10.2166/ws.2022.430>.
- Kaushik, V. & Kumar, M. 2023b *Sustainable gene expression programming model for shear stress prediction in nonprismatic compound channels*. *Sustainable Energy Technologies and Assessments* **57**, 103229. <https://doi.org/10.1016/j.seta.2023.103229>.
- Kaushik, V. & Kumar, M. 2023c *Water surface profile prediction in non-prismatic compound channel using support vector machine (SVM)*. *AI in Civil Engineering* **2**, 6. <https://doi.org/10.1007/s43503-023-00015-1>.
- Khatua, K. K., Patra, K. C. & Mohanty, P. K. 2012 *Stage-discharge prediction for straight and smooth compound channels with wide floodplains*. *Journal of Hydraulic Engineering* **138**(1), 93–99.
- Klipalo, E., Besharat, M. & Kuriqi, A. 2022 *Full-scale interface friction testing of geotextile-based flood defence structures*. *Buildings* **12**(7), 990. <https://doi.org/10.3390/buildings12070990>.
- Knight, D. W., Tang, X., Sterling, M., Shiono, K. & McGahey, C. 2010 *Solving open channel flow problems with a simple lateral distribution model*. *River Flow* **1**, 41–48.
- Kuriqi, A. & Ardiclioglu, M. 2018 *Investigation of hydraulic regime at middle part of the Loire River in context of floods and low flow events*. *Pollack Periodica* **13**(1), 145–156. <https://doi.org/10.1556/606.2018.13.1.13>.

- Kuriqi, A., Pinheiro, A. N., Sordo-Ward, A. & Garrote, L. 2019 Influence of hydrologically based environmental flow methods on flow alteration and energy production in a run-of-river hydropower plant. *Journal of Cleaner Production* **232**, 1028–1042.
- Leandro, J., Chen, A. S., Djordjević, S. & Savić, D. A. 2009 Comparison of 1D/1D and 1D/2D coupled (Sewer/surface) hydraulic models for urban flood simulation. *Journal of Hydraulic Engineering* **135**(6), 495–504. [https://doi.org/10.1061/\(ASCE\)HY.1943-7900.0000037](https://doi.org/10.1061/(ASCE)HY.1943-7900.0000037).
- Liu, Z. & Merwade, V. 2018 Accounting for model structure, parameter and input forcing uncertainty in flood inundation modeling using Bayesian model averaging. *Journal of Hydrology* **565**, 138–149. <https://doi.org/10.1016/j.jhydrol.2018.08.009>.
- Mowinckel, E. 2011 *Flood Capacity Improvement of San Jose Creek Channel Using HEC-RAS*. Calif Polytech State Univ, California.
- Myers, W. R. C. & Elsayy, E. M. 1975 Boundary shears in channel with flood plain. *Journal of the Hydraulics Division ASCE* **101**(7), 933–946.
- Naik, B. & Khatua, K. K. 2016 Water surface profile computation for compound channels with narrow flood plains. *Arabian Journal for Science and Engineering* **42**(3), 941–955. doi:10.1007/s13369-016-2236-x.
- Naik, B., Kaushik, V. & Kumar, M. 2022 Water surface profile in converging compound channel using gene expression programming. *Water Supply* **22**(5), 5221–5236. <https://doi.org/10.2166/ws.2022.172>.
- Parhi, P. K., Sankhua, R. N. & Roy, G. P. 2012 Calibration of channel roughness for Mahanadi River, (India) using HEC-RAS model. *Journal of Water Resource and Protection* **04**(10), 847–850. <https://doi.org/10.4236/jwarp.2012.410098>.
- Patel, V. C. 1965 Calibration of the Preston tube and limitations on its use in pressure gradients. *Journal of Fluid Mechanics* **231**, 85–208.
- Proust, S., Rivière, N., Bousmar, D., Paquier, A. & Zech, Y. 2006 Flow in the compound channel with abrupt floodplain contraction. *Journal of Hydraulic Engineering* **132**(9), 958–970.
- Ramesh, R., Datta, B., Bhallamudi, S. M. & Narayana, A. 2000 Optimal estimation of roughness in open-channel flows. *Journal of Hydraulic Engineering* **126**(4), 299–303. [https://doi.org/10.1061/\(ASCE\)0733-9429\(2000\)126:4\(299\)](https://doi.org/10.1061/(ASCE)0733-9429(2000)126:4(299)).
- Rezaei, B. 2006 *Overbank Flow in Compound Channels with Prismatic and Nonprismatic Floodplains*. PhD Thesis, University of Birmingham, UK.
- Rezaei, B. & Knight, D. W. 2009 Application of the Shiono and Knight Method in the compound channel with nonprismatic floodplains. *Journal of Hydraulic Research* **47**(6), 716–726.
- Rezaei, B. & Knight, D. W. 2011 Overbank flow in compound channels with non-prismatic floodplains. *Journal of Hydraulic Engineering* **137**, 815–824.
- Sellin, R. H. J. 1964 A laboratory investigation into the interaction between flow in the channel of a river and that of its flood plain. *La Houille Blanche* **7**, 793–801.
- Shiono, K. & Knight, D. W. 1991 Turbulent open channel flows with variable depth across the channel. *Journal of Fluid Mechanics* **222**, 617–646.
- Subramanya, K. 2015 *Flow in Open Channels*, 4th edn. McGraw Hill, India.
- Timbadiya, P., Patel, P. L. & Porey, P. 2011 Calibration of HEC-RAS model on prediction of flood for lower Tapi River, India. *Journal of Water Resource and Protection* **03**(11), 805–811. <https://doi.org/10.4236/jwarp.2011.311090>.
- Wark, J. B., Samuels, P. C. & Ervine, D. A. 1990 A practical method of estimating velocity and discharge in compound channels. *Proc. River Flood Hydraulics*, pp. 163–172.
- Wormleaton, P. R. & Merrett, D. J. 1990 An improved method of the calculation for steady uniform flow in prismatic main channel/flood plain sections. *Journal of Hydraulic Research* **28**(2), 157–174.
- Wormleaton, P. R., Allen, J. & Hadjipanous, P. 1982 Discharge assessment in compound channel flow. *Journal of Hydraulics Division ASCE* **108**(9), 975–994.
- Yonesi, H. A., Omid, M. H. & Ayyoubzadeh, S. A. 2013 The hydraulics of flow in nonprismatic compound channels. *Journal of Civil Engineering and Urbanism* **3**(6), 342–356. [https://doi.org/10.1061/\(ASCE\)0733-9429\(2000\)126:4\(299\)](https://doi.org/10.1061/(ASCE)0733-9429(2000)126:4(299)).

First received 30 May 2023; accepted in revised form 31 August 2023. Available online 14 September 2023

# Modified High Gain Non-Isolated Boost DC-DC Converter for Electric Vehicles

Mohd Adib  
Department of Electrical Engineering  
Delhi Technological University  
Delhi, India  
mohdadib\_ee20a14\_11@dtu.ac.in

Saurabh Mishra, *Member IEEE*  
Department of Electrical Engineering  
Delhi Technological University  
Delhi, India  
saurabhmishra@dtu.ac.in

**Abstract**— This paper delves into the examination and simulation analysis of a modified HGBC that is intended for use in electric vehicles (EVs). The key objective of the converter presented is to boost the voltage of electrical energy stored within the battery pack, allowing it to power the electric motor and auxiliary EV ecosystem. The proposed converter input voltage range is 48V, and output voltage is 350V. The converter results demonstrate high efficiency at full load, low ripple voltage of less than 1%, and steady-state error of less than 1%. This paper showcases the efficiency and output voltage regulation effectiveness of the proposed topology. The HGBC proposed is anticipated to have a substantial impact on the development of efficient and reliable electric vehicle systems. The MATLAB/Simulink platform is used to validate the efficacy of the proposed system.

**Keywords**—*Electric Vehicle (EV), High Gain DC-DC Converter (HGDC), Modified High Gain NI DC-DC Converter (M-HGNIDC), High Gain Boost Converter (HGBC), Steady State (SS), Non-Isolated (NI).*

## I. INTRODUCTION

Electric vehicles (EVs) have picked up critical consideration in later a long time as an implication to decrease dependence on non-renewable vitality and control nursery gas emanations. In any case, optimizing power utilization may be a key challenge in the EV plan. One basic component in numerous EV frameworks is the HGBC, which steps up the voltage of electrical vitality put away within the battery to control the electric engine and other vehicle systems. HGBC could be a sort of DC-DC converter that employs electronic components like inductors, capacitors, and switches to change over a low-voltage input flag into a higher-voltage yield flag. Its key advantage is the capacity to attain tall voltage pick-up, making it a successful arrangement for applications requiring tall yield voltages.

In EVs, HGBC is utilized to boost the voltage of electrical vitality put away within the battery pack, ordinarily around 400-800 V, to control electric engines that require thousands of volts to function proficiently. By utilizing HGBC, the battery pack voltage can be optimized to the level required by the electric engine, empowering top performance.

HGBC offers tall productivity, minimizing misfortunes amid vitality exchange from the battery pack to the electric engine. This can be vital in maximizing vehicle extension, diminishing vitality utilization, and minimizing running costs. Additionally, HGBC's flexibility and adaptability permit it to be custom fitted to meet the particular needs of an EV framework, optimizing execution and efficiency.

The tall degree of control over the output voltage makes HGBC a well-known choice for numerous EV applications.

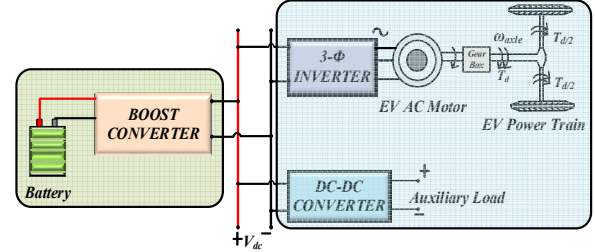


Fig. 1: Block diagram of EV ecosystem

Furthermore, it can be utilized in other scenarios where tall voltage yield is required, such as making a DC grid to control customer hardware, which can be changed over to AC power by an inverter. This approach offers a few focal points, counting expanded productivity and the capacity to utilize renewable vitality sources like sun-powered boards to control the DC framework. Generally, HGBC's flexibility and adaptability make it a profitable innovation for a extend of applications.

In rundown, HGBC plays a vital part in numerous EV frameworks by effectively boosting the voltage of electrical vitality stored in the battery pack to meet the voltage prerequisites of electric engines. Known for their proficiency, versatility, and adaptability, HGBCs have gotten to be a key innovation empowering the advancement and arrangement of electric vehicles. [1],[2] describe different non-inverting HGDC topologies aimed for solar PV systems, featuring advancements in efficiency, duty cycle range, and voltage gain. [3],[4] present novel NI DC-DC converter topologies for microgrid applications, which offer high voltage gain and reduced stress on switches and diodes. Additionally, [5],[6] discuss a boost converter topology based on the switched-inductor and single switch for high step-up DC-DC conversion in PV systems, which achieves improved efficiency. [7] conducts a comparison of various DC-DC converters. Finally, [8] provides specific details on the design of an onboard battery charger utilizing an interleaved Luo converter topology cascaded with a fly-back converter.

This paper presents an analytical assessment of a M-HGNIDC designed for use in electric vehicles. The system schematic is outlined in section II, while section III elaborates on the converter's topology, mode of operation, duty cycle calculation, and component sizing. To gain insight into the converter's behavior under different operating conditions, a small signal analysis is conducted in section IV. Section V presents the results of the complete analysis, while the conclusion of the analytical study is outlined in the subsequent section. This section provides insights into the potential applications and future developments of high gain modified boost converters in the field of electric vehicles.

## II. SYSTEM SCHEMATIC

The system schematic, illustrated in Fig.2, shows the configuration of the proposed converter. The voltage output of the converter is greater than the voltage input and varies according to the switching frequency. The circuit is composed of two switches, three diodes, three capacitors, and two inductors, all of which are illustrated in Fig. 2.

### III. ANALYSIS & CONTROL OF M-HGNIDC CONVERTER

The presented system in Fig.2 is modelled and analyzed for a fixed duty ratio. The control of the presented system is also presented with dual loop control.

#### A. Modes of Operation

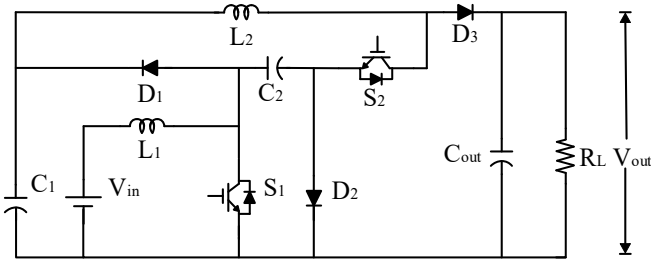


Fig.2 Modified High Gain NI DC-DC Converter (M-HGNIDC)

The proposed converter is analyzed two modes (Mode-I and Mode-II) depending on the switching of the switches, which is controlled from the dual loop control logic.

##### i) MODE-I Operation: $S_1$ ON, $S_2$ ON

The proposed converter has been analyzed for the duration of  $0 < t < DT$  while switches  $S_1$  and  $S_2$  are closed, resulting in an equivalent circuit as depicted in Fig.3.

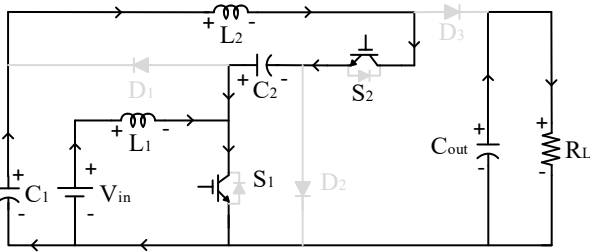


Fig. 3 Mode I operation of M-HGNIDC

During this time, inductors  $L_1$  and  $L_2$  are charging, with  $L_1$  charged by the input DC source and  $L_2$  charged by capacitors  $C_1$  and  $C_2$ . The equations for this circuit are obtained using Kirchhoff's Voltage Law (KVL).

$$V_{L1} = L_1 \frac{di_{L1}}{dt} = V_{in} \quad [1]$$

$$V_{L2} = L_2 \frac{di_{L2}}{dt} = V_{C1} + V_{C2} \quad [2]$$

##### ii) MODE-II Operation: $S_1$ OFF, $S_2$ OFF

When the time interval  $DT < t < T$ , the switches  $S_1$  and  $S_2$  are open, and the circuit enters into a discharging state as presented in Fig.4.. At this stage, the inductors  $L_1$  and  $L_2$  release energy, with  $L_1$  transferring its energy to capacitor  $C_2$

and  $L_2$  releasing energy to the output voltage side. The voltage equations for this circuit are expressed by applying Kirchhoff's Voltage Law (KVL).

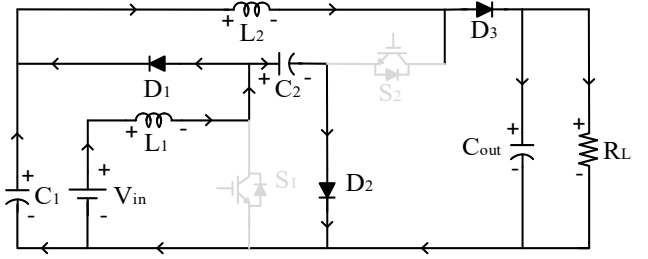


Fig. 4 Mode II operation of M-HGNIDC

$$V_{L1} = L_1 \frac{di_{L1}}{dt} = V_{in} - V_{C1} = V_{in} - V_{C2} \quad [3]$$

$$V_{L2} = L_2 \frac{di_{L2}}{dt} = -V_{out} + V_{C1} = -V_{out} + V_{C2} \quad [4]$$

Voltage across capacitor voltages  $C_1$  &  $C_2$  are expressed as,

$$V_{C1} = V_{C2} + \frac{V_{in}}{1-D} \quad [5]$$

The output voltage  $V_{out}$  of the presented converter is estimated as,

$$V_{out} = V_{in} * \frac{1+D}{(1-D)^2} \quad [6]$$

The presented converter system waveforms are presented in Fig.5. The turn ON period of the presented converter is marked by  $DT$  and the turn OFF period is marked by  $(1-D)T$ .

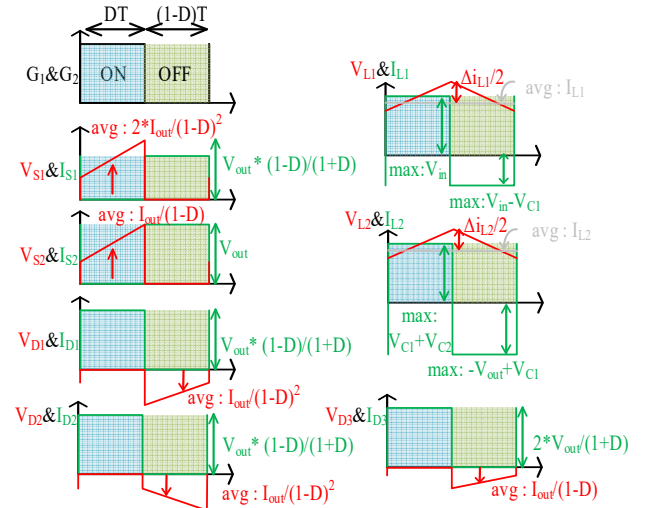


Fig.5. System parameters waveforms of M-HGNIDC Converter

#### B. Control of the M-HGNIDC Converter

The control of the presented system is outlined in Fig.7. The output voltage of the presented system is sensed and fed to the outer loop of the control block to generate the current reference of the battery and then the inner loop is used to generate the switching pulses to the two switches.



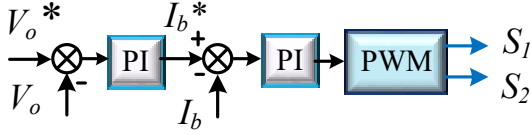


Fig.6. Control Logic of the system.

### C. Derivation of the Duty cycle (D)

The converter presented in this study defines the duty cycle as a function of the input and output voltages, as shown by the following equation;

$$D = \frac{1}{2} \left( 2 * V_{out} + V_{in} - \frac{\sqrt{V_{in}^2 + 8 * V_{in} * V_{out}}}{V_{out}} \right) \quad [7]$$

The ripple current in inductors L1 and L2 is as follows:

$$\Delta i_{L1} = \frac{D}{L1 * f_s} * V_{in} \quad [8]$$

$$\Delta i_{L2} = \frac{2 * D}{L2 * f_s * (1-D)} * V_{in} \quad [9]$$

### D. Semiconductor Device Rating

The proposed converter has specific voltage and current ratings for its semiconductor devices, which are listed as follows:

$$V_{S1} = V_{out} * \frac{1-D}{1+D} = V_{D1} = V_{D2} \quad [10]$$

$$V_{S2} = V_{out} \quad [11]$$

$$V_{D3} = V_{out} * \frac{2}{1+D} \quad [12]$$

$$I_{S1} = 2I_{out} * \frac{1}{(1-D)^2} \quad [13]$$

$$I_{S2} = I_{out} * \frac{1}{(1-D)} = I_{D3} \quad [14]$$

$$I_{D1} = I_{out} * \frac{1}{(1-D)^2} \quad [15]$$

$$I_{D2} = I_{out} * \frac{D}{(1-D)^2} \quad [16]$$

### E. Component Selection

Below are the values of inductors and capacitors used in the design of the converter:

$$L1 \geq \frac{DV_{in}}{\Delta i_{L1} f_s} = 2mH \quad [17]$$

$$L2 \geq \frac{2DV_{in}}{\Delta i_{L2} f_s (1-D)} = 2mH \quad [18]$$

$$C1 = C2 \geq \frac{D(1-D)I_{in}}{\Delta V_{C1,2} f_s (1+D)} = 100\mu F \quad [19]$$

$$C_{out} \geq \frac{D(1-D)^2 V_{in}}{\Delta V_{C_{out}} f_s (1+D)} = 220\mu F \quad [20]$$

## IV. SMALL SIGNAL ANALYSIS OF THE CONVERTER

This paper presents an examination of the state signal for a modified HGBC. The modified boost converter, which could be a variety of the commonly utilized DC-DC power conversion procedure, offers improved effectiveness and stability compared to the conventional boost converter. Due to its high gain property, the modified boost converter is well-suited for applications that require high output voltage. The state signal investigation is performed to way better get the dynamic behavior of the converter and to help in planning control methodologies to upgrade its execution. The investigation discoveries give a premise for future inquiries about optimizing modified HGBCs.

The state space averaging method is utilized to obtain the small signal analysis. In this method, the input variable is denoted as  $v_{in}(t)$ , the control variable as  $d$ , and the output variable as  $v_{out}(t)$ . The state variables consist of  $i_{L1}(t)$ ,  $i_{L2}(t)$ ,  $v_{C1}(t)$ ,  $v_{C2}(t)$ , and  $v_{Cout}(t)$ .

In the state where the switch is ON, the state space average model is expressed as,

$$\begin{bmatrix} \frac{di_{L1}(t)}{dt} \\ \frac{di_{L2}(t)}{dt} \\ \frac{dv_{C1}(t)}{dt} \\ \frac{dv_{C2}(t)}{dt} \\ \frac{dv_{Cout}(t)}{dt} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{L2} & \frac{1}{L2} & 0 \\ 0 & \frac{1}{C1} & 0 & 0 & 0 \\ 0 & \frac{1}{C2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{-1}{C_{out} * R} \end{bmatrix} \begin{bmatrix} i_{L1}(t) \\ i_{L2}(t) \\ v_{C1}(t) \\ v_{C2}(t) \\ v_{Cout}(t) \end{bmatrix} + \begin{bmatrix} \frac{2}{L1} \\ \frac{-1}{L2} \\ 0 \\ 0 \\ 0 \end{bmatrix} [v_{in}(t)] \quad [21]$$

The output voltage where the switch is ON in terms of the state space variables is expressed as,

$$[v_{out}(t)] = [0 \ 0 \ 0 \ 0 \ 1] \begin{bmatrix} i_{L1}(t) \\ i_{L2}(t) \\ v_{C1}(t) \\ v_{C2}(t) \\ v_{Cout}(t) \end{bmatrix} \quad [22]$$

In the state where the switch is OFF, the state space average model is expressed as,

$$\begin{bmatrix} \frac{di_{L1}(t)}{dt} \\ \frac{di_{L2}(t)}{dt} \\ \frac{dv_{C1}(t)}{dt} \\ \frac{dv_{C2}(t)}{dt} \\ \frac{dv_{Cout}(t)}{dt} \end{bmatrix} = \begin{bmatrix} 0 & 0 & \frac{-1}{L1} & 0 & 0 \\ 0 & 0 & \frac{1}{L2} & 0 & \frac{-1}{L2} \\ \frac{-1}{C1} & \frac{1}{C1} & 0 & \frac{1}{C1 * r} & 0 \\ 0 & 0 & 0 & \frac{1}{C2 * r} & 0 \\ 0 & \frac{1}{C_{out}} & 0 & 0 & \frac{-1}{C_{out} * R} \end{bmatrix} \begin{bmatrix} i_{L1}(t) \\ i_{L2}(t) \\ v_{C1}(t) \\ v_{C2}(t) \\ v_{Cout}(t) \end{bmatrix} + \begin{bmatrix} \frac{2}{L1} \\ \frac{-1}{L2} \\ 0 \\ 0 \\ 0 \end{bmatrix} [v_{in}(t)] \quad [23]$$

The output voltage where the switch is OFF in terms of the state space variables is expressed as,

$$[v_{out}(t)] = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} i_{L1}(t) \\ i_{L2}(t) \\ v_{C1}(t) \\ v_{C2}(t) \\ v_{Cout}(t) \end{bmatrix} \quad [24]$$

The average model of the Boost Converter can be obtained by combining equations (21) and (23), as well as equations (22) and (24), which leads to the representation of the model as equation (25) and (26).

$$\begin{bmatrix} \frac{di_{L1}(t)}{dt} \\ \frac{di_{L2}(t)}{dt} \\ \frac{dv_{C1}(t)}{dt} \\ \frac{dv_{C2}(t)}{dt} \\ \frac{dv_{Cout}(t)}{dt} \end{bmatrix} = \begin{bmatrix} 0 & 0 & \frac{-1}{L_1} & 0 & 0 \\ 0 & 0 & \frac{1}{L_2} & 0 & \frac{-1}{L_2} \\ \frac{d(t)}{C_1} & \frac{1}{C_1} & 0 & \frac{1-d(t)}{C_1 * r} & 0 \\ 0 & \frac{d(t)}{C_2} & 0 & \frac{1-d(t)}{C_2 * r} & 0 \\ 0 & \frac{1-d(t)}{C_{out}} & 0 & 0 & \frac{d(t)-1}{C_{out} * R} \end{bmatrix} \begin{bmatrix} i_{L1}(t) \\ i_{L2}(t) \\ v_{C1}(t) \\ v_{C2}(t) \\ v_{Cout}(t) \end{bmatrix} + \begin{bmatrix} \frac{2}{L_1} \\ \frac{-1}{L_2} \\ 0 \\ 0 \\ 0 \end{bmatrix} [v_{in}(t)] \quad [25]$$

$$[v_{out}(t)] = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} i_{L1}(t) \\ i_{L2}(t) \\ v_{C1}(t) \\ v_{C2}(t) \\ v_{Cout}(t) \end{bmatrix} \quad [26]$$

The introduction of small-signal disturbance variables allows for the description of the state variables, input variable, output variable, and control variable as follows:

$$\begin{cases} i_{L1}(t) = I_{L1} + \hat{i}_{L1}(t) \\ i_{L2}(t) = I_{L2} + \hat{i}_{L2}(t) \\ v_{C1}(t) = V_{C1} + \hat{v}_{C1}(t) \\ v_{C2}(t) = V_{C2} + \hat{v}_{C2}(t) \\ v_{Cout}(t) = V_{Cout} + \hat{v}_{Cout}(t) \\ v_{in}(t) = V_{in} + \hat{v}_{in}(t) \\ v_{out}(t) = V_{out} + \hat{v}_{out}(t) \\ d(t) = D + \hat{d}(t) \end{cases} \quad [27]$$

Where  $I_{L1}$ ,  $I_{L2}$ ,  $V_{C1}$ ,  $V_{C2}$ ,  $V_{Cout}$ ,  $V_{in}$ ,  $V_{out}$  &  $D$  represent the S S components. The small-signal disturbance variables are represented as  $\hat{i}_{L1}(t)$ ,  $\hat{i}_{L2}(t)$ ,  $\hat{v}_{C1}(t)$ ,  $\hat{v}_{C2}(t)$ ,  $\hat{v}_{Cout}(t)$ ,  $\hat{v}_{in}(t)$ ,  $\hat{v}_{out}(t)$  &  $\hat{d}(t)$ . By combining equations (25), (26), and (27), the converter's small-signal demonstration can be determined, as appeared in equations (28) and (29). This demonstration offers profitable data on the converter's energetic reaction, helping with the advancement of successful control methodologies to upgrade its general performance.

$$\begin{bmatrix} \frac{d\hat{i}_{L1}(t)}{dt} \\ \frac{d\hat{i}_{L2}(t)}{dt} \\ \frac{d\hat{v}_{C1}(t)}{dt} \\ \frac{d\hat{v}_{C2}(t)}{dt} \\ \frac{d\hat{v}_{Cout}(t)}{dt} \end{bmatrix} = \begin{bmatrix} 0 & 0 & \frac{D-1}{L_1} & 0 & 0 \\ 0 & 0 & \frac{1}{L_2} & \frac{D}{L_2} & \frac{D-1}{L_2} \\ \frac{D}{C_1} & \frac{1}{C_1} & 0 & \frac{1-D}{C_1 * r} & 0 \\ 0 & \frac{D}{C_2} & 0 & \frac{1-D}{C_2 * r} & 0 \\ 0 & \frac{1-D}{C_{out}} & 0 & 0 & \frac{D-1}{C_{out} * R} \end{bmatrix} \begin{bmatrix} \hat{i}_{L1}(t) \\ \hat{i}_{L2}(t) \\ \hat{v}_{C1}(t) \\ \hat{v}_{C2}(t) \\ \hat{v}_{Cout}(t) \end{bmatrix} + \begin{bmatrix} 0 & 0 & \frac{1}{L_1} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{L_2} & \frac{1}{L_2} \\ \frac{1}{C_1} & 0 & 0 & \frac{-1}{C_1 * r} & 0 \\ 0 & \frac{1}{C_2} & 0 & \frac{-1}{C_2 * r} & 0 \\ 0 & \frac{-1}{C_{out}} & 0 & 0 & \frac{1}{C_{out} * R} \end{bmatrix} \begin{bmatrix} I_{L1}(t) \\ I_{L2}(t) \\ V_{C1}(t) \\ V_{C2}(t) \\ V_{Cout}(t) \end{bmatrix} \begin{bmatrix} \hat{d}(t) \end{bmatrix} + \begin{bmatrix} \frac{2}{L_1} \\ \frac{-1}{L_2} \\ 0 \\ 0 \\ 0 \end{bmatrix} [v_{in}(t)] \quad [28]$$

$$[\hat{v}_{out}(t)] = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{i}_{L1}(t) & \hat{i}_{L2}(t) & \hat{v}_{C1}(t) & \hat{v}_{C2}(t) & \hat{v}_{Cout}(t) \end{bmatrix}^T \quad [29]$$

Within the system examined in this paper, the closed-loop control depends on two parameters, to be specific, the duty ratio and input voltage. These parameters are utilized to decide the transfer function of the system. Specifically, the transfer function within the s-domain is communicated as takes after when the duty ratio is considered as the input variable:

$$G_{do}(s) = \left. \frac{\hat{u}_o(s)}{\hat{d}(s)} \right|_{\hat{u}_{in}(s)=0} \quad [30]$$

$$G_{do}(s) = \frac{8963.6 (s - 4.6e12) (s + 9.369e04) (s - 1046) (s + 1011)}{(s - 4.6e12)(s - 1746)(s + 1744)(s^2 + 22.95s + 1.671e05)} \quad [31]$$

Moreover, the transfer function of the system with input voltage as the variable is expressed as,

$$G_{io}(s) = \left. \frac{\hat{u}_o(s)}{\hat{u}_{in}(s)} \right|_{\hat{d}(s)=0} \quad [32]$$

$$G_{io}(s) = \frac{-5.0182e07(s - 4.6e12)(s^2 + 3.542e06)}{(s - 4.6e12)(s - 1746)(s + 1744)(s^2 + 22.95s + 1.671e05)} \quad [33]$$

The dual close loop system of the presented converter has the following derived proportional and integral gain,

Current controller gains:  $K_P = 1.88$ ,  $K_I = 0.45$

Voltage controller gains:  $K_P = 2.5$ ,  $K_I = 1.25$

## V. RESULTS AND DISCUSSION

The MATLAB/SIMULINK platform was used to simulate the discussed system at a switching frequency of 10 kHz. The simulation was conducted by applying an input voltage range of 48V and obtaining an output voltage of 350V. For the



purpose of emulating an electric vehicle battery, a battery with a rating of 48V and 100Ah was used as the input. The presented system of M-HGNIDC converter SS analysis is presented in Fig.7. The performance of the presented system from the input end (battery) is presented in Fig. 6(a) with input voltage, input current, State of charge (SOC) and power as variables. The negative slope is an indicator of battery discharging for the time  $t=0.1s$  to  $0.1001s$  as shown. The converter SS results in terms of inductor voltages and currents are presented in Fig. 7(b). The current ripples in the inductor currents are less than 1%, which is very low as per the power rating of the presented system. The voltage stress on the inductor  $L_1$  and  $L_2$  are also under operating limits. Fig. 7(c) presents the switch voltage and current performance during the different modes of operation. As per the modes of operation presented in the paper, the switch voltages and currents are under the operating limits. The switch ( $S_1$ ) is marked for high current stress and low voltage stress as compared to the switch ( $S_2$ ) accordingly as per the functioning of the presented converter. The SS performance of the converter diodes with the output voltage and in terms of voltages and current is presented in Fig. 7(d) and Fig. 7(e) respectively. The diodes performance is also as per the required performance of the presented system. The ripples in the output voltage and output current are also minimal (less than 0.5%) with negligible SS error. The capacitor voltages and currents of the presented system with minimal ripples are presented in Fig. 7(f). The presented performance of the converter is shown for small time interval for better efficacy. Fig. 8 outlines the yield voltage execution of the converter displayed. For best performance of the M-HGNIC, the duty ratio range is from 0.5-0.8. The comparison of the presented system based on the component level with the other topologies are presented in Fig. 9. The presented system is with less number of power electronic components for a substantial high performance and gain of the converter.

## VI. CONCLUSION

The HGBC is an effective solution for enhancing the efficiency and power density of electric vehicle (EV) systems. Its innovative design deviates from the traditional boost converter and is anticipated to gain significant traction in the EV industry. The converter topology outlined in this paper integrates a boost converter and a buck-boost converter to achieve high gain and optimize overall system performance. The reenactment comes about to have affirmed the viability of the proposed topology in different viewpoints such as productivity, control thickness, yield voltage control, and exchanging voltage. Besides, the utilization of two-loop control has driven superior execution and solidness of the framework, which is basic for accomplishing precise control of the converter. This paper's displayed adjusted M-HGNIDC could be a profitable commitment to electric vehicle control hardware, and it is anticipated to essentially affect the advancement of productive and solid electric vehicle systems. The transformer has effectively accomplished the specified execution in terms of voltage pick-up and framework productivity, making it a promising arrangement for electric vehicles.

## ACKNOWLEDGMENT

The authors are thankful to Department of Electrical Engineering and Centre of Excellence for Electric Vehicle and Related

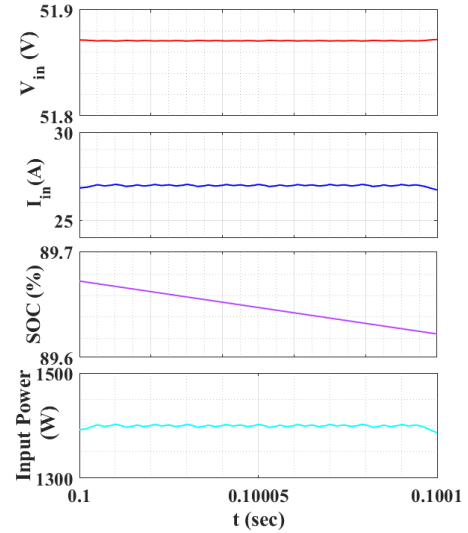


Fig 7 (a) SS results of Battery Voltage,  $V_{in}$  in V, Battery Current,  $I_{in}$  in A, State of Charge (SOC) in % and Input Power in W.

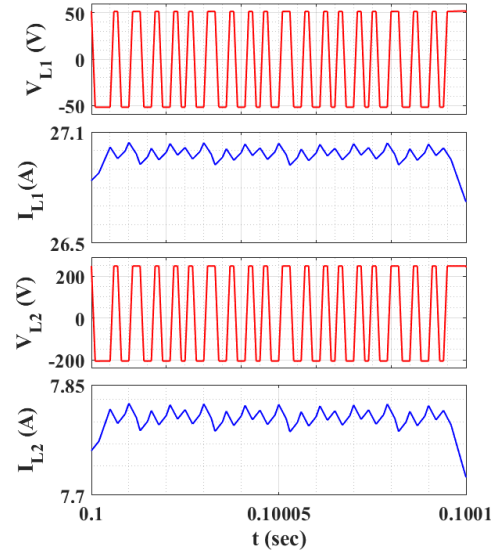


Fig 7 (b) SS results of Inductor Voltages,  $V_{L1}$  and  $V_{L2}$  in V, and Inductor Currents,  $I_{L1}$  and  $I_{L2}$  in A.

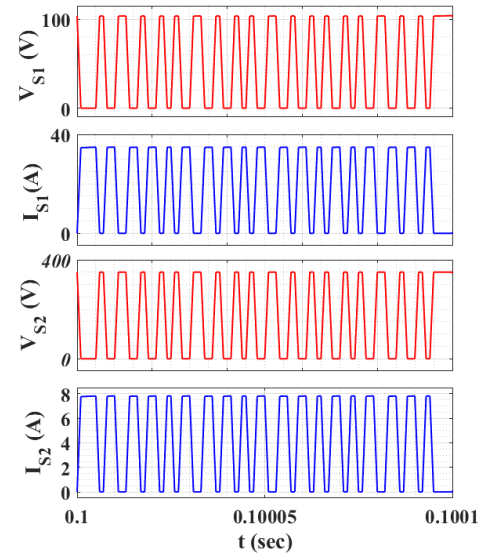


Fig 7 (c) SS results of Switch Voltages,  $V_{S1}$  and  $V_{S2}$  in V, Switch Currents,  $I_{S1}$  and  $I_{S2}$  in A.

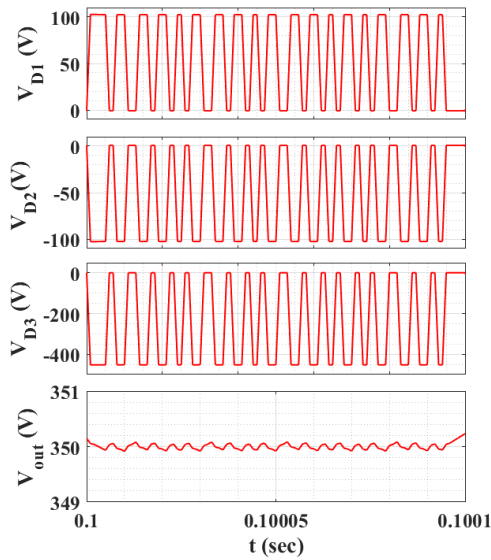


Fig 7 (d) SS results of Diode Voltages,  $V_{D1}$ ,  $V_{D2}$  and  $V_{D3}$  in V and the output voltage  $V_{out}$  in V.

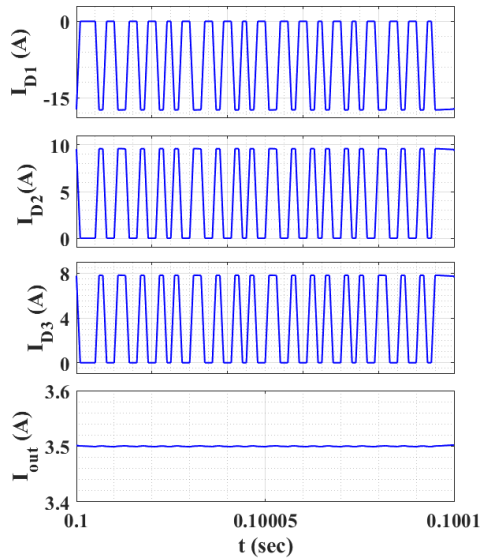


Fig 7 (e) SS results of Diode Currents,  $I_{D1}$ ,  $I_{D2}$ ,  $I_{D3}$  in A and the output current  $I_{out}$  in A.

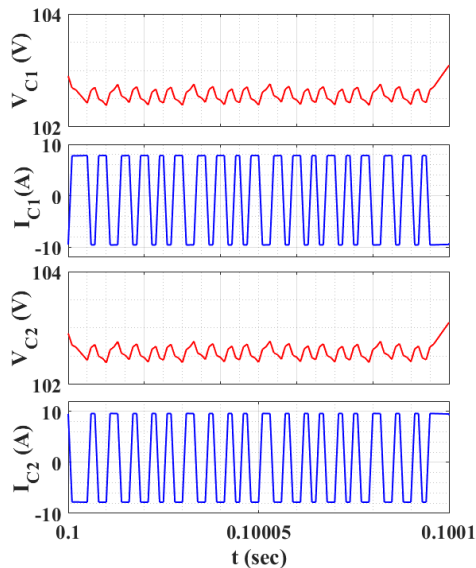


Fig 7 (f) SS results of Capacitor Voltages,  $V_{C1}$  and  $V_{C2}$  in V, Capacitor Currents,  $I_{C1}$  and  $I_{C2}$  in A.

Technologies (COEVRT), Delhi Technological University for providing the necessary support.

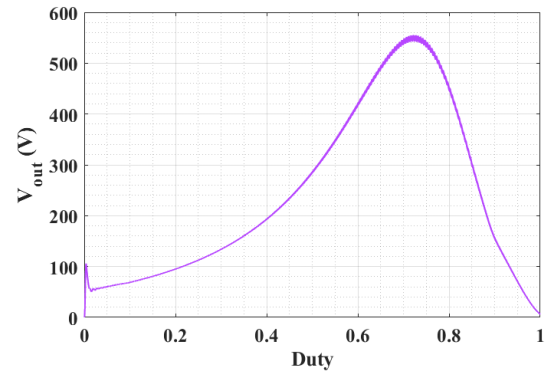


Fig 8. Output voltage variation with respect to duty ratio.

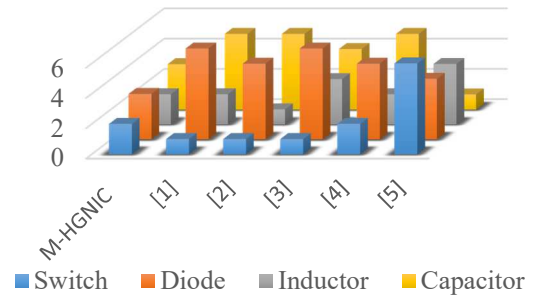


Fig 9. Comparison of the presented M-HGNIDC converter with other topologies.

## REFERENCES

- [1] Mahmood, A. , Zaid, M. , Ahmad, J. , Khan, M. A. , Khan, S. , Sifat, Z. , Lin, C. H. , Sarwar, A. , Tariq, M. , and Alamri, B, "A Non-Inverting High Gain DC-DC Converter With Continuous Input Current," in *IEEE Access*, vol. 9, pp. 54710-54721, 2021.
- [2] R. Suriyakulnaayudhya, "A Bootstrap Charge-Pump Technique for High Gain Boost Converter Applications," *2018 2nd European Conference on Electrical Engineering and Computer Science (EECS)*, Bern, Switzerland, 2018, pp. 533-537.
- [3] M. Zaid, J. Ahmad, A. Sarwar, Z. Sarwer, M. Tariq and A. Alam, "A Transformer less Quadratic Boost High Gain DC-DC Converter," *2020 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES)*, Jaipur, India, 2020, pp. 1-6.
- [4] F. A. A. Meinagh and G. Mirzaeva, "Detailed Design of a High Voltage Gain Quasi Switched-Boost DC-DC Converter," *2020 Australasian Universities Power Engineering Conference (AUPEC)*, Hobart, Australia, 2020, pp. 1-6.
- [5] A. Siadatan, S. Tahzibi, R. Babaloo and J. Gotlieb, "Simple Switched-Inductor High-Gain Boost converter," *2020 International Symposium on Power Electronics, Electrical Drives, Automation and Motion (SPEEDAM)*, Sorrento, Italy, 2020, pp. 733-737.
- [6] L. Qin, L. Zhou, W. Hassan, J. L. Soon, M. Tian and J. Shen, "A Family of Transformer-Less Single-Switch Dual-Inductor High Voltage Gain Boost Converters With Reduced Voltage and Current Stresses," in *IEEE Transactions on Power Electronics*, vol. 36, no. 5, pp. 5674-5685, May 2021.
- [7] S. Yadav, S. Mishra and Garima, "Comparative Analysis of NI DC-DC Converters for Solar-Photovoltaic System," *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, Noida, India, 2021, pp. 880-885.
- [8] C. Chaudhary, S. Mishra and U. Nangia, "Unidirectional Onboard Battery Charging for Electric Vehicle using Interleaved Luo Converter," *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, Noida, India, 2021, pp. 874-879.



# Monoclinic to cubic structural transformation, local electronic structure, and luminescence properties of Eu-doped HfO<sub>2</sub>

Rajesh Kumar<sup>1</sup> · Jitender Kumar<sup>2</sup> · Ramesh Kumar<sup>3</sup> · Akshay Kumar<sup>4</sup> · Aditya Sharma<sup>5</sup> · S. O. Won<sup>6</sup> · K. H. Chae<sup>6</sup> · Mukhtiyar Singh<sup>1</sup> · Ankush Vij<sup>7</sup>

Received: 8 June 2023 / Accepted: 8 September 2023

© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE part of Springer Nature 2023

## Abstract

We report the photoluminescence (PL) properties of Eu<sup>3+</sup> doped HfO<sub>2</sub> nanoparticles prepared using co-precipitation method and annealed at 600 °C. X-ray diffraction results revealed the monoclinic phase in undoped HfO<sub>2</sub> and show mixed phase formation at lower concentration and a dominant cubic phase achieved at 5 mol% doping of Eu in HfO<sub>2</sub>. The phase transition anticipated by the density functional theory is in excellent agreement with experimental findings. The oxygen K-edge XANES spectra clearly depicts the diverse hybridization of O 2p orbitals in M–O7 (for monoclinic) and M–O8 (for cubic) polyhedra of HfO<sub>2</sub>. Hf L-edge XANES confirms Hf<sup>4+</sup> ions in cubic and monoclinic structured HfO<sub>2</sub>. The Eu<sup>3+</sup> ions are dominantly present in the Eu-doped HfO<sub>2</sub> nanoparticles. PL study demonstrates the emission in red region with high color purity under different excitation wavelengths from near UV to blue light. PL emission spectra show four emission bands at 594 nm, 609 nm, 650 nm, and 716 nm corresponding to 4f–4f transitions of Eu<sup>3+</sup> under excitation wavelengths of 361 nm, 383 nm, 394 nm and 465 nm. The reddish PL emission with high color purity under different excitation wavelengths from near-UV to blue region may be exploited in solid state lighting-based applications.

**Keywords** HfO<sub>2</sub> · Phase transition · XANES · PL · DFT

## 1 Introduction

The technological areas of the phosphor-based materials have gathered great attention due to numerous enthralling potential applications which includes solid state lighting, light emitting diodes, and scintillation [1, 2]. The contemporary light emitting diodes (LEDs) have received much attention over the conventional light sources due to their exclusive properties, such as low operating voltage, longer lifetime, compactness, high efficiency, and diverse applicability, makes it next generation illumination sources [3, 4]. Researchers also particularly drawn to RE-doped luminous materials because of their notable qualities including, high color rendering index (CRI), good color purity, great chemical stability and high luminescence efficiency [5, 6]. The RE doping has the ability to improve electron transport between 4f levels and their wavelength ranges from ultraviolet (UV) to infrared (IR) [5]. Numerous oxide-based materials such as TiO<sub>2</sub>, ZrO<sub>2</sub>, HfO<sub>2</sub>, SnO<sub>2</sub> have been exploited with the RE doping to enrich the field of luminescence [6–9].

Among these oxide-based compounds, HfO<sub>2</sub> is practically important due to its wide bandgap (~5.7 eV), high dielectric

✉ Mukhtiyar Singh  
msphysik09@gmail.com

Ankush Vij  
vij\_anx@yahoo.com

<sup>1</sup> Department of Applied Physics, Delhi Technological University, Delhi 110042, India

<sup>2</sup> Department of Physics, Amity University Haryana, Gurugram 122413, India

<sup>3</sup> Department of Physics, Guru Jambheshwar University of Science and Technology, Hisar 125001, India

<sup>4</sup> School of Material Science and Engineering, Changwon National University, Changwon, Gyeongnam 51140, South Korea

<sup>5</sup> Department of Physics, Manav Rachna University, Faridabad, Haryana 124001, India

<sup>6</sup> Advance Analysis Centre, Korea Institute of Science and Technology (KIST), Seoul 02792, South Korea

<sup>7</sup> Department of Physics and Astrophysics, Central University of Haryana, Mahendergarh 123031, India

constant, dynamical stability, high thermal stability and melting point [10–13]. The high refractive index (1.8–2.1) and low absorption to the UV to mid-IR light photon makes it suitable for anti-reflective multi-layer coatings [14]. The high effective atomic number ( $Z_{\text{eff}} \sim 67.2$ ) and high atomic density ( $\sim 9.7 \text{ g/cm}^3$ ) of  $\text{HfO}_2$  are suitable for modern scintillators [15]. Along with the development of high-performance devices of this material, the  $\text{HfO}_2$  also drawn much research attention for its structural re-adjustments. The  $\text{HfO}_2$  obey three crystallographic phases; cubic (*c*) (space group;  $\text{Fm } \bar{3} \text{ m}$ ), tetragonal (*t*) (space group;  $\text{P4}_2/\text{nmc}$ ) and monoclinic (*m*) (space group;  $\text{P2}_1/\text{c}$ ). The monoclinic phase is stable at temperature ( $\sim 1100^\circ \text{C}$ ) and undergoes a phase transformation at high temperatures ( $\sim 1720^\circ \text{C}$ ) and converted into a tetragonal phase [16]. At a higher temperature ( $\sim 2600^\circ \text{C}$ ), the tetragonal phase can also be transformed into cubic phase [17]. Instead of high temperature annealing-based phase transformation studies, the monoclinic to cubic phase transformation of  $\text{HfO}_2$  has been achieved through yttrium doping at moderate temperatures [18]. Likewise, doping of other aliovalent ions such as  $\text{Dy}^{3+}$  and  $\text{Sm}^{3+}$  and isovalent ions such as  $\text{Ti}^{4+}$  and  $\text{Zr}^{4+}$  have been reported for stabilizing the cubic or tetragonal phases at lower annealing temperatures ( $500\text{--}800^\circ \text{C}$ ) [8, 19, 20]. Dopant incorporation into  $\text{HfO}_2$  results in defects and oxygen vacancies due to charge and ionic radius variations between the dopants and  $\text{Hf}^{4+}$ . Increase in the doping concentration of  $\text{RE}^{3+}$  ions lead to rise the concentration of oxygen vacancy for charge compensation. These vacancies are randomly arranged in the local surroundings of  $\text{Hf}^{4+}$ , creating several slightly different sites with diverse coordination [21]. The oxygen vacancies are responsible for reducing the repulsive force between neighbouring sites, creating a modification in the lattice parameters, and causing the ions to be arranged in a new crystal structure [22–24]. It has been observed that formation of oxygen ion vacancies and their coordination with  $\text{Hf}^{4+}$  cations helped to stabilize the tetragonal and cubic phases when the oversized aliovalent rare earth cations were incorporated in  $\text{HfO}_2$  [25].  $\text{Eu}^{3+}$  doping effect on the structure modification of  $\text{HfO}_2$  have been observed under the temperature variation and a tetragonal phase appears at high temperature [26]. Density functional theory calculations have shown that oversized trivalent ions preferentially stabilize the cubic phase and not the tetragonal phase [27]. Experimental studies have shown mixed results on the formation of tetragonal and cubic phases with the doping of lanthanides in  $\text{HfO}_2$ . For example, based on X-ray diffraction (XRD) results, 20% Tb could stabilize the tetragonal phase in  $\text{HfO}_2$  thin films [28]. A few studies have been reported on the local bonding arrangement and modification in the crystal system [29, 30]. X-ray absorption spectroscopy (XAS) is commonly used to identify the defect type and formal valance states and local symmetry of dopant ions using X-ray

absorption near edge structure (XANES). Intrinsic luminescence has been reported in undoped  $\text{HfO}_2$  which ascribed to the oxygen vacancy which acts as luminescence centres in the crystal lattice [10]. The wide bandgap and low phonon frequency makes it promising host for doping of RE activator ions [11, 31]. Efforts have been intensified for the PL properties of  $\text{HfO}_2$  via doping of RE activator ions such as  $\text{Sm}^{3+}$ ,  $\text{Dy}^{3+}$ ,  $\text{Gd}^{3+}$ ,  $\text{Pr}^{3+}$ , and  $\text{Er}^{3+}$  which leads to energy transfer between different energy levels of host and activator [8, 25, 32, 33]. The RE-doped  $\text{HfO}_2$  nanocrystals have been found to alter in size and exhibit phase-dependent luminous characteristics at varying temperatures and pH levels [34]. Under the suitable doping condition  $\text{Eu}^{3+}$ -doped amorphous  $\text{HfO}_2$  leads to PL due to defects which reside in the crystal sites with inversion centres [30]. It is demonstrated that doping of  $\text{Sm}^{3+}$  ions in  $\text{HfO}_2$  leads to phase transformation and shows the emission in near green and red region [35], whereas  $\text{Dy}^{3+}$  and  $\text{Sm}^{3+}$  binary-doped  $\text{HfO}_2$  nanophosphors shows PL emission in blue, yellow and red region and used as latent finger print imaging application [8].

With the above impulse, in the present study, we have demonstrated structure, morphological, XANES, PL response of  $\text{HfO}_2$  with varying  $\text{Eu}^{3+}$  concentration. The phase transformation of  $\text{HfO}_2$  powder at lower temperature obtained by chemical co-precipitation method. The effect of phase transition on electronic and optical properties have investigated by XANES and PL.

## 2 Experimental and theoretical details

### 2.1 Synthesis method

The chemical co-precipitation method was employed to synthesize the  $\text{HfO}_2$  doped with varying  $\text{Eu}^{3+}$  ion concentrations [10, 30]. To synthesize the powder sample,  $\text{Eu}^{3+}$  doping ratio was defined as molar ratio of  $\text{HfO}_2$ . The stoichiometric proportion of  $\text{HfCl}_4 \cdot 5\text{H}_2\text{O}$  and  $\text{Eu}(\text{CH}_3\text{COO})_3 \cdot \text{H}_2\text{O}$  were weighted and separately dissolved into de-ionized (DI) water with the help of magnetic stirring for 2 h. Ammonium hydroxide solution (0.1 M) was added to the mixture of  $\text{HfCl}_4 \cdot 5\text{H}_2\text{O}$  and  $\text{Eu}(\text{CH}_3\text{COO})_3 \cdot \text{H}_2\text{O}$  solutions. The pH of the solution was kept constant at  $7 \pm 0.4$  during the synthesis of all samples, resulting in precipitates. The sample was dried overnight at  $50^\circ \text{C}$  and the resulting precipitates were washed multiple times with DI water. The yellow color precipitates were finely crushed and for further annealing sample was placed in muffle furnace at  $600^\circ \text{C}$  for 2 h.

### 2.2 Theoretical methodology

The structural optimization of  $\text{Hf}_{1-x}\text{Eu}_x\text{O}_2$  ( $x = 0, 0.03, 0.06, 0.07$ ) in monoclinic and cubic phases were obtained



using first-principles method as implemented in VASP [36]. The relaxation was carried out within the parametrization of generalised gradient approximation (GGA) of Perdew–Burke–Ernzerhof (PBE) [37]. The convergence criteria for self-consistent field energy were set to  $10^{-7}$  eV and the atomic force relaxation was carried out till 0.01 eV/Å. The plane wave cutoff energy of 550 eV and a gamma centered k-mesh of  $9 \times 9 \times 7$  was used for the calculations. A  $2 \times 2 \times 2$  and  $1 \times 1 \times 7$  supercell with 96 and 84 atoms, respectively, was constructed to obtain the Eu doping concentrations, i.e., 0.03, 0.06, 0.07.

## 2.3 Characterization techniques

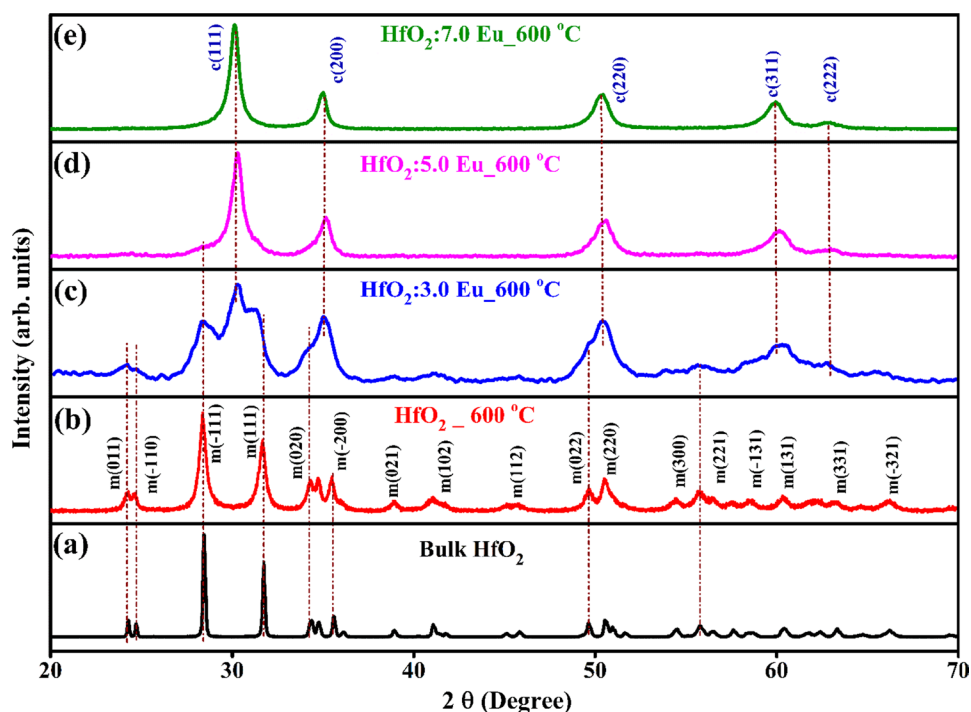
The X-ray diffraction (XRD) profile of the synthesized sample were registered using the high resolution Rigaku ultima IV X-ray diffractometer (Cu-K $\alpha$ ,  $\lambda = 1.5417$  Å). XANES spectra for the synthesized samples were collected at the soft X-ray beam line of Pohang Accelerator laboratory (PLS-II). Hf L $_1$ -edge XANES were collected at 1 D XRS KIST-PAL beamline, and O K-edge and Eu M $_{5,4}$ -edge XANES spectra were collected at the soft X-ray beam line (10D). Detailed procedure of XANES data collection, used gasses, chamber vacuum, monochromator details, and data normalization/background removal details are provided elsewhere [38]. PLE and PL spectra for Eu-doped HfO $_2$  samples were measured via employing Horiba spectrophotometer.

## 3 Results and discussion

### 3.1 X-ray diffraction and transmission electron microscopy (TEM)

Figure 1 shows the XRD data of different HfO $_2$  samples. Figure 1a shows the XRD patterns from reference/bulk HfO $_2$  powders (Aldrich, 99.99% pure). Figure 1b, c, d and e shows the XRD patterns of undoped HfO $_2$ , HfO $_2$ :3.0 mol% Eu, HfO $_2$ :5.0 mol% Eu, and HfO $_2$ :7.0 mol% Eu samples, respectively. It is noticeable that the XRD patterns of HfO $_2$  at 600 °C matches with the standard profile of HfO $_2$  (JCPDS card no. 78–0049) with a monoclinic structured unit cell (space group P2 $_1$ /c) with lattice parameters;  $a = 5.12$  Å,  $b = 5.17$  Å,  $c = 5.29$  Å,  $\alpha = \gamma = 90^\circ$  and  $\beta = 99.19^\circ$  [8, 10]. This strengthened the formation of monoclinic phase of HfO $_2$  at 600 °C. Noticeable changes are seen in the XRD patterns of Eu-doped samples. There is evolution of a few new peaks at  $\sim 30.2^\circ$ ,  $35.1^\circ$ ,  $50.3^\circ$  and  $60^\circ$  in the HfO $_2$ :3.0 mol% Eu along with the peaks of monoclinic structured HfO $_2$ . The intensity of newly evolved XRD peaks is improved in the HfO $_2$ :5.0 mol% Eu sample and the intensity of XRD peaks of monoclinic structured HfO $_2$  is significantly diminished. Eventually, the XRD peaks of monoclinic structured HfO $_2$  are eradicated and the intensity of XRD peaks at  $\sim 30.2^\circ$ ,  $35.1^\circ$ ,  $50.3^\circ$  and  $60^\circ$  is dominated the XRD patterns of HfO $_2$ :7.0 mol% Eu sample. The peak position of newly evolved XRD peaks ( $\sim 30.2^\circ$ ,  $35.1^\circ$ ,  $50.3^\circ$  and  $60^\circ$ ) is fairly matched with the previously reported XRD patterns

**Fig. 1** XRD patterns of pure and Eu-doped HfO $_2$  (at 3.0, 5.0 and 7.0 mol%)



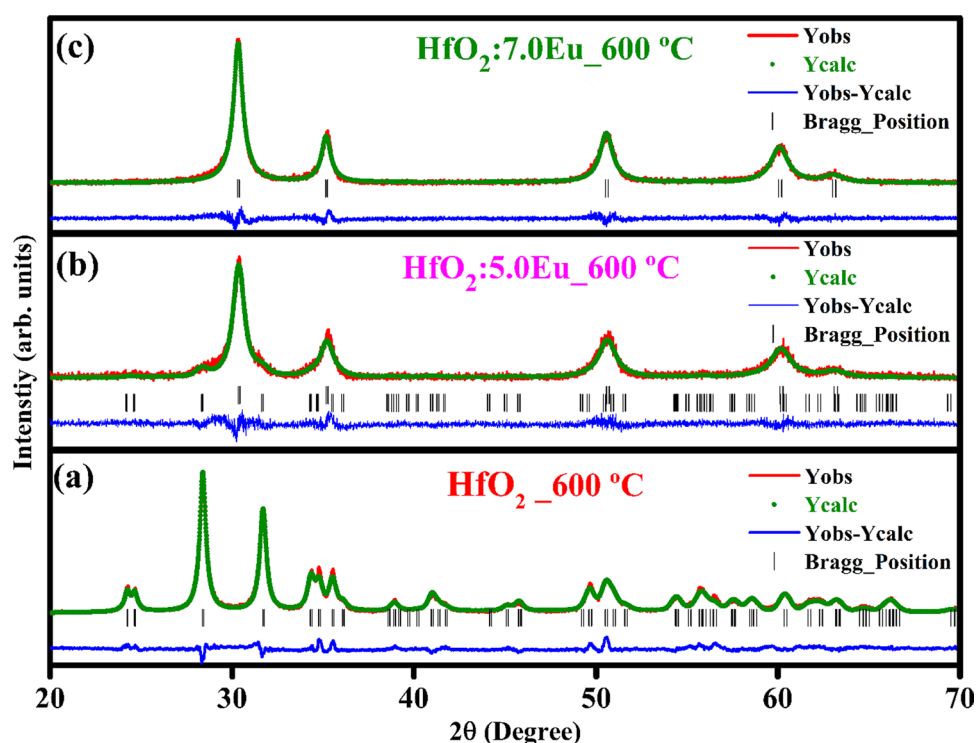
of cubic phase of  $\text{HfO}_2$  nanoparticles. Likewise, the JCDPF card no. 53–0560 also match with the XRD findings of  $\text{HfO}_2$ :7.0 mol% Eu sample and convey cubic phase of  $\text{HfO}_2$  (space group  $\text{Fm } \bar{3}m$   $a=b=c=5.105 \text{ \AA}$ ,  $\alpha=\beta=\gamma=90^\circ$ ).

For the analysis of the content of phases and more information regarding the XRD pattern of  $\text{HfO}_2$  samples. We have performed the rietveld refinement using the Fullprof suite with pseudo-Voigt shape function [39]. The rietveld refinement peaks fitted well with the XRD pattern, as shown in Fig. 2. With increase in the Eu concentration the peaks become more dominant and intense. The rietveld refinement parameters are summarised in Table 1. Average crystallite size is calculated using the Scherrer relation [8]:

$$D = \frac{0.9\lambda}{\beta \cos \theta} \quad (1)$$

where  $\lambda$  is the wavelength of the X-rays,  $D$  is the crystallite size,  $\beta$  is the full width at half maximum of the diffraction peak). The calculated crystallite size from  $\text{HfO}_2$ ,  $\text{HfO}_2$ :3.0 mol% Eu,  $\text{HfO}_2$ :5.0 mol% Eu, and  $\text{HfO}_2$ :7.0 mol% Eu samples were observed 32.99 nm, 11.32 nm, 12.69 nm and 13.60 nm, respectively. The slope of W–H plot (supplementary information Fig. S1) was used to measure the strain of the all studied samples. The strain for the synthesized  $\text{HfO}_2$ , 3.0, 5.0 and 7.0 mol% Eu-doped  $\text{HfO}_2$  samples were found to be 2.76, 8.34, 7.39, 4.52, respectively. The strain is reduced in the samples are having single phase (or nearly the

**Fig. 2** Rietveld refinement of pure and Eu-doped  $\text{HfO}_2$  (at 5.0 and 7.0 mol%)



**Table 1** Refined parameters, convergence indicator (chi square) and phase for pure and doped  $\text{HfO}_2$

Sample	$R_p$	$R_{wp}$	$R_{exp}$	$\chi^2$	GOF	Phase (in %)	Normalized occupancy
$\text{HfO}_2$	16.0	17.5	16.2	1.26	1.08	$m-100$	Hf—1.04 O1—1.22 O2—0.98
$\text{HfO}_2$ :5.0 Eu	24.6	28.7	26.60	1.17	1.07	$c-96.58$ $m-3.42$	Hf—0.95 O1—0.97 O2—1.12 Eu—0.05
$\text{HfO}_2$ :7.0 Eu	20.1	20.9	18.33	1.29	1.14	$c-100$	Hf—0.93 O1—0.92 O2—1.07 Eu—0.07



single-phase nature) and leads higher XRD peak intensity. The dislocation density was correlated with the crystallite size as follows:

$$\delta = \frac{1}{D^2} \quad (2)$$

where  $\delta$  is dislocation density and  $D$  is the crystallite size. The measured values of  $\delta$  for HfO<sub>2</sub>, 3.0, 5.0 and 7.0 mol% Eu-doped HfO<sub>2</sub> samples are 0.0009, 0.007, 0.006, and 0.005, respectively. The structural transformation, monoclinic to cubic phase, may arise due to the substitution of Hf<sup>4+</sup> (ionic radii=0.078 nm) by the larger sized Eu<sup>3+</sup> ions (ionic radii=0.106 nm), which promotes the formation of oxygen vacancies. The strain produced in the lattice, due to larger sized Eu<sup>3+</sup> ions, may push O<sup>2-</sup> ions towards the Hf ions leading to Hf–O8 kind of polyhedron formation in the HfO<sub>2</sub>:Eu<sup>3+</sup> compound and leading to alteration in the lattice parameters of the resultant unit cell, which cause the arrangement of ions in the cubic structure [7].

Figure 3a, b shows the TEM image and elemental mapping, respectively, from HfO<sub>2</sub> sample. Figures S2 and S3 (supplementary information) show the elemental mapping pure and HfO<sub>2</sub>:7.0 mol% Eu, respectively. It is noticeable from Fig. 3a, b that both of the samples have exhibited the unusual morphology of agglomerated spherical particles. The elemental mapping profiles convey the even distribution of elements and are nullifying the formation of segregated phase of a particular element. XRD results have also ruled out the formation of other phases and strengthened the formation of single-phase compound with high crystallinity.

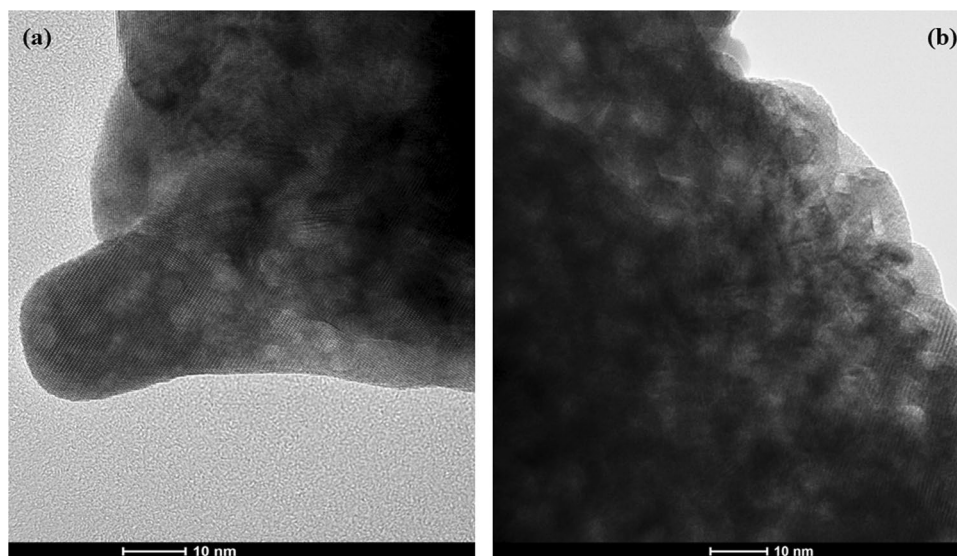
### 3.2 First-principles study of structural phase transition

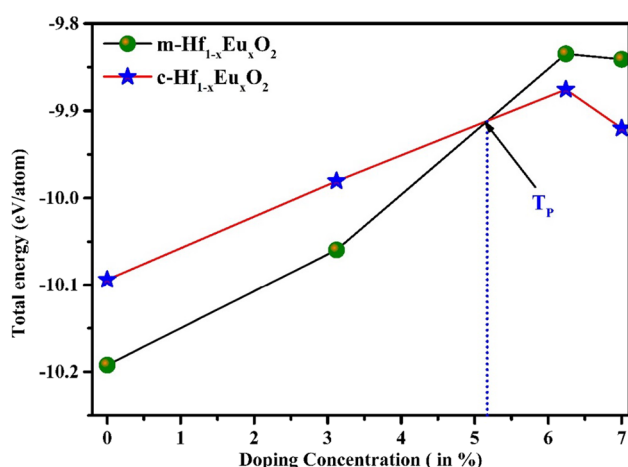
The *first-principles* calculations have been used to analyse the effect of doping on the structural phase transition in HfO<sub>2</sub>. The calculated atomic positions and cell parameters of monoclinic as well as cubic phases are in good agreement with the experimental findings (Table 2). The variation of the total energy vs doping concentration is shown in Fig. 4. The crossover point in the energy vs doping concentration is called as transition point ( $T_p$ ). It clearly shows that Eu doping concentrations up to  $T_p$ , i.e., 5.11% (close to experimental result of 5%) acquire minimum amount of energy for the m-HfO<sub>2</sub> resulting in a stable structure. After  $T_p$ , the cubic structure is more stable than monoclinic one. The difference in the energies of monoclinic and cubic phases increases with higher doping concentration. Hence, we theoretically validate the phase transition and stability of cubic phase after  $T_p$  which is in excellent agreement with our experimental results.

**Table 2** Theoretical and experimental lattice parameters of HfO<sub>2</sub>

HfO <sub>2</sub> phases	Space group	Lattice parameters (in Å)	
		Theoretical	Experimental
Monoclinic	P2 <sub>1</sub> /c	$a=5.14$ , $b=5.19$ , $c=5.33$	$a=5.12$ , $b=5.17$ , $c=5.29$
Cubic	Fm $\bar{3}$ m	$a=5.09$	$a=5.104$

**Fig. 3** a, b TEM image pure and HfO<sub>2</sub>:7.0 mol% Eu, respectively





**Fig. 4** Total energy vs doping concentration of  $\text{Hf}_{1-x}\text{Eu}_x\text{O}_2$  (at 0, 3%, 6% and 7%)

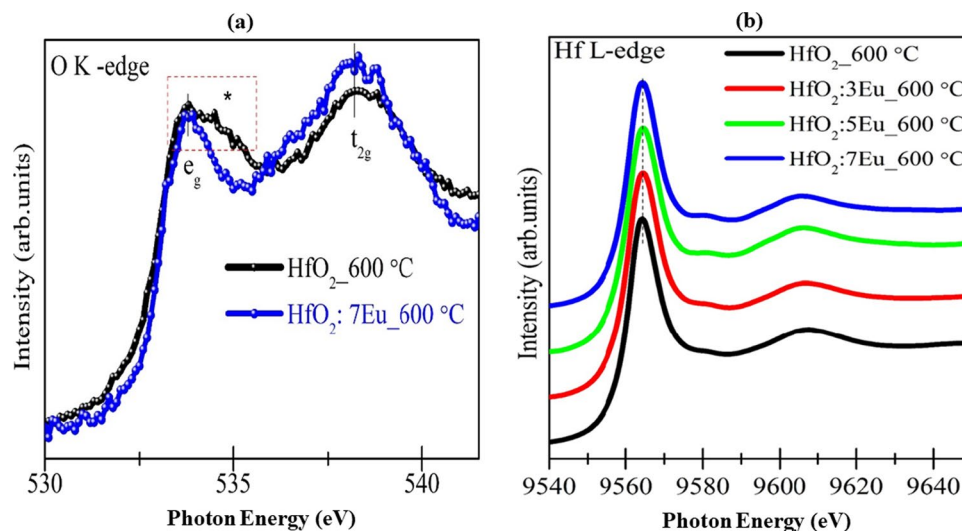
### 3.3 X-ray absorption spectroscopy

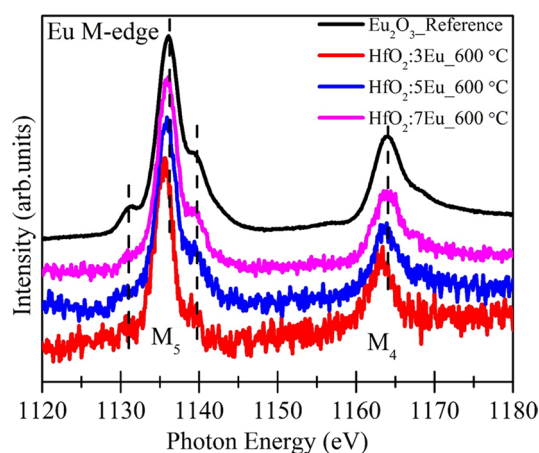
Figure 5a shows the O K-edge spectra of  $\text{HfO}_2$  and  $\text{HfO}_2$ :7.0 mol% Eu samples. In previous reports, it has been reported that the cubic phasic  $\text{CeO}_2$  and pseudo-cubic  $\text{HfO}_2$  and  $\text{ZrO}_2$  compounds experience vivid crystal field effects [40, 41]. The  $dx^2-y^2$  and  $dz^2$  orbitals (i.e.,  $e_g$  orbitals) of metal elements align between the O ligand fields and experience weak interaction on the other hand metal's  $d_{xy}$ ,  $d_{xz}$ , and  $d_{yz}$  orbitals (i.e.,  $t_{2g}$  orbitals) directed towards the oxygen ligands, and experience strong interaction. The ligand field interaction impulses the  $t_{2g}$  orbitals at higher energy and the  $e_g$  orbitals are placed at lower energy [42]. It is noticeable from the Fig. 5a that both of the samples have exhibited two sharp features at 533.8 eV and 538.2 eV for  $e_g$  and  $t_{2g}$  orbitals, respectively. O K edge spectra of  $\text{HfO}_2$  is quite similar to the spectra of bulk  $\text{HfO}_2$  with monoclinic

crystal structure [41, 43]. In case of monoclinic unit cell of  $\text{HfO}_2$ , there is no the centre of symmetry in M–O7 (M is metal and O is Oxygen atoms) polyhedra and the different  $d$  orbitals interact, uncommonly, with the crystal field of O ligands. In case of monoclinic structured  $\text{ZrO}_2$ , the orbital degeneracy was uninvolved and resulted complex splitting of  $d$  orbitals [41]. The monoclinic phase of  $\text{HfO}_2$  has been reported with fragmented  $e_g$  and sharp  $t_{2g}$  spectral features in the O K-edge XANES spectra. On the other hand, the tetragonal phase of  $\text{HfO}_2$  has reported with sharp  $e_g$  and splitted  $t_{2g}$  features in the O K-edge XANES spectra [43]. The M–O8 polyhedra also exists in cubic phasic  $\text{HfO}_2$  and the crystal field splitting values (i.e., energy separation between  $e_g$  and  $t_{2g}$  spectral features) remains the same for both cubic and tetragonal phase of  $\text{HfO}_2$  [43, 44]. In the present case the O K-edge of  $\text{HfO}_2$ :7.0 mol% Eu sample exhibited the splitting/broadening in the  $e_g$  peak (marked by \*) and better symmetric  $t_{2g}$  spectral feature (within the spectral resolution of used beam line). On the other hand, the  $\text{HfO}_2$ :7.0 mol% Eu sample shows the sharp or less splitted  $e_g$  spectral feature and broadened  $t_{2g}$  feature. This strengthened the monoclinic phase in  $\text{HfO}_2$ :7.0 mol% Eu sample and cubic phase in  $\text{HfO}_2$ :7.0 mol% Eu sample. XRD results have confirmed the formation of monoclinic phase in  $\text{HfO}_2$ :7.0 mol% Eu sample and stabilization of cubic phase in  $\text{HfO}_2$ :7.0 mol% Eu sample. Therefore, a difference has been observed in the O K-edge XANES of both of the samples which arises due to the distinct hybridization of frontier orbitals of Hf and O atoms in metal–oxygen polyhedra.

Figure 5b shows the Hf L-edge XANES spectra of  $\text{HfO}_2$ ,  $\text{HfO}_2$ :3.0 mol% Eu,  $\text{HfO}_2$ :5.0 mol% Eu,  $\text{HfO}_2$ :7.0 mol% Eu samples. The Hf  $L_1$ -edge XANES spectrum arises due to the 2 s core level transitions to the p-type unoccupied states. In Fig. 6, there is no measurable change in the pre-edge, main edge and post edge features of Eu-doped samples. This

**Fig. 5** **a** O K-edge XANES spectra of  $\text{HfO}_2$  and  $\text{HfO}_2$ :7.0 mol% Eu **b** Hf L-edge XANES spectra of  $\text{HfO}_2$  and  $\text{HfO}_2$ :3.0, 5.0 and 7.0 mol% Eu





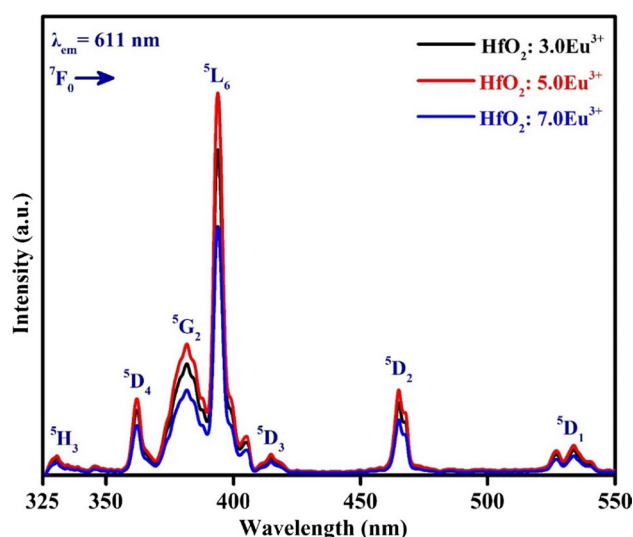
**Fig. 6** Eu  $M_{5,4}$ -edge XANES spectra of 3.0 mol%, 5.0 mol% and 7.0 mol% Eu-doped  $\text{HfO}_2$

indicates oxidation state of  $\text{Hf}^{4+}$  ions and local coordination of Hf and O atoms is not affected under the Eu doping conditions in  $\text{HfO}_2$  lattice.

To understand the valence state of Eu ions in  $\text{HfO}_2$ , the Eu  $M_{5,4}$  edge XANES measurements are performed and are shown in Fig. 6. Reference  $\text{Eu}_2\text{O}_3$  sample was also scanned under the same conditions for collecting the Eu  $M_{5,4}$  edge XANES spectrum. It is noticeable that main spectral features of reference  $\text{Eu}_2\text{O}_3$  and Eu-doped  $\text{HfO}_2$  samples are quite similar and showing two intense peaks at  $\sim 1136$  eV and  $\sim 1164$  eV, which are corresponding to the  $M_5$  and  $M_4$  edges, respectively. The  $M_5$  and  $M_4$  edges originate from the Eu  $3d_{5/2}$  and  $3d_{3/2}$  electronic transitions to the 4f states, respectively, and their intensities are proportional to the density of unoccupied 4f states [38]. The energy difference between  $M_5$  and  $M_4$  edges is  $\sim 28$  eV and is consistent with previous reports of  $\text{Eu}^{3+}$  ions containing sample [38]. A closer view of Fig. 6 conveys that the lower Eu concentration-doped  $\text{HfO}_2$  samples are showing  $M_{5,4}$  edge features at lower energy. This indicates that, at lower Eu doping concentrations, the Eu ions may coordinate with oxygen atom with  $\text{EuO}$  kind of geometry with +2 valence state. However, we have not seen distinct  $\text{EuO}$  crystalline phases in the XRD data. Therefore, it is anticipated that  $\text{EuO}$  kind of clusters, if formed, are very diluted in  $\text{HfO}_2$  lattice or present in the amorphous form. The  $M_{5,4}$  peaks of higher concentration Eu-doped  $\text{HfO}_2$  sample are fairly matched with the spectral features of reference  $\text{Eu}_2\text{O}_3$  sample and confirm the dominant  $\text{Eu}^{3+}$  ions in  $\text{HfO}_2$ :7 mol% Eu sample.

### 3.4 Photoluminescence properties of $\text{Eu}^{3+}$ -doped $\text{HfO}_2$

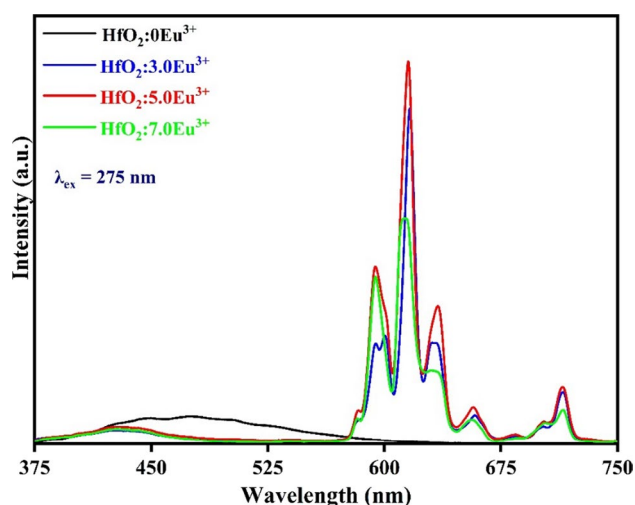
The photoluminescence excitation (PLE) spectra for  $\text{HfO}_2$  doped with varying  $\text{Eu}^{3+}$  ion concentrations were measured



**Fig. 7** Photoluminescence excitation spectra for  $\text{HfO}_2$  doped with varying  $\text{Eu}^{3+}$  ion concentrations under 611 nm emission wavelength

at room temperature in the spectral wavelength range from 325 to 550 nm by monitoring 611 nm emission wavelength, as demonstrated in Fig. 7. The PLE spectra consists of several excitation peaks related to f–f electronic transitions located at 331, 361, 383, 394, 414, 465 and 532 nm initiating from ground energy level ( $^7F_0$ ) to various excited energy levels, such as  $^5H_3$ ,  $^5D_4$ ,  $^5G_2$ ,  $^5L_6$ ,  $^5D_3$ ,  $^5D_2$  and  $^5D_1$  of  $\text{Eu}^{3+}$  ions, respectively. Among all these excitation peaks, the sharp intense excitation peak was observed at 394 nm corresponds to  $^7F_0 \rightarrow ^5L_6$  transition, which well matches the emission profile of the n-UV LED chip. The various excitation peaks, including 361, 383, 394 and 465 nm were opted to record the emission spectral profiles of  $\text{Eu}^{3+}$ -doped  $\text{HfO}_2$  samples.

The emission spectra of  $\text{HfO}_2$  with varying  $\text{Eu}^{3+}$  ion concentrations were observed in the 375–750 nm spectral range by monitoring the excitation wavelength of 275 nm presented in Fig. 8. In addition, the emission spectra for  $\text{HfO}_2$  doped with differing  $\text{Eu}^{3+}$  ion concentrations were measured in the wavelength region from 550 to 750 nm at room temperature under various excitation wavelengths, including 361, 383, 394 and 465 nm, respectively, as shown in Fig. 9a–d. Using above-mentioned different excitations, the emission spectra of  $\text{Eu}^{3+}$ -doped  $\text{HfO}_2$  samples reveal that several emission peaks starting from higher energy excited level ( $^5D_0$ ) to various lower energy levels, such as  $^7F_J$  (where  $J=0, 1, 2, 3$  and 4), which represents the emission peak observed at 578, 587, 611, 654 and 710 nm, respectively. A minute variation in the emission peak intensity with varying the different excitation wavelengths was noticed. The emission peak observed at 611 nm corresponds to the  $^5D_0 \rightarrow ^7F_2$  transition and was highly intense as compared to other



**Fig. 8** Emission spectra of  $\text{Eu}^{3+}$ -doped  $\text{HfO}_2$  samples pumping under excitation wavelength of 275 nm

emission peaks, which ascribed to forced electric dipole (ED) transition. The forced ED transition is hypersensitive in behaviour and obeys the following selection rule, such as  $\Delta J = 2$  [43, 44]. Furthermore, the emission intensity of the  $^5\text{D}_0 \rightarrow ^7\text{F}_2$  transition remains dominant through the crystal field strength of the local environment [45, 46]. The emission peak located at 587 nm corresponds to the  $^5\text{D}_0 \rightarrow ^7\text{F}_1$  transition was a magnetic dipole (MD) transition in behaviour and following the Laporte selection rule, such as  $\Delta J = 1$ , which ascribed to the insensitive to the crystal field environment of the  $\text{Eu}^{3+}$  ions in the synthesized samples [46]. The emission intensity ratio of ED ( $^5\text{D}_0 \rightarrow ^7\text{F}_2$ ) to MD ( $^5\text{D}_0 \rightarrow ^7\text{F}_1$ ) is referred to as the asymmetric fraction, which explains the degree of distortion concerning the inversion symmetry of the  $\text{Eu}^{3+}$  ions. In the present work, the emission intensity of ED transition was greater than the emission intensity of MD transition, which indicates the presence of  $\text{Eu}^{3+}$  ions at sites without inversion symmetry. The emission intensity varies with  $\text{Eu}^{3+}$  ion concentrations in the synthesized samples, as shown in inset of Fig. 9a–d. As increasing the  $\text{Eu}^{3+}$  concentration in the synthesized samples from 3.0 to 5.0 mol%, the emission intensity increases and beyond that the emission intensity decreases with an increase in the concentration of  $\text{Eu}^{3+}$  ions up to 7.0 mol%. This behaviour occurs because of concentration quenching phenomena. The quenching phenomena can be taken place owing to a decrease in the distance between the dopant ( $\text{Eu}^{3+}$ ) ions, which increases the non-radiative energy transfer (NRET) and multipolar interaction or resonant energy transfer (RET). The optimal dopant ion concentration in  $\text{Eu}^{3+}$ -doped  $\text{HfO}_2$  samples was discovered to be 5.0 mol%.

Furthermore, the partial energy level diagram that includes excitation and emission as well as possible some

other type of non-radiative transition (NRT) of  $\text{Eu}^{3+}$ -doped  $\text{HfO}_2$  sample are presented in Fig. 10. By absorbing the specific photon energy, the  $\text{Eu}^{3+}$  ions at ground energy level ( $^7\text{F}_0$ ) were excited to the higher energy level ( $^5\text{L}_6$ ) of  $\text{Eu}^{3+}$  ions. The excited  $\text{Eu}^{3+}$  ions have been de-excited by emitting the phonon or photon via NRT before carrying out the downward radiative transitions, including  $^5\text{D}_0 \rightarrow ^7\text{F}_J$  ( $J = 0, 1, 2, 3$  and 4) of  $\text{Eu}^{3+}$  ions.

### 3.4.1 Estimation of CIE coordinates, correlated color temperature (CCT) and color purity

The Commission International de l'Eclairage (CIE) 1931 parameters, such as ( $x, y$ ) color coordinates, CCT values may be calculated via employing the emission data to understand the colorimetric features of the prepared samples. To reveal the color tone of the emission, the calculated color coordinates were represented on the CIE 1931 graph. The calculated color coordinates shown in the CIE chromaticity graph for 5.0 mol%  $\text{Eu}^{3+}$ -doped  $\text{HfO}_2$  sample (i.e., optimized 5.0 mol% doped) excited along with different wavelengths, including 361 nm, 383 nm, 393 nm and 465 nm situated in the reddish region (shown in Fig. 11).

The estimated color coordinates for optimized  $\text{HfO}_2:5.0 \text{ mol\% } \text{Eu}^{3+}$  were found close to the standard National television Standard Committee (NTSC) and the red-emitting commercial phosphors. Furthermore, the CCT is a crucial parameter that reveals the distinct color of the light emitted by the luminous components and estimated in kelvin (K). The McCamy empirical relation has been used to compute the CCT values for  $\text{HfO}_2:x\text{Eu}^{3+}$  samples [47];

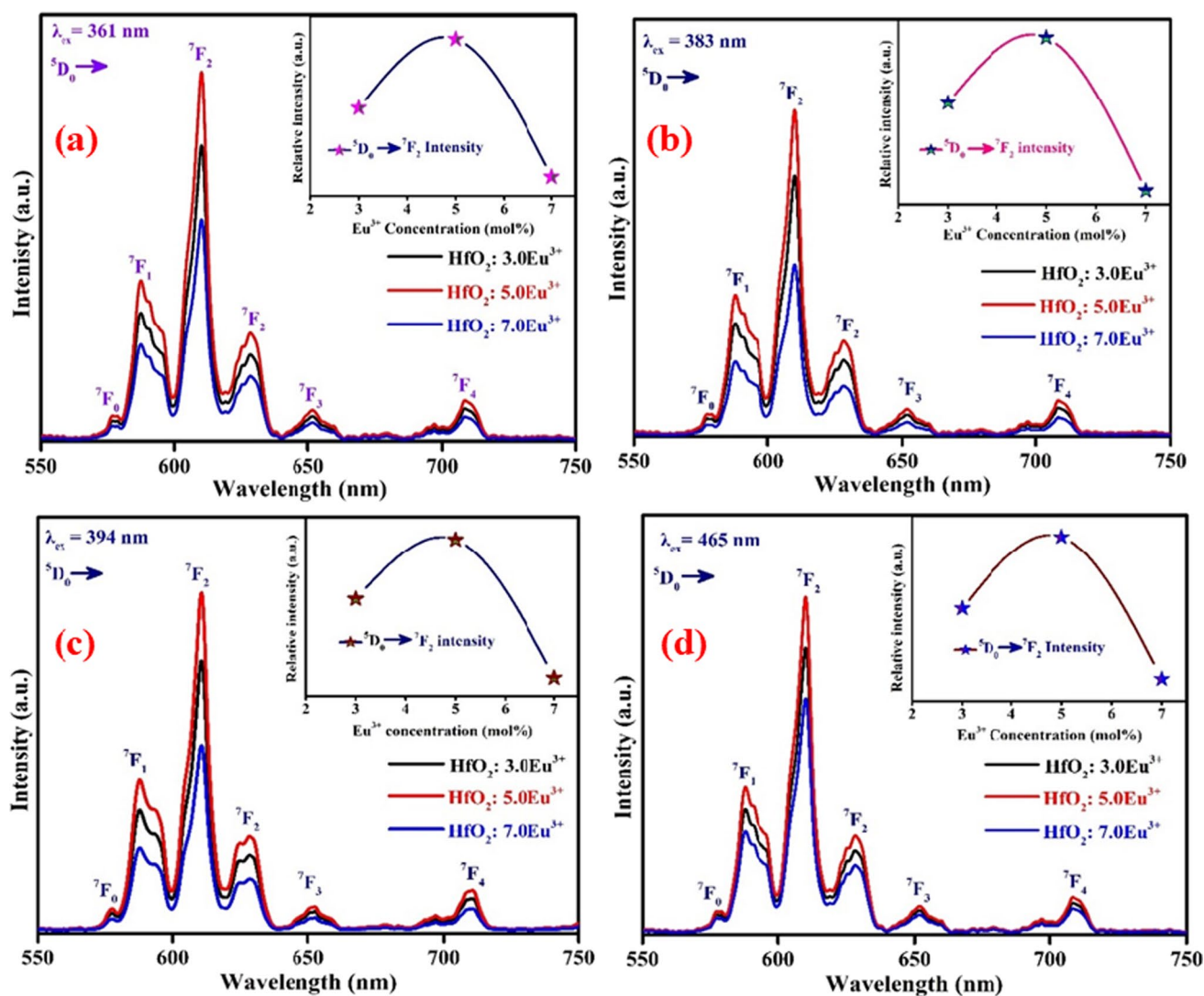
$$\text{CCT} = -449n^3 + 3525n^2 - 6823n + 5520.3 \quad (3)$$

in above relation  $n = \frac{X-X_e}{Y-Y_e}$ ;  $X_e = 0.332$  and  $Y_e = 0.186$  is the epicentre, respectively. The estimated CCT for optimized 5.0 mol%  $\text{Eu}^{3+}$ -doped  $\text{HfO}_2$  sample was found to be 1221 K, 1225 K, 1232 K and 1224 K under different excitation wavelengths. The estimated CCT values for the optimized  $\text{HfO}_2:x\text{Eu}^{3+}$  ( $x = 5.0 \text{ mol\%}$ ) samples are below 4000 K signifying the aptness of  $\text{HfO}_2:5.0 \text{ mol\% } \text{Eu}^{3+}$  sample for warm lighting applications. The CP of the as-synthesized  $\text{HfO}_2:x\text{Eu}^{3+}$  samples have been evaluated from the given relation below [47]:

$$\text{CP} = \frac{\sqrt{(x - x_{ee})^2 + (y - y_{ee})^2}}{\sqrt{(x_d - x_{ee})^2 + (y_d - y_{ee})^2}} \quad (4)$$

in the above relation, ( $x, y$ ) ( $x_{ee}, y_{ee}$ ) and ( $x_d, y_d$ ) shows the chromaticity coordinates, equal energy point and dominant wavelength of as-synthesized samples, respectively. The CP for the optimized 5.0 mol%  $\text{Eu}^{3+}$ -doped  $\text{HfO}_2$  sample was





**Fig. 9** Emission spectra of  $\text{Eu}^{3+}$ -doped  $\text{HfO}_2$  samples pumping under various excitation wavelengths, including 361 nm, 383 nm, 394 nm and 465 nm

found to 97.2%, 97.8%, 98.5% and 98.1% under different excitation wavelengths. Hence, the aforementioned results show that the direct utility of  $\text{Eu}^{3+}$ -doped  $\text{HfO}_2$  samples may be a suitable candidate for the red emitting element for photonic device applications.

## 4 Conclusion

In the present study,  $\text{Eu}$ -doped  $\text{HfO}_2$  nanoparticles have synthesised using chemical co-precipitation method and annealed at 600 °C. X-ray diffraction results have revealed the structure evolution of  $\text{HfO}_2$  from the monoclinic phase to cubic phase. A mixed phase formation has occurred

at lower concentration and a dominant cubic phase achieved at 5.0 mol% and a perfect cubic phase achieved at 7.0 mol% doping of  $\text{Eu}$  in  $\text{HfO}_2$ . The theoretical calculations have validated the phase transition and stability of cubic phase at ~5% which is in excellent agreement with our experimental results. XANES spectra of  $\text{Hf}$  L-edge and  $\text{Eu}$  M-edge clearly depicted the  $\text{Hf}^{4+}$  and  $\text{Eu}^{3+}$  ions environment of doped  $\text{HfO}_2$  nanoparticles. Oxygen K-edge has shown the diverse hybridization of O 2p orbitals in M–O7 (for monoclinic) and M–O8 (for cubic) polyhedra of  $\text{HfO}_2$ . The PL emission in  $\text{Eu}^{3+}$ -doped  $\text{HfO}_2$  nanocrystals have observed in red region under different excitations with a high color purity, which may be exploited in solid state lighting-based applications.

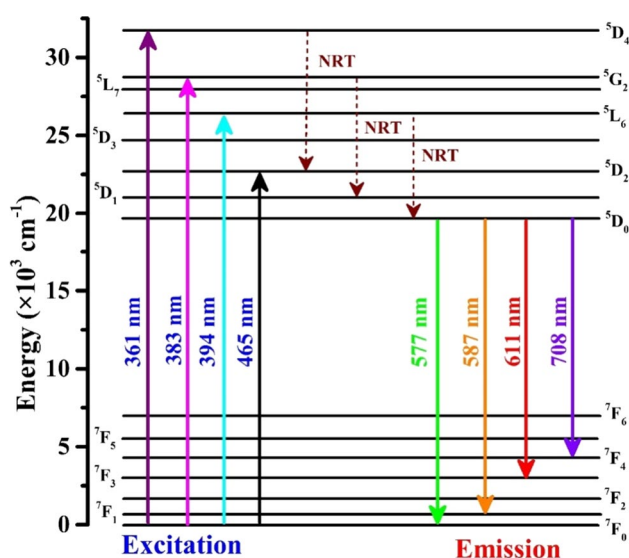


Fig. 10 Partial energy level diagram of  $\text{Eu}^{3+}$ -doped  $\text{HfO}_2$  samples

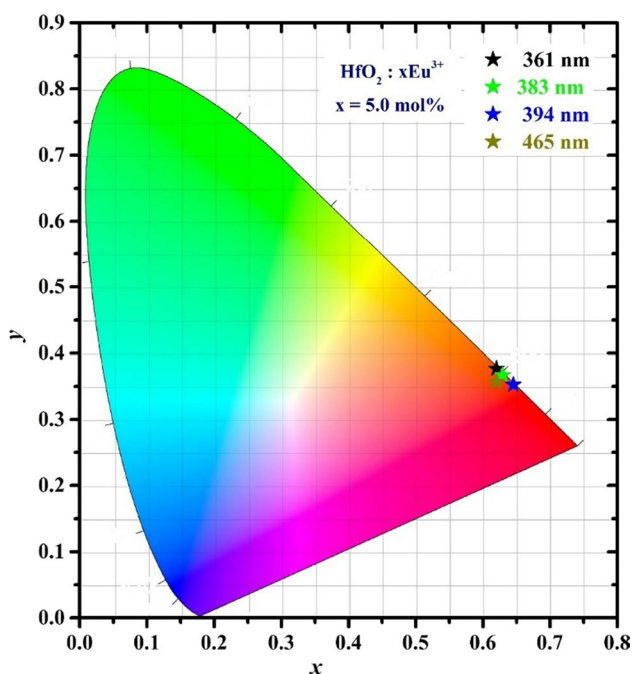


Fig. 11 CIE chromaticity diagram for optimized  $\text{HfO}_2$ :5.0 mol%  $\text{Eu}^{3+}$  sample under various excitation wavelengths, including 361 nm, 383 nm, 394 nm and 465 nm

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s00339-023-06997-0>.

**Author contributions** RK: investigation, data curation, writing—original draft, preparation. AS and KHC: experimental and analysis support for X-ray absorption spectroscopy study. AK and JK: experimental support during synthesis of samples and manuscript preparation. RK: software, methodology. SOW: experimental support for XRD and

TEM. MS: supervision, conceptualization, methodology, writing—review and editing. AV: supervision, visualization, validation, writing—review and editing.

**Data availability** Data available upon request.

## Declarations

**Conflict of interest** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

1. S. Liu, X. Li, X. Yu, Z. Chang, P. Che, J. Zhou, W. Li, A route for white LED package using luminescent low-temperature co-fired ceramics. *J. Alloy. Compd.* **655**, 203–207 (2016)
2. Z. Xia, Z. Xu, M. Chen, Q. Liu, Recent developments in the new inorganic solid-state LED phosphors. *Dalton Trans.* **45**, 11214–11232 (2016)
3. J. Sun, X. Zhang, Z. Xia, H. Du, Luminescent properties of  $\text{LiBaPO}_4$ : RE (RE =  $\text{Eu}^{2+}$ ,  $\text{Tb}^{3+}$ ,  $\text{Sm}^{3+}$ ) phosphors for white light-emitting diodes. *J. Appl. Phys.* **111**, 013101 (2012)
4. Z. Wang, J. Ha, Y.H. Kim, W.B. Im, J. McKittrick, S.P. Ong, Mining unexplored chemistries for phosphors for high-color-quality white-light-emitting diodes. *Joule* **2**, 914–926 (2018)
5. Manju, M. Jain, P. Vashishtha, G. Gupta, M. Gupta, P. Rajput, A. Vij, A. Thakur, Charge transfer-induced fast blue emission in  $\text{SrZnO}_2$ :Ce. *Appl. Phys. Lett.* **119**, 121108 (2021)
6. K.R.V. Babu, C.G. Renuka, R.B. Basavaraj, G.P. Darshan, H. Nagabhushana, One pot synthesis of  $\text{TiO}_2$ : $\text{Eu}^{3+}$  hierarchical structures as a highly specific luminescent sensing probe for the visualization of latent fingerprints. *J. Rare Earths* **37**, 134–144 (2019)
7. L.X. Lovisa, V.D. Araújo, R.L. Tranquilin, E. Longo, M.S. Li, C.A. Paskocimas, M.R.D. Bomio, F.V. Motta, White photoluminescence emission from  $\text{ZrO}_2$  co-doped with  $\text{Eu}^{3+}$ ,  $\text{Tb}^{3+}$  and  $\text{Tm}^{3+}$ . *J. Alloys Comp.* **674**, 245–251 (2016)
8. S. Kumar, T. Dehury, C. Rath, Stabilization of cubic phase at room temperature and photoluminescence properties of Dy and Sm co-doped  $\text{HfO}_2$  nanoparticles. *ECS J. Solid State Sci. Technol.* **10**, 081009 (2021)
9. T.N.L. Tran, A. Szczurek, A. Lukowiak, A. Chiasera, A review on rare-earth activated  $\text{SnO}_2$ -based photonic structures: Synthesis, fabrication and photoluminescence properties. *Opt. Mater.* **13**, 100140 (2022)
10. R. Kumar, A. Vij, M. Singh, Defects assisted luminescence in  $m$ - $\text{HfO}_2$  nanocrystals: an experimental and theoretical study. *Optik* **248**, 168121 (2021)
11. R. Kumar, R. Kumar, M. Singh, D. Meena, A. Vij, Carrier concentration mediated enhancement in thermoelectric performance of various polymorphs of hafnium oxide: a plausible material for high temperature thermoelectric energy harvesting application. *J. Phys. D: Appl. Phys.* **55**, 495302 (2022)
12. M. Baik, H.K. Kang, Y.S. Kang, K.S. Jeong, Y. An, S. Choi, H. Kim, J.D. Song, M.H. Cho, Electrical properties and thermal stability in stack structure of  $\text{HfO}_2/\text{Al}_2\text{O}_3/\text{InSb}$  by atomic layer deposition. *Sci. Rep.* **7**, 11337 (2017)
13. H.L. Leiming, F.S. Guan, F. Peng, Z. Zhang, H. Chen, W. Zhang, C. Lu, Insights into the bond behavior and mechanical properties of hafnium carbide under high pressure and high temperature. *Inorg. Chem.* **60**, 515–524 (2021)



14. P. Torchio, A. Gatto, M. Alvisi, G. Albrand, N. Kaiser, C. Amra, High-reflectivity  $\text{HfO}_2/\text{SiO}_2$  ultraviolet mirrors. *Appl. Opt.* **41**, 3256–3261 (2002)
15. M. Kirm, J. Aarik, M. Jurgens, I. Sildos, Thin films of  $\text{HfO}_2$  and  $\text{ZrO}_2$  as potential scintillators. *Nucl. Instrum. Methods Phys. Res. A* **5337**, 251–255 (2004)
16. X. Luo, W. Zhou, S.V. Ushakov, A. Navrotsky, A.A. Demkov, Monoclinic to tetragonal transformations in hafnia and zirconia: a combined calorimetric and density functional study. *Phys. Rev. B* **80**, 134119 (2009)
17. T. Tobase, A. Yoshiasa, H. Arima, K. Sugiyama, O. Ohtaka, T. Nakatani, K. Funakoshi, S. Kohara, Pre-transitional behavior in tetragonal to cubic phase transition in  $\text{HfO}_2$  revealed by high temperature diffraction experiments. *Phys. Status Solidi B* **255**, 1800090 (2018)
18. N. Kumar, B.P.A. George, H. Abrahamse, V. Parashar, S.S. Ray, J.C. Ngila, A novel approach to low-temperature synthesis of cubic  $\text{HfO}_2$  nanostructures and their cytotoxicity. *Sci. Rep.* **7**, 9351 (2017)
19. T. Song, H. Tan, N. Dix, R. Moalla, G. Saint-Girons, R. Bachelet, F. Sánchez, I. Fina, Stabilization of ferroelectric phase in epitaxial  $\text{Hf}_{1-x}\text{Zr}_x\text{O}_2$  enabling coexistence of ferroelectric and enhanced piezoelectric properties. *ACS Appl. Electron. Mater.* **3**, 2106–2113 (2021)
20. S. Phokriyal, S. Biswas, Tuning of dielectric properties in Ti-Doped granular  $\text{HfO}_2$  nanoparticles for high-k applications. *Ceram. Int.* **48**, 11199–11208 (2022)
21. F.H. Borges, D.S.D.H. Oliveira, G.P. Hernandez, S.J.L. Ribeiro, R.R. Gonçalves, Highly red luminescent stabilized tetragonal rare earth-doped  $\text{HfO}_2$  crystalline ceramics prepared by sol-gel. *Opt. Mater. X* **16**, 100206 (2022)
22. E.R. Andrievskaya, Phase equilibria in the refractory oxide systems of zirconia, hafnia and yttria with rare-earth oxides. *J. Eur. Ceram. Soc.* **28**, 2363–2388 (2008)
23. A. Lauria, I. Villa, M. Fasoli, M. Niederberger, A. Vedda, Multifunctional role of rare earth doping in optical materials: non-aqueous sol-gel synthesis of stabilized cubic  $\text{HfO}_2$  luminescent nanoparticles. *ACS Nano* **7**, 7041–7052 (2013)
24. S. Kumar, S.B. Rai, C. Rath, Latent fingerprint imaging using Dy and Sm codoped  $\text{HfO}_2$  nanophosphors: structure and luminescence properties. *Part. Part. Syst. Character.* **36**, 1–11 (2019)
25. N. Sekar, B. Ganesan, H.R.A.S. Khilafath, P. Aruna, S. Ganesan, Synthesis and characterization of  $\text{Gd}^{3+}$  Doped  $\text{HfO}_2$  nanoparticles for radiotherapy applications. *J. Nanosci. Nanotechnol.* **20**, 819–827 (2020)
26. M.V. Ibañez, C.L. Luyer, O. Marty, J. Mugnier, Annealing and doping effects on the structure of europium-doped  $\text{HfO}_2$  sol-gel material. *Opt. Mater.* **24**, 51–57 (2003)
27. C.K. Lee, E. Cho, H.-S. Lee, C.S. Hwang, S. Han, First-principles study on doping and phase stability of  $\text{HfO}_2$ . *Phys. Rev. B* **78**, 012102 (2008)
28. E. Montes, P. Cerón, T.R. Montalvo, J. Guzmán, M.G. Hipólito, A.B.S. Guzmán, R.G. Salcedo, C. Falcony, Thermoluminescent characterization of  $\text{HfO}_2:\text{Tb}^{3+}$  synthesized by hydrothermal route. *Appl. Rad. Isotop.* **83**, 196–199 (2014)
29. M. Nanda, S. Tripathia, P. Rajputa, M. Kumar, Y. Kumar, S.K. Mandald, R. Urkuded, M. Guptae, A. Dawarg, S. Ojhag, S.K. Raif, S.N. Jha, Different polymorphs of Y doped  $\text{HfO}_2$  epitaxial thin films: Insights into structural, electronic and optical properties. *J. Alloys Compd.* **928**, 167099 (2022)
30. A. Sharma, M. Varshney, H.J. Shin, K.H. Chae, S.O. Won, XANES, EXAFS and photoluminescence investigation on the amorphous Eu:  $\text{HfO}_2$ . *Mol. Biomol. Spectrosc.* **173**, 549–555 (2017)
31. S. Stojadinović, N. Tadić, R. Vasilic, Photoluminescence properties of  $\text{Er}^{3+}/\text{Yb}^{3+}$  doped  $\text{ZrO}_2$  coatings formed by plasma electrolytic oxidation. *J. Lumin.* **208**, 296–301 (2019)
32. T. Dehury, S. Kumar, C. Rath, Structural transformation and band-gap engineering by doping Pr in  $\text{HfO}_2$  nanoparticles. *Mater. Lett.* **302**, 130413 (2021)
33. C.L.O. Romero, J.C. Flores, J.A. Hernández, E.G. Camarillo, E.B. Cabrera, M.G. Hipólito, H.S. Murrieta, Effects of the  $\text{HfO}_2$  sinterization temperature on the erbium luminescence. *J. Lumin.* **145**, 713–716 (2014)
34. E. Montes, I. Martínez-Merlín, J.C. Guzmán-Olguín, J. Guzmán-Mendoza, I.R. Martín, M. García-Hipólito, C. Falcony, Effect of pH on the optical and structural properties of  $\text{HfO}_2:\text{Ln}^{3+}$ , synthesized by hydrothermal route. *J. Lumin.* **175**, 243–248 (2016)
35. S. Kumar, S.B. Rai, C. Rath, Monoclinic to cubic phase transformation and photoluminescence properties in  $\text{Hf}_{1-x}\text{Sm}_x\text{O}_2$  ( $x = 0-0.12$ ) nanoparticles. *J. Appl. Phys.* **123**, 055108 (2018)
36. G. Kresse, D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–17598 (1999)
37. J.P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996)
38. A. Sharma, J.P. Singh, S.O. Won, K.H. Chae, S.K. Sharma, S. Kumar, Introduction to X-Ray absorption spectroscopy and its applications in material science, in *Handbook of materials characterization*. ed. by S. Sharma (Springer, 2018), pp.497–548
39. J. Rodriguez-Carvaial “Fullprof: A Program for Rietveld Refinement and Pattern Matching Analysis,” *Abstract of the Satellite Meeting on Powder Diffraction of the XV Congress of the IUCr*. (Toulouse, France, 1990), pp. 127
40. V.R. Mastelaro, V. Briois, D.P.F. de Souza, C.L. Silva, Structural studies of a  $\text{ZrO}_2\text{--CeO}_2$  doped system. *J. Eur. Ceram. Soc.* **23**, 273–282 (2003)
41. C.H. Lu, J.M. Raitano, S. Khalid, L. Zhang, S.-W. Chan, Cubic phase stabilization in nanoparticles of hafnia-zirconia oxides: particle-size and annealing environment effects. *J. Appl. Phys.* **103**, 124303 (2008)
42. S. Chaudhary, D. Hsieh, G. Refael, Orbital Floquet engineering of exchange interactions in magnetic materials. *Phys. Rev. B* **100**, 220403 (2019)
43. D.Y. Cho, H.S. Jung, C.S. Hwang, Structural properties and electronic structure of  $\text{HfO}_2\text{--ZrO}_2$  composite films. *Phys. Rev. B* **82**, 094104 (2010)
44. J.I. Beltran, M.C. Muñoz, J. Hafner, Structural, electronic and magnetic properties of the surfaces of tetragonal and cubic  $\text{HfO}_2$ . *New J. Phys.* **10**, 063031 (2008)
45. N.F. Santos, J. Rodrigues, T. Holz, N. Ben Sedrine, A. Sena, A.J. Neves, F.M. Costa, T. Monteiro, Luminescence studies on  $\text{SnO}_2$  and  $\text{SnO}_2$ : Eu nanocrystals grown by laser assisted flow deposition. *Phys. Chem. Chem. Phys.* **17**, 13512–13519 (2015)
46. K. Smits, L. Grigorjeva, D. Millers, A. Sarakovskis, A. Opalinska, J.D. Fidelus, W. Lojkowski, Europium doped zirconia luminescence. *Opt. Mater.* **32**, 827–831 (2010)
47. M. Vikas, Jayasimhadri, Thermally stable red luminescence from  $\text{Eu}^{3+}$ -activated telluro zinc phosphate glass under near-ultraviolet light excitation for photonic applications. *Luminescence* **37**, 2059–2066 (2022)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

# MTSO: Multi-Target Search Optimisation based on Probability Map

Aman Virmani

Department of Electrical Engineering  
Delhi Technological University  
Delhi, India  
virmaniaman4@yahoo.com

Vayam Jain

Department of Applied Physics  
Delhi Technological University  
Delhi, India  
jainvayam@gmail.com

Nilesh Aggarwal

Department of Applied Physics  
Delhi Technological University  
Delhi, India  
nilesh.ggoo@gmail.com

Arjun Gupta

Department of Electronics Engineering  
Delhi Technological University  
Delhi, India  
arjun222gupta@gmail.com

Anunay

Department of Mechanical Engineering  
Delhi Technological University  
Delhi, India  
anunay2608@gmail.com

Dr. Anup Kumar Mandpura

Department of Electrical Engineering  
Delhi Technological University  
Delhi, India  
kanup@dtu.ac.in

**Abstract**—This report presents an optimized, decentralized way of searching for targets using swarms of rotary based unmanned aerial vehicles with information available from onboard EO/IR sensors. The are three main objectives of the proposed algorithm: time-optimized multi-target search, maximum and optimized payload drop, and maximum area coverage. The real-time application of these UAVs is in Human Aid and Disaster Relief scenarios mostly where the survivors have to be identified and given relief material. The proposed controller is inspired by the multi-target particle swarm optimized method(MTPSO) and its limitations of convergence towards Pbest and Gbest which may not be the best option in HADR scenarios. The algorithm involves the target probability distribution of the cells present in the target area which keeps on updating dynamically as the UAVs dive deeper into the target area. The deep learning-based model for target detection is deployed on each UAV using a dual camera with a limited field of view and its target discriminability varies as a function of the Class of the target, environmental conditions, etc. These parameters are given input to the probability distribution method for generating a probability map using which the UAVs optimize their path. The algorithm is evaluated on Ardupilot's SITL platform for parameter tuning and simulation in various scenarios which are later compared with existing search methods.

**Keywords**—Swarm, Unmanned Aerial Vehicles, Multi-target Search, Multi-agent systems, Particle Swarm Optimization, Probability map, k-means, Hungarian algorithm.

## I. INTRODUCTION

The demand for Unmanned Aerial vehicles has increased in the past years, as they are capable of rapidly covering areas, and providing better surveillance and efficient monitoring areas which gives them an edge over other alternatives. Because of these characteristics, the role of UAVs in civilian and military applications has become popular. The ground target search problem is one of the most important and popular applications of UAVs. Over time, the problem has become more complex as the number of targets and the search area has grown; to solve these huge UAV swarms are used. In this type of mission, it is essential to provide aid quickly and find multiple targets as quickly as possible.

There have been multiple solutions[1] put forward by various researchers for the problem of multi-target acquisition. The most common is a simple exhaustive search, i.e. scanning the complete area using pre-defined trajectories. Although used successfully, this method has certain limitations. First, as the trajectories are pre-fed, they fail to adapt according to the dynamic situations, resulting in more time taken. Second, as the decisions are not being made autonomously, there is a need for the operations team to look after the mission which is both expensive and time-consuming.

There are also unique approaches based on heuristic methods like multiple target search area optimization [2] which have proven to give good results during operations. The algorithm begins with an exhaustive search, identifies the global best and personal best during the search, and changes their behavior according to them. This algorithm satisfies all three main objectives but has its limitations like the guidance of UAVs away from multiple targets if the global best is identified somewhere else. Also, the percentage of the area covered can be improved if additional lines of UAVs are added.

There are also other centralized approaches similar to our algorithm which involves dividing the target area into small n-cells [3],[4] and assigning a probability to each cell. As the UAVs move further into the search area the probability of the cells keeps on updating depending on the targets found by them. This creates a dynamic updating probability map of the search area. This method adapts according to the situation but has its limitations. First, the centralized controller stores data from all agents and this requires a robust communication architecture.

In this report, we propose a decentralized swarm controller which is based on the probability method which has been proven to give good results but its application in multi-target searches has been limited due to a lack of adequate structure, multi-UAV collisions, and poor navigation in unfamiliar search locations. The proposed controller has a defined structure to optimize search and payload drops and incorporate inter-UAV collision avoidance. The algorithm is scalable, and fault-tolerant but is computationally expensive and hence requires a strong OnBoard Computer. In our case, we chose the NVIDIA Jetson Nano as the OBC as the requirement was to deploy a CPU-intensive algorithm for search and GPU intensive deep learning-based model for object detection. Also, for the communication architecture requirement, we integrated a software-defined Radio with mesh topology to test the controller in a real-world environment.

The limitation of the probability-based method was the need for a centralized controller so the probability map is the same for all the UAVs, but the current controller estimates velocity based on the weight assigned to the cells making the process dynamic and real-time hence eliminating the issue of uniformity. There are two main terms in the velocity vector term, first the inertial term which controls the factor exploration vs exploitation, and the second the derivative of the probability distribution curve. The weights of these were tuned visually such that there was no saturation in the local minima of the probability distribution curve. The exploration and exploitation characteristics are also analyzed over different functions which have given different performances and can be used for different missions. [5]-[11].

## II. PRELIMINARIES

### 2.1 Problem Formulation

We envision a swarm of quadcopter UAVs moving across a search region specified by a starting waypoint  $W_{start}$ , an ending waypoint  $W_{end}$ , and a breadth  $B$ , seeking numerous targets strewn throughout the landscape. Because all of the UAVs are at the same altitude, they all travel in the same direction. The position vector  $p_i^t = [x_i, y_i]^T$  and the velocity vector  $v_i^t = [x_i, y_i]^T$  where  $N$  is the number of UAVs, may be employed for the UAV  $i$  ( $i = 1, 2, \dots, N$ ) at time  $t$ .

Using an onboard software stack. Each UAV has an onboard camera that can identify and locate objects inside its field of vision (FOV). Each UAV is also equipped with a single payload that may be dropped at one of many different target locations. The ultimate purpose of UAVs is to scan the surveillance area, find as many targets as possible, drop as many payloads as possible, and reach the end of the search area while avoiding inter-UAV collisions.

A partially connected ad-hoc network [12] is used to mimic the network architecture of all the agents. The model can be represented mathematically by an undirected graph  $G = (V, E)$ , where  $V = 1, 2, \dots, N$  is the set of nodes and  $E = \{(i, j) : i, j \in V, \|p_i - p_j\| \leq DC\}$  is the edge set. A single UAV directly connects only to its neighboring agents denoted by  $NG = \{j \in V, (i, j) \in E\}$

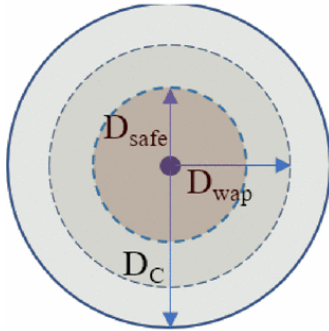


Fig.1: Pictorial representation of  $D_{safe}$ ,  $D_{wap}$ , and  $D_C$

Agents only share the bare minimum of data, such as their location data for inter UAV-collision avoidance, the GPS location and probability of detected targets for cooperative convergence on target clusters, and their payload status for smart and cooperative payload delivery a pictorial representation of the radii are given in fig.1..

The following bounding requirements (in order of priority) confine the algorithm:

- Avoid colliding with other UAVs.
- Identify every target in the search area
- Drop as many cargoes as feasible as fast as possible.

### 2.2 Probability Map Formulation

Initially the search area is divided into  $N$ -cells, where the dimension of each cell is chosen on the basis of the computational capability of the Onboard Computer of the UAV. In our case, the NVIDIA Jetson nano can support up to 30,000-40,000 grid cells, hence according to this limitation the cell dimension is decided.

An object detection algorithm has been used to execute aerial detection of the objects using the images from the EO/IR camera. After deploying and testing the leading models for object detection, YOLOV5 was chosen for its high accuracy while also maintaining decent FPS when deployed on the Jetson Nano. You Only Look Once (YOLO), takes in the images from the EO/IR

sensor and predicts the classes and the bounding boxes along with each detection's confidence value. The object is subsequently geotagged using the field of view of the EO/IR sensor.

The probability and area of influence are the two inputs that are estimated for the construction of a probability map. The probability of a target being at the specific cell can be specified by the confidence level that the YOLOv5 model returns and has a range of (0.7,1). The area of influence on the other hand has various factors which need to be accounted for and have been discussed below briefly.

- **Class of the Object:-** There are multiple types of targets that we took in our test case e.g. humans, fuel dumps, type-A vehicles, and type-B vehicles. Certain classes such as Type A vehicles are physically larger than other objects such as Humans and easily detected. Moreover, they should have a much higher area of influence than humans as the presence of vehicles implies a greater concentration of possible targets.
- **Altitude and Size of the object:-** For a given class the relationship between the altitude of the UAV and the size of the object in the image plane ( in pixels ) is directly proportional and can be calculated using the field of view of the camera system and the resolution of the camera using basic trigonometry.
- **Cluster of targets:-** In real-world environments targets tend to stick together and travel in clusters, hence the probability of a specific grid cell needs to be determined increased if the other targets are also dedicated in its vicinity. This can be represented mathematically just by adding the probability over the cells again.

## III. ALGORITHM DESIGN

### 3.1 Initialization of UAVs

We begin the target-search method in a thorough manner, with UAVs placed in a line at the start of the search region, to optimize the initial exploration. The overlap between each UAV's FOV is limited to a bare minimum and remains consistent ( $O_{min}$ ). For repeated runs, if the search region is large enough, it can be divided into many portions.

### 3.2 Inertial Term

The inertial velocity term, as its name implies, aids in maintaining the swarm unit's original course and prevents abrupt changes in each UAV's trajectory. It's the result of multiplying the current velocity by the inertial function. The algorithm's exploration vs. exploitation features is controlled by the inertial function, which is generally calibrated to give better exploration at first and exploitation as the swarm reaches deeper into the area and has already collected sufficient information about the area.

### 3.3 Probability Calculation

The concept is based on the fact that targets stick together and travel in clusters, hence a probability distribution is proposed. The mathematical model used to estimate the probability distribution is the Gaussian / Normal distribution curve which is represented below:-

$$f(x) = e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2} \div \sigma \sqrt{2\pi}$$

where,  
 $\sigma$  = Standard Deviation

$\mu$  = Mean

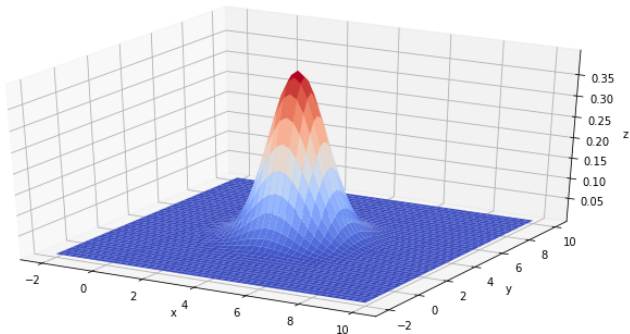


Fig.2: Graphical representation of gaussian curve

$\sigma$  represents the area of influence which as specified in the previous section has to be selected according to the class of targets identified during the search. The below table represents the area of influence for each class:-

S.No	Type of target	Range of Influence
1	Human	20-40 meters
2	Fuel Dumps	40-60 meters
3	Type-B vehicles	60-100 meters
4	Type-A vehicle	100-140 meters

Table 1: Limits of  $\sigma$  for the type of class

The reason for taking the limits was the second factor which corresponds to increase in probability if the number of pixels occupied by that class of target is more. For YOLOV5 there is a relationship between the number of pixels occupied by the target and the accuracy of the detection given as:

$$m(x) = \frac{1}{1 + e^{-s(x-m)}}$$

Where:-

$f(x)$  = Accuracy of the detection

$x$  = Size of the object

$s$  = Steepness

$m$  = Shift

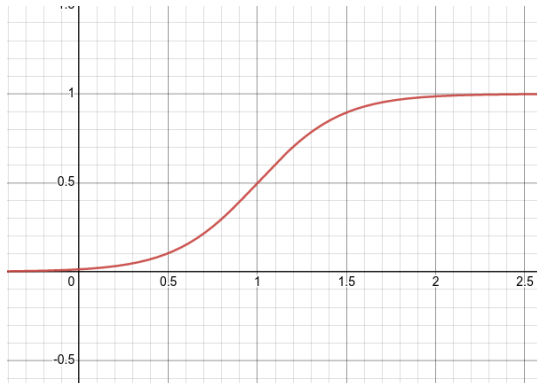


Fig.3: The S-shaped curve of  $m(x)$

The graph for the above function  $m(x)$  is shown in fig.3.

The YOLOV5 model was tested and the value of the hyperparameters came out be:

$S = -0.1$

$M = 16$

These parameters represent that the YOLOV5 model can only detect targets with good accuracy if they occupy more than 16 pixels in the input image. Now, this behavior of the graph is imposed between the limits to increase the  $\sigma$  if the number of pixels of the same target class increases, hence making the algorithm more operationally robust.

$$\sigma = m(x) * (Upper\ Limit - Lower\ Limit) + Lower\ Limit$$

This process is repeated for each detected object to generate an independent probability map for each target and the corresponding probability maps are added up to find the final probability map  $P(x)$  countering the third point that probability needs to be increased further if more targets keep on detecting in high probable areas. This map can be visualized as a surface plot where the x and y-axis represent the search grid and the z-axis represents the probability value at that specific grid point. The map gives a representation like shown in fig.4.

$$P(x) = \sum_{i=0}^n f_i(x)$$

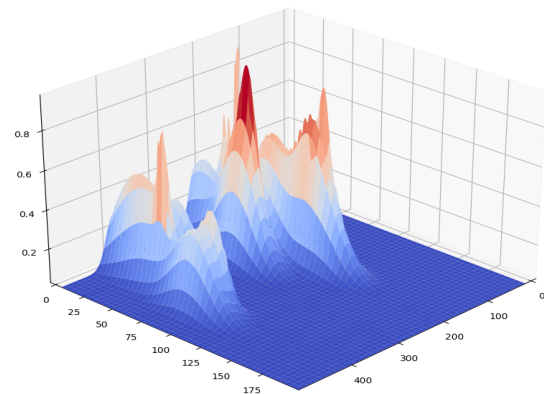


Fig.4: Gaussian-based probability distribution curve of the target area

### 3.4 Probability-Based Velocity

The probability-based velocity term aims to guide the UAVs towards the high probability areas using the probability map generated earlier. The partial derivative/gradient of the probability map is taken to get the following surface plots representing the partial derivative in the x and y-axis respectively as shown in fig.5.

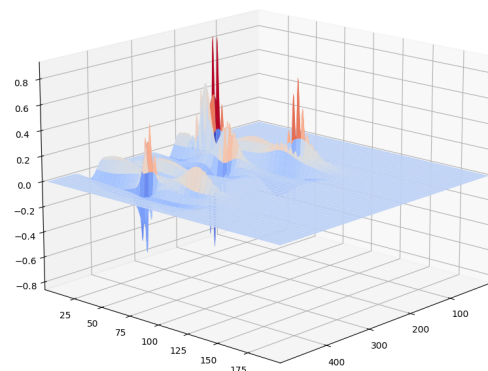


Fig.5: Partial derivative of  $P(x)$  in the x and y-axis respectively

$$D(x) = \frac{dP(\sum_{i=0}^n f_i(x))}{dx}, D(y) = \frac{dP(\sum_{i=0}^n f_i(y))}{dy}$$

The probability-based velocity term uses this gradient to guide it towards the high probability areas using the formula:-

$$V_{(x,y)} = w * D(x, y)$$

where,

w = Tunable constant

### 3.5 Payload Optimization

The most crucial duty in the quick reaction approach is the payload drop. For job assignment and optimal delivery, we begin by assigning the UAVs to one of three states, which aids in determining which UAVs should continue to search and which should proceed for payload drop:

State	Meaning
State '1'	Payload available
State '0'	Going to deliver the payload
State '-1'	Payload has been dropped

Table 2: Payload drop states

As long as the target is within the maximum displacement range, the locations of the identified targets are continuously relayed between the UAVs, and the payload delivery job is assigned to the UAV nearest to the target.

$$v_{drop}^{t+1} = C_3 * (p_{drop} - p^t)$$

Furthermore, if the UAV detects a target B that is closer to target A while dropping a payload on target A, it moves its drop location to target B, re-qualifying target B for payload drop. While delivering the payload, the UAV continues to seek targets and communicate with other UAVs. It does not return its steps after delivering the goods, instead of continuing the hunt from the drop spot.

### 3.6 Inter-UAV Collision Avoidance

To maintain the safety of any swarm system, inter-UAV collision avoidance is essential. As a result, we use the consensus equation to generate an extra velocity term that may be vectorially added to each agent's final velocity [13]. If the distance between any nodes gets lower than  $D_{safe}$ , avoidance is initiated. This method is simple to include in our framework and is represented as follows:

$$D = \|p_i - p^t\|_2$$

$$v_c^{t+1} = \begin{cases} C_4 * \sum_{i=1}^N (p_i - p^t) * \left(1 - \frac{D_{safe}}{D}\right), & D \leq D_{safe} \\ 0, & D > D_{safe} \end{cases}$$

This method is suited for our application since it just considers the agents coming for collision. The  $D_{safe}$  parameter is usually set to match the maximum swarm velocity, although it can be manually adjusted if necessary.

### 3.7 Net Equivalent Velocity

The UAVs communicate with each other to share information about the targets and allow for the decentralized construction of probability maps, such that the final velocity vector can be calculated onboard each UAV independently. Hence the final velocity vector is as follows:

$$V = v_{Inertial} + v_{Probability} + v_{Obstacle Avoidance} + v_{Payload Drop}$$

### 3.8 Return Mission

As the mission in MTSO progressed a lot of empty unsearched gaps in the search area were left behind to optimize the time of the mission. Thus when the MTSO mission is over another grid search mission is plotted on the unsearched regions so as to be sure that no other targets are left behind.

#### 3.8.1 K-Means Clustering Algorithm

In the return mission the first task is to calculate the total unsearched area left behind during the search as shown in fig.6. Once this is calculated all the points in the area are segregated in clusters using the K-means clustering algorithm. The K-Means Clustering technique divides the input dataset into multiple clusters based on their distances from the centroids. It's a clustering-based approach in which each cluster has its own centroid, which is used to characterize the clusters. The input given to the Kmeans is the coordinates of the unsearched area which are represented as d-dimensional real vectors as  $(x_1, x_2, \dots, x_n)$ . v K means to organize the n groups to k groups where  $k \leq n$ . The main equation of k means is

$$\arg \min_S \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|^2 = \arg \min_S \sum_{i=1}^k |S_i| \text{Var } S_i$$

where  $\mu_i$  is the mean of points in  $S_i$ . This is equivalent to minimizing the pairwise squared deviations of points in the same cluster

$$\arg \min_S \sum_{i=1}^k \frac{1}{|S_i|} \sum_{x, y \in S_i} \|x - y\|^2$$

Which can be reduced to

$$S_i \sum_{x \in S_i} \|x - \mu_i\|^2 = \sum_{x \neq y \in S_i} \|x - y\|^2$$

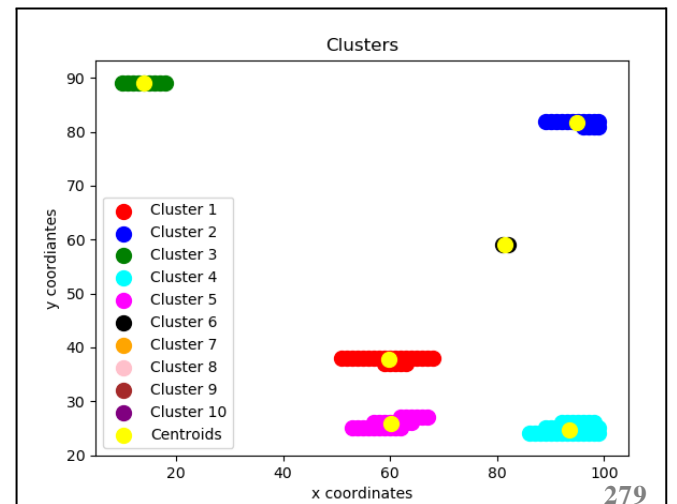


Fig.6: Image showing the division of clusters and centroids using the K-means algorithm



Because the total variance is constant, this is identical to iteratively maximizing the sum of squared deviations across points in distinct clusters.

The number of clusters in which the unsearched area is to be divided is decided by the elbow function. The elbow method runs k-means clustering on the dataset for a range of values for k and then for each value of k computes an average score for all clusters. Using these scores a graph is plotted as shown in fig.7 and the line with the highest changes in slope indicates the optimal number of clusters that need to be made.

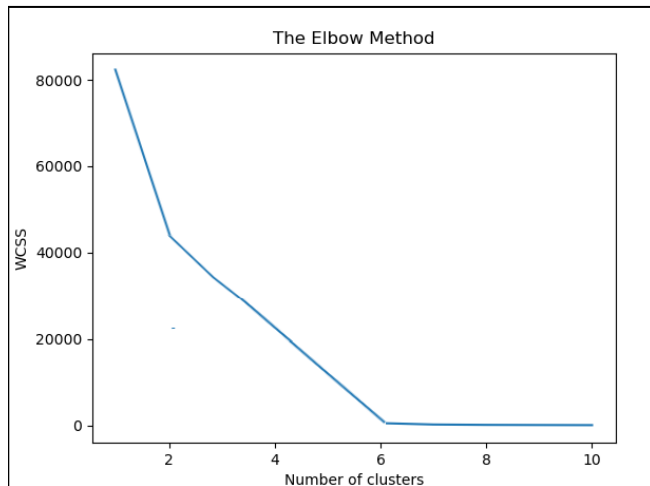


Fig.7: Elbow method showing the max slope change at 6 clusters

### 3.8.2 Health Function

Once the clusters are divided, UAVs are also segregated based on the priority order given below -:

- Payload dropped or not-: As the number of targets and the number of UAVs are not always equal, so in a case if the number of targets is less than the number of UAVs the UAVs that haven't dropped their payload yet are listed first for the return mission.
- Health function -: The rest of the UAVs with the highest battery percentage and proximity to targets are chosen first for the return mission.

### 3.8.3 Hungarian Algorithm

The clusters have been constructed, and the UAVs that are qualified for the mission have been separated; now each cluster must be assigned to a UAV that will conduct the search. The nearest possible cluster should be assigned to each UAV for mission robustness and time optimization, and the Hungarian method was utilized for this. The Hungarian Method is a technique based on the idea that if the same value is added or subtracted from every member of a matrix's row or column, the new assignment issue's optimal solution should be the same as the original problem.

### 3.8.4 Polygon Grid Search

The Hungarian algorithm provided each UAV its cluster. Now each UAV is to be assigned waypoints to search in its area. To assign the waypoints a self-implemented polygon grid search was invented that can work on any polygon. Our algorithm takes in the coordinate points of the boundary of the polygon and forms a perfect rectangle, around the polygon completely enclosing the polygon inside it. Now the Grid search method is applied on the rectangle providing initial and final waypoints to the UAV as it

continues the mission. To overcome the extra space provided by the rectangle that is not in the polygon a line intersect function is implemented to keep the UAV inside the polygon.

## IV. SIMULATION AND TESTING

This section firstly describes the methods of simulations used to test our code. Then, a comparative study of the tuning of different inertial functions according to different FOVs is given. Lastly, the effectiveness of MTSO is compared with other methods in various simulated scenarios.

### A. Simulation Environment

To simulate the algorithm in various scenarios and later on compare it to other algorithms, Ardupilot SITL was utilized. To visualize the targets, a self-created mavproxy module is used, which selects the targets at random and visualizes them in the SITL map as shown in fig. 8-11. If the target is detected, the probability map is updated and presented continuously using OpenCV and matplotlib libraries. Because the SITL only accepts GPS data, conversion to UTM coordinates were conducted.

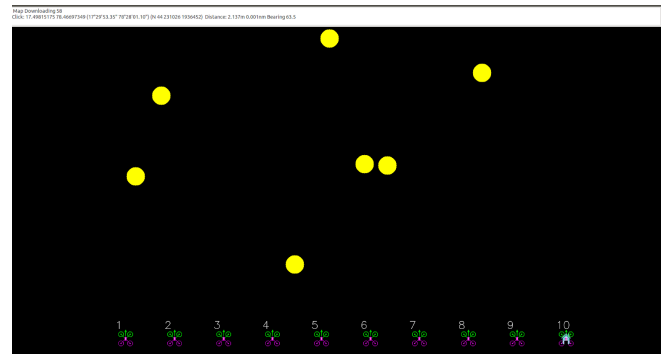


Fig.8: Simulation Environment and Initialization of UAVs

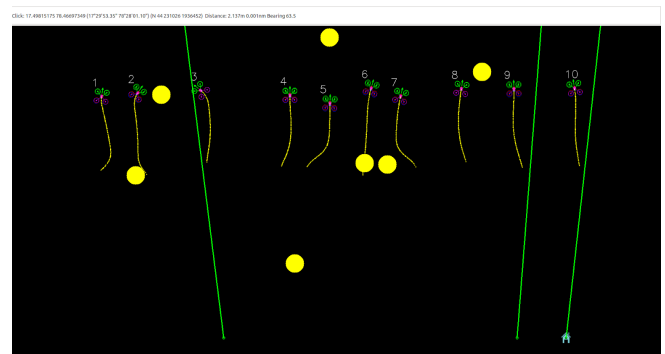


Fig.9: Convergence of UAVs on targets in search area

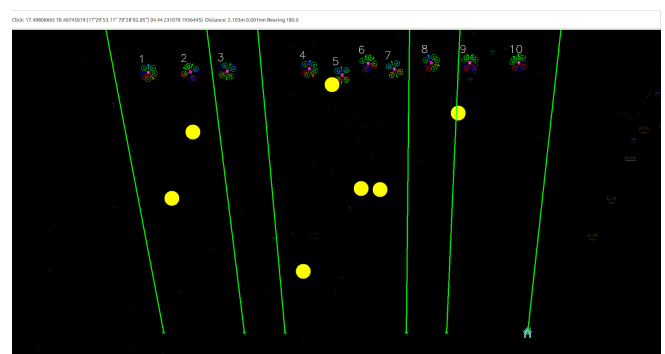


Fig.10: UAVs halting and performing clustering



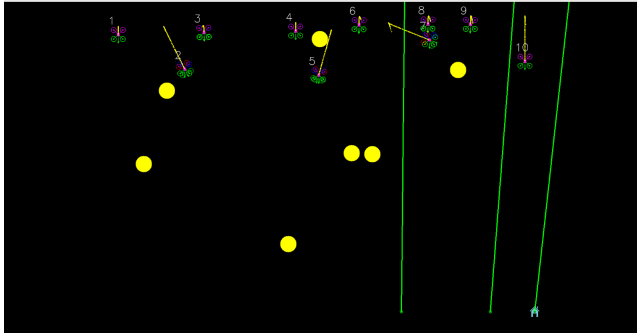


Fig.11: UAVs searching their assigned cluster according to the Hungarian algorithm

### B. Path Map of the Swarm

In order to evaluate the mission later, a real-time map of each UAVs travel trajectory is created at the start of the flight. This map also shows the target that was discovered. The map is created using a straightforward openCV function and is shown with a real-time mission updates shown in fig.12.

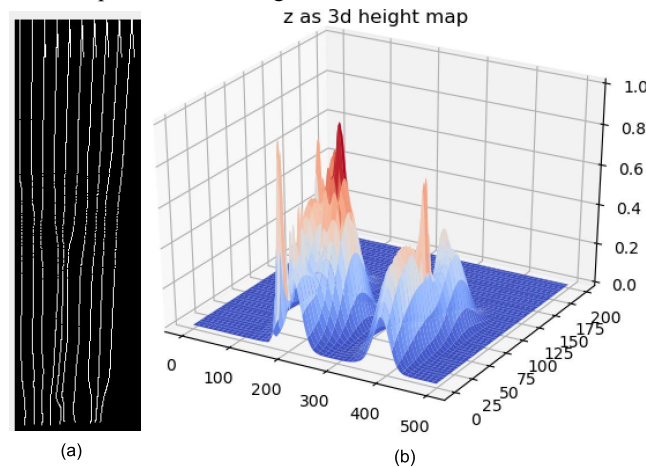


Fig.12: (a)Path map of the swarm and (b) Probability Map for the test case of FOV 140x80

### C. Quantitative Comparison with Different FOVs

The algorithm of MTSO with probability map was rigorously tested against different parameters for FOV of the camera to find its performance and discrepancies if any. These tests were carried out while the camera module's FOV was changed. The FOV settings used in the experiment were chosen with the least amount of overlap in mind (0 min).

Table III summarizes the outcomes of the tests. The data clearly illustrates that expanding the FOV enhances exploration since the majority of the search region is searched, resulting in an improved overall number of discovered targets but also increasing the total time needed to scan the environment. So it is essential to find an optimum balance of efficiency, time optimization and total area scanned to get the best results for the mission.

Metrics	FOV (in meters)		
	55X32	70 X40	140X80
Target Detected	15	17	20
Payload Dropped	14	17	19

Time For Mission(s)	188.14	182.24	176.89
Total Area Scanned	92.46	96.23	98.91

Table 3: Comparison between different FOVs

### D. Quantitative Comparison with other approaches

This part shows the comparison between MTPSO, Exhaustive Search and our approach. Below in table.4 is a summary of the results of different qualities of the search algorithm by changing fov in different scenarios by all three algorithms.

Metrics	Method	FOV		
		55X32	70X40	140X80
Targets Detected	Exhaustive Search	19	20	20
	MTPSO	15	16	19
	New Approach	15	17	20
Payload Dropped	Exhaustive Search	18	18	19
	MTPSO	15	15	17
	New Approach	14	17	19
Time For Mission	Exhaustive Search	402.63	422.75	454.56
	MTPSO	207.86	200.01	193.45
	New Approach	188.14	182.24	176.89
Total Area Scanned	Exhaustive Search	99.99	99.99	99.99
	MTPSO	92.61	96.38	98.32
	New Approach	92.46	96.23	98.91

Table 4: Performance metrics evaluated for different approaches with different FOVs

In each of the scenarios outlined above, MTSO with probability map outperformed exhaustive search and MTPSO. The time used by MTSO was roughly half that of the exhaustive search in terms of time optimization. Because the probability map aids in swarm convergence rather than selecting a Pbest and Gbest, the convergence towards the goal is finer, resulting in a smaller scanned region. The processing power required by MTSO utilizing a probability map is significantly lower than that required by any other probability graph approach, and it is easily implemented in a decentralized version on very light hardware platforms with minimum connectivity.

Hyper parameter	Value	Hyper parameter	Value
$D_{\text{wap}}$	20m	$v_{\text{min}}$	0m/s
$D_c$	500m	N	20
$D_{\text{max}}$	2.5*FOV	$D_{\text{safe}}$	15m
$N_s$	5	w	30
$\alpha$	105°	$\mu$	0
$v_{\text{max}}$	15m/s	m	16

**Table 5: Hyperparameters used for simulations****V. CONCLUSION**

The new algorithm, unlike the earlier MTPSO algorithm, does not limit the movement of UAVs on the basis of only PBEST and GBEST, which often led to UAVs getting stuck in local minima and being unable to contribute for the rest of the mission. The new approach allows the movement of the UAVs on the basis of all the detected objects using a probability map, hence areas with a much higher concentration of objects take precedence. Moreover, the impact of each object is not constant unlike the previous approaches and various factors like Altitude, Class and Model Confidence which actually affect the ideal behavior in real-world environments are taken into account for assigning the probability which leads to better convergence of the swarm. This also allows for considerably more payloads to be dropped quickly, when compared to exhaustive search and MTPSO increasing the effectiveness of the UAV swarms in real-world disaster relief scenarios. To overcome the limitation of the new algorithm of limited coverage of the search area a return mission is proposed which can be used when the endurance of the UAVs allows, to exhaustively search the left out areas and confirm any objects missed by the initial pass of the algorithm.

The various hyperparameters used in the new algorithm have been tuned manually but instead, they can be tuned using advanced techniques like evolutionary hyperparameter tuning which can further improve the performance of the algorithm.

In future work, the research aims to improve:-

- Refining the revert mission to do an iterative MTSO.
- Further improving MTSO to decrease time and increase efficiency.

**REFERENCES**

- [1] Cabreira, T.M.; Brisolara, L.B.; Ferreira Jr., P.R. Survey on Coverage Path Planning with Unmanned Aerial Vehicles. *Drones* 2019, 3, 4.
- [2] L. Bertuccelli and J. How, "Robust UAV search for environments with imprecise probability maps," in *Proc. IEEE Conf. Decision Control*, Dec. 2005, pp. 5680–5685.
- [3] Y. Yang, A. Minai, and M. Polycarpou, "Decentralized cooperative search by networked UAVs in an uncertain environment," in *Proc. Amer. Control Conf.*, Jun.–Jul. 2004.
- [4] J. Hu, L. Xie, K. Lum and J. Xu, "Multiagent Information Fusion and Cooperative Control in Target Search," in *IEEE Transactions on Control Systems Technology*, vol. 21, no. 4, pp. 1223-1235, July 2013.
- [5] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of IEEE International Conference on Neural Networks*, pages 1942-1948, IEEE Press, Piscataway, NJ, 1995.
- [6] Y. Shi and R. Eberhart., "A modified particle swarm optimizer", In *Evolutionary Computation Proceedings*, 1998. *IEEE World Congress on Computational Intelligence.*, The 1998 IEEE International Conference on, pages 69–73.
- [7] IEEE, 2002. R.C. Eberhart and Y. Shi., "Tracking and optimizing dynamic systems with particle swarms", In *Evolutionary Computation*, 2001. *Proceedings of the 2001 Congress on*, volume 1, pages 94–100.
- [8] IEEE, 2002. A.Nikabadi, M.Ebadzadeh , "Particle swarm optimization algorithms with adaptive Inertia Weight : A survey of the state of the art and a Novel method", *IEEE Journal of evolutionary computation* , 2008
- [9] R.F. Malik, T.A. Rahman, S.Z.M. Hashim, and R. Ngah, "New Particle Swarm Optimizer with Sigmoid Increasing Inertia Weight", *International Journal of Computer Science and Security (IJCSS)*, 1(2):35, 2007.
- [10] J. Xin, G. Chen, and Y. Hai., "A Particle Swarm Optimizer with Multistage Linearly-Decreasing Inertia Weight", In *Computational Sciences and Optimization*, 2009. *CSO 2009. International Joint Conference on*, volume 1, pages 505–508. IEEE, 2009.
- [11] Y. Feng, G.F. Teng, A.X. Wang, and Y.M. Yao., "Chaotic Inertia Weight in Particle Swarm Optimization", In *Innovative Computing, Information and Control*, 2007. *ICICIC'07*.
- [12] I. Bekmezci, O. K. Sahingoz, and Ş. Temel, "Flying AdHoc Networks (FANETs): A survey," *Ad Hoc Networks*, vol. 11.
- [13] Anuj Agrawal, Aniket Gupta, Joyraj Bhowmick, Anurag Singh, Raghava Nallanthighal, "A Novel Controller of Multi-Agent System Navigation and Obstacle Avoidance", *Procedia Computer Science*, Volume 171, 2020, Pages 1221-1230
- [14] A. Gupta, A. Virmani, P. Mahajan and R. Nallanthighal, "A Particle Swarm Optimization-Based Cooperation Method for Multiple-Target Search by Swarm UAVs in Unknown Environments," 2021 7th International Conference on Automation, Robotics and Applications (ICARA), 2021, pp. 95-100, doi: 10.1109/ICARA51699.2021.9376529.
- [15] N. Aggarwal, Anunay, V. Jain, T. Singh and D. K. Vishwakarma, "DLVS: Time Series Architecture for Image-Based Visual Servoing," 2023 8th International Conference on Control and Robotics Engineering (ICRE), Niigata, Japan, 2023, pp. 101-107, doi: 10.1109/ICRE57112.2023.10155585.
- [16] Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. *Proceedings of IEEE International Conference on Neural Networks*, 4, 1942-1948. DOI: 10.1109/ICNN.1995.488968.
- [17] Coello Coello, C. A., Lamont, G. B., & Van Veldhuizen, D. A. (2007). *Evolutionary algorithms for solving multi-objective problems* (2nd ed.). Springer.
- [18] Tan, Y., Zhu, J., & Ding, X. (2013). A hybrid multi-objective particle swarm optimization algorithm. *Soft Computing*, 17(6), 1013-1031. DOI: 10.1007/s00500-013-0983-3.
- [19] Zeng, J., Chen, J., Zhang, Q., & Zhang, B. (2018). Multi-objective particle swarm optimization with decomposition for optimization problems with variable linkages. *IEEE Transactions on Evolutionary Computation*, 23(2), 217-231. DOI: 10.1109/TEVC.2018.2817386.
- [20] López-Ibáñez, M., Paquete, L., & Stützle, T. (2010). Hybridization of evolutionary multiobjective algorithms: A survey. *ACM Computing Surveys (CSUR)*, 42(4), 1-37. DOI: 10.1145/1830761.1830763.
- [21] N.Roberto, Z.Qian, L.Yanmin H.Huayao, Y.Meilan, S.Xiaoli, 02/03/2022 "Multi-Objective Particle Swarm Optimization with Multi-Archiving Strategy", 1058-9244, doi:10.1155/2022/7372450.

- [22] Raida SELLAMI, Farooq SHER, Rafik NEJI, An improved MOPSO algorithm for optimal sizing & placement of distributed generation: A case study of the Tunisian offshore distribution network (ASHTART), Energy Reports, Volume 8, 2022, Pages 6960-6975, ISSN 2352-4847, <https://doi.org/10.1016/j.egy.2022.05.049>.
- [23] D.Van, W.Shihua, L.Yanmin, Z.Kangge, L.Nana, W.Yaowei, 2022, "Multiobjective Particle Swarm Optimization Based on Ideal Distance", 2022, DOI:10.1155/2022/3515566
- [24] Yang M, Liu Y, Yang J. A Hybrid Multi-Objective Particle Swarm Optimization with Central Control Strategy. Comput Intell Neurosci. 2022 Mar 9;2022:1522096. doi: 10.1155/2022/1522096. PMID: 35310587; PMCID: PMC8926491.

# Multimodal Sarcasm Recognition by Fusing Textual, Visual and Acoustic content via Multi-Headed Attention for Video Dataset

Sajal Aggarwal

Biometric Research Laboratory,  
Department of Information Technology,  
Delhi Technological University  
Delhi, India  
authorsajal@gmail.com

Ananya Pandey

Biometric Research Laboratory,  
Department of Information Technology,  
Delhi Technological University  
Delhi, India  
ananyaphdit08@gmail.com

Dinesh Kumar Vishwakarma

Biometric Research Laboratory,  
Department of Information Technology,  
Delhi Technological University  
Delhi, India  
dvishwakarma@gmail.com

**Abstract**— Multimodal sarcasm recognition uses a combination of acoustic, video, and text-based cues to detect sarcasm. Though multimodal approaches have outperformed unimodal detection by providing a more comprehensive description of the speaker's sentiment, they are particularly challenging as cues may not be consistent across the multiple modalities. In our research study, we propose a system that first extracts multimodal features from the input provided and then applies a bimodal multi-head attention mechanism to them. Subsequently, the features are concatenated and passed through a softmax layer for detection. The proposed model is evaluated on the MUSTARD dataset for multimodal sarcasm recognition. For the speaker-dependent configuration, the proposed model beats cutting-edge methods in terms of accuracy, precision, recall, and F1-score by 0.75%, 2.9, 2.82, and 3.1, respectively.

**Keywords**—Sarcasm, Multimodal, MUSTARD dataset, Multi-headed attention, Multimodal Sarcasm Recognition

## I. INTRODUCTION

Sarcasm is a specific kind of irony that is used to ridicule or convey contemptuous remarks towards a person or situation. The literal interpretation of a sarcastic remark does not convey its intended meaning; rather, it is quite often the opposite. This makes sarcasm detection a somewhat arduous task. For example, "I really appreciate how you always show up to meetings unprepared. It really adds to the overall productivity of the group.", is a remark wherein a word indicative of a positive sentiment, i.e., 'appreciate' has been used to mock the listener for 'showing up to meetings unprepared'. A combination of verbal and nonverbal cues, including spoken words, tone of voice, facial expressions, and body language, can be used to convey sarcasm.

For a long time, researchers have studied sarcasm recognition on unimodal data (textual inputs), such as news headlines, tweets [1]–[3], or Reddit comments. However, textual input alone may not suffice in situations where additional contextual knowledge is required to identify sarcasm. Consequently, the emergence of multimodal datasets for sarcasm recognition ensued, encompassing a fusion of textual, acoustic, and video data. Multimodal inputs for sarcasm recognition in the form of videos and associated captions provide a more holistic representation of the emotion of the speaker in the conversational context as compared to just text or audio and thus aid more accurate sarcasm recognition. Figure 1 clearly demonstrates how non-verbal cues supplement the textual data and help us understand that a particular utterance is sarcastic.

Initially, statistical or rule-based techniques were adopted for text-based sarcasm recognition, like the method proposed in [3] for a set of manually annotated tweets. This was followed by the use of feature-dependent ML classifiers such as SVM, KNN, decision trees, random forests, and regression. Newer research suggests using deep learning techniques to automatically extract sarcasm-specific traits. The integration of attention mechanisms with deep-learning approaches has further augmented the accuracy of sarcasm recognition methods [4].

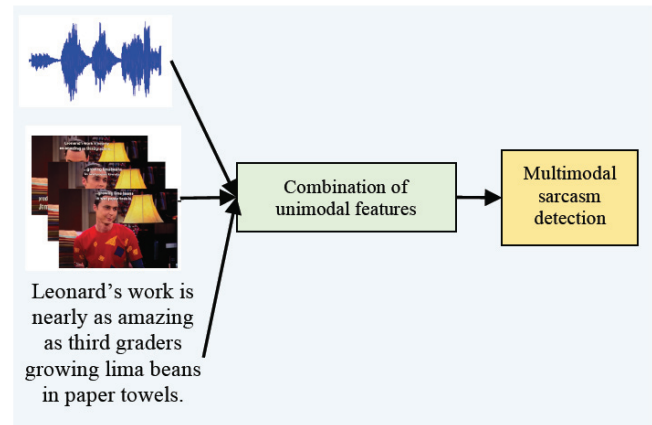


Fig. 1. Multimodal sarcasm recognition

The significant contributions of our research study are as follows:

- We present a unique framework for the recognition of sarcasm using multiple modalities of video utterances by applying a multi-head bimodal attention mechanism to pairs of text, video; text, acoustic, and acoustic, video characteristics. This allows us to recognise sarcasm in video utterances using multiple modalities.
- To test the robustness of our proposed architecture, an experimental analysis has been conducted on one of the benchmark datasets, MUSTARD [5].

In the following sections, we elaborate on our objective and methodology and provide additional information. The remaining sections of this paper are organised as follows: Section 2 focuses on the existing research conducted in the field of sarcasm recognition, with subsections discussing the various data modalities and data employed for the task. Section 3 delineates the proposed framework. Section 4

presents the results derived from the application of the proposed framework. Section 5 concludes our study and makes recommendations for future studies.

## II. RELATED WORK

### A. Unimodal Sarcasm Recognition

#### 1) Using text

In [1], Davidov et al. proposed a semi-supervised sarcasm identification algorithm for sarcasm recognition on two textual datasets, the first comprising 5.9 million tweets collected from Twitter and the second, 66,000 product reviews from Amazon. González-Ibáñez et al. [2] analyzed the effect of employing lexical and pragmatic features on the efficiency of machine learning methods for automatic sarcasm recognition on a dataset comprising 900 tweets.

#### 2) Using acoustic data

In [6], Tepperman et al. experimented with sarcasm recognition on 131 uninterrupted occurrences of the phrase ‘yeah right’ collected from the Switchboard and Fisher corpora using prosodic, spectral, and contextual cues.

### B. Multimodal Sarcasm Recognition

#### 1) Using image-text pairs

Multimodal datasets comprising image-text pairs have been used for sentiment analysis, such as by Pandey et al. in [7], [8], using attention mechanisms integrated with convolutional neural networks. Similar approaches are also being studied for sarcasm detection, which have been discussed further in this section.

Schifanella et al. [9] performed sarcasm recognition on image-text data collected from posts on three social media websites – Twitter, Tumblr, and Instagram, and demonstrated how images in a post help a user understand the situational context and decipher a sarcastic tone. The authors suggested two novel multimodal fusion frameworks to integrate text and visual features and concluded that the combination of the two modalities helped achieve better results for sarcasm recognition.

In [10], Sangwan et al. compiled two multimodal sarcasm detection datasets: the Silver-Standard dataset, comprising 10K sarcastic and non-sarcastic posts each, classified on the basis of hashtags, and the Gold-Standard dataset, comprising 1600 randomly selected sarcastic posts from the first dataset and annotated first using only the text modality and then reannotated using both modalities. The authors came up with an RNN-based deep learning system to take advantage of the fact that text and images are interdependent in order to recognize sarcasm.

Pan et al. [11] propose a BERT architecture-based model for sarcasm recognition on the image-text Twitter dataset introduced in [12]. The authors engaged inter-modal and intra-modal attention between images and text and within the text, respectively. In [13], Liang et al. proposed a novel cross-modal graph convolutional network (CMGCN) architecture for multimodal sarcasm recognition on the same dataset. Instead of considering the image as a whole, the authors suggested employing object recognition methods to recognize the important regions in the image and then

correlating the textual tokens with these regions for sarcasm recognition.

#### 2) Using videos

Chauhan et al. [14] presented a multi-task framework for multimodal sarcasm, emotion, and sentiment recognition. After extending the MUSTARD dataset to include appropriate emotion and sentiment labels, they utilized both emotion and sentiment for sarcasm recognition. Two novel attention mechanisms,  $I_e$  and  $I_a$  attention, provided a better combination of data across different modalities.

In [15], Pramanick et al. proposed the MuLOT system that combines the use of self and cross-attention modules for multimodal sarcasm recognition on the MUSTARD dataset [5]. While the purpose of self-attention was to study intra-modal relationships and highlight the most important features from each of the unimodal feature sequences, the pairwise mapping of unimodal features using cross attention provided the ability to study inter-modal relationships.

Bavkar et al. [16] suggested a model for multimodal sarcasm recognition, wherein after preprocessing the unimodal data, and extracting features from textual data using improved bag of words, TF-IDF, emojis, and n-grams from video data using CLM & SLBT and from acoustic data using MFCC, chroma, spectral features, and jitter, an improved feature fusion technique was used to concatenate the features. The final classification was carried out using a hybrid classifier of LSTM and Bi-GRU. The authors concluded that the combination of a hybrid classifier with the newly proposed OLAO algorithm used for tuning the weights of LSTM produced better detection results as compared to cutting edge methods.

Wu et al. [17] proposed the IWAN model, which uses word and utterance-level multimodal features for sarcasm recognition. Textual features were extracted using BERT and visual features using a pretrained ResNet-50 model after recognition of the speaker’s face from all the faces detected. Low-level speech features were extracted using OpenSmile. The authors used a scoring mechanism to measure the inconsistency between the sentiments of the words and non-verbal cues and assigned higher weights to words with contradicting modalities for sarcasm recognition.

In [18], Ray et al. extended the variant of the MUSTARD dataset with sentiment annotations provided by Chauhan et al. [14] to twice its size by adding new utterances from similar sources while maintaining the class balance. They incorporated labels indicating arousal, valence, and the type of sarcasm corresponding to each utterance. The authors released this newly formulated dataset as the MUSTARD++ dataset. Various configurations of pretrained feature extractors and multimodal fusion algorithms were used to benchmark the dataset.

Chauhan et al. [19] proposed another variant of the MUSTARD dataset, SEEmoji MUSTARD, wherein each utterance was annotated with a relevant emoji and the emoji’s emotion and sentiment. The authors suggested a deep learning-based emoji-aware multitask learning framework with sarcasm recognition as the primary task and emotion and sentiment analysis as the secondary.



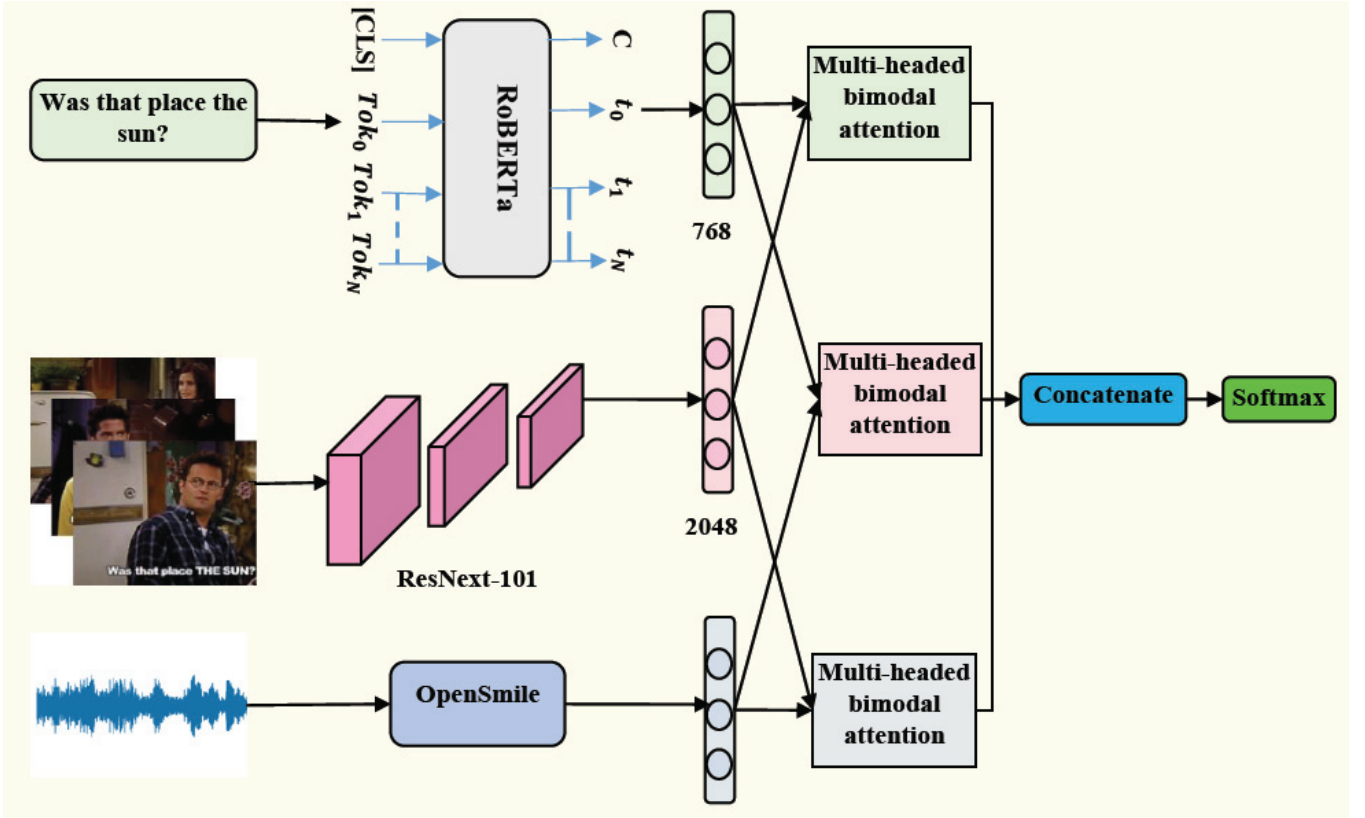


Fig. 2. Proposed Framework

### III. PROPOSED FRAMEWORK

This section outlines the method proposed for multimodal sarcasm detection on the MUSTARD dataset, as depicted in Figure 2.

#### A. Multimodal Feature Extraction

**Text Features:** The pre-trained RoBERTa [20] model with 12 hidden layers and a hidden size of 768 is used to obtain a unique representation corresponding to each utterance. The RoBERTa transformer model has been trained on 160 GB of long sequences as compared to just 16 GB for BERT. If each sample comprises  $n$  words,  $w_1, w_2, \dots, w_n$ , then  $w_n \in R^{d_T}$ , where  $d_T = 768$ .

**Acoustic or Speech Features:** The low-level acoustic features corresponding to the acoustic data stream for each utterance are extracted using OpenSmile [21]. These features describe the pitch and tone of the speaker. The features, including zero crossing rate and mel frequency cepstrum coefficients, are obtained corresponding to each acoustic segment, and their mean is taken to yield the final acoustic feature set. The acoustic feature set for each utterance is represented by  $a_u$ , where  $a_u \in R^{d_a}$  and  $d_a = 1000$ .

**Video Features:** We used the output of the pool 5 layer of the ResNext-101 [22] model pretrained on the ImageNet dataset for image classification to extract the video features for all frames present in each utterance. By incorporating an extra cardinality dimension in addition to the depth and width present in the conventional ResNet models, ResNext has diminished the number of hyperparameters required for training. The features are extracted after initial preprocessing of video frames, including resizing and normalizing frames. Corresponding to each sample utterance, the final visual

feature representation is obtained by taking the average of the 2048-dimensional feature vector.

### IV. EXPERIMENTAL OUTCOMES

Table 1 and Table 2 illustrate the experiment outcomes of our proposed approach. Our approach outperforms the current cutting-edge methods, indicating improved accuracy and robustness. The graphical representations in Figure 3 and Figure 4 illustrate how our technique compares to past methods in speaker-dependent and speaker-independent settings.

TABLE I. EXPERIMENT ANALYSIS OF THE MUSTARD DATASET FOR SPEAKER-DEPENDENT SETTING

Modalities (text + visual + acoustic)	Methods	Speaker-Dependent			
		$\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F1}$
	IWAN [17]	-	75.2	74.6	74.5
	[23]	-	73.8	73.62	73.58
	MuLOT [15]	78.57	-	-	-
	[5]	-	71.9	71.4	71.5
	Proposed (Ours)	79.32	78.1	77.42	77.6

TABLE II. EXPERIMENTAL ANALYSIS OF THE MUSTARD DATASET FOR SPEAKER-INDEPENDENT SETTING

Modalities (text + visual + acoustic)	Methods	Speaker-Independent			
		$\mathcal{A}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F1}$
	IWAN [17]	-	71.9	71.3	70.0
	[23]	-	63.85	63.09	62.45
	MuLOT [15]	-	-	-	-
	[5]	-	64.3	62.6	62.8
	Proposed (Ours)	77.32	74.21	73.7	73.94

#### 1) Dataset



The proposed model was tested on the MUSTARD dataset [5]. This multimodal sarcasm recognition dataset comprises 690 samples (50% each of sarcastic and non-sarcastic utterances), collected from four famous TV shows: Sarcasmoholics (2.03%), The Golden Girls (5.8%), The Big Bang Theory (40.58%), and Friends (51.6%). Each utterance (or video) and its context comprise data corresponding to three modalities: textual transcription, audio and video. The context has been provided to deal with cases where the sarcastic tone of an utterance cannot be judged without knowledge of the conversational context. The proposed model is tested in two different experimental configurations—speaker-dependent and speaker-independent. All utterances belonging to the Friends series are used for testing and all others for training in the speaker-independent setting.

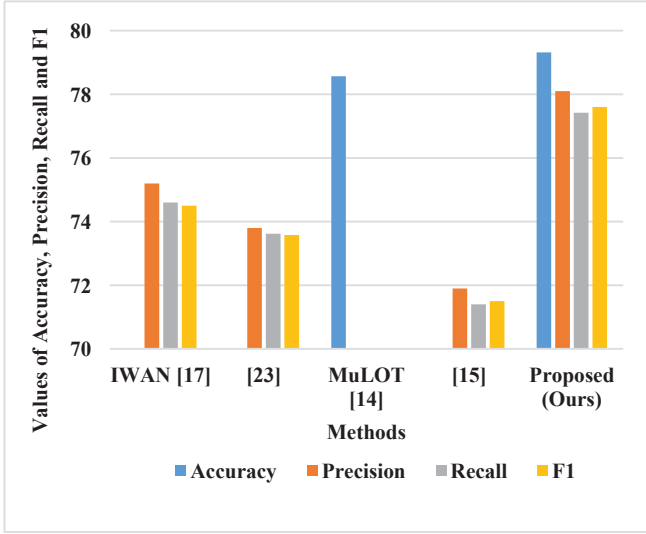


Fig. 3. Comparison of our proposed method with the existing cutting-edge approaches for speaker-dependent scenario

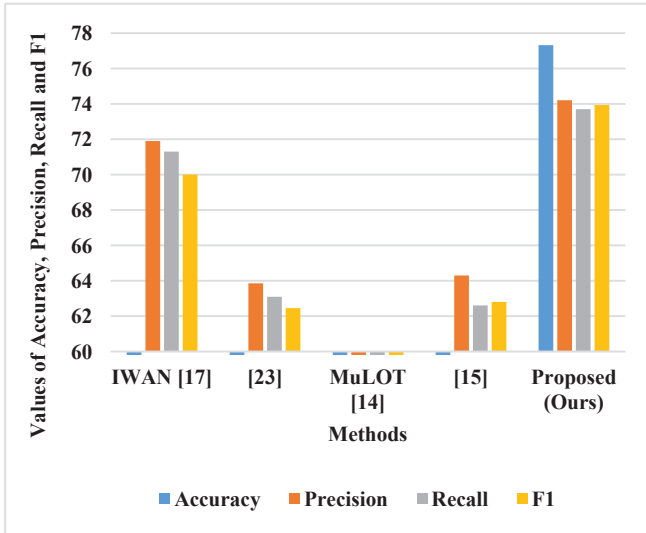


Fig. 4. Comparison of our proposed method with the existing cutting-edge approaches for speaker-independent scenario

## 2) Implementation Details

The model was constructed and trained using the Python programming language with Keras libraries. The model was trained for 100 epochs using the Adam optimizer with cross-

entropy categorical loss calculations. The learning rate was set to 0.0001.

## B. Evaluation parameters

**Accuracy:** Accuracy is the measure of how often the classifier is correct. It represents the proportion of accurate predictions to the total number of predictions.

$$Accuracy, A = \frac{TN + TP}{FP + FN + TP + TN}$$

**Precision:** Precision, also known as positive predictive value, quantifies the percentage of correct positive predictions. It represents the proportion of correct predictions made relative to the total.

$$Precision, P = \frac{TP}{FP + TP}$$

**Recall:** The proportion of true positives that the model accurately predicted is known as the recall, hit rate, or true positive rate. It is the ratio of true positives to all the positives in the ground truth.

$$Recall, R = \frac{TP}{FN + TP}$$

**F-Score:** The F-score combines the two-performance metrics, R & P. The F1 score is the harmonic mean of the R & P scores. If the F1 score is high, then both R & P are also high.

$$F1 - score = \frac{2 \times R \times P}{R + P}$$

## V. CONCLUSION & FURTHER PROSPECTS

In this paper, we first demonstrate a comprehensive overview of the sarcasm recognition problem and the main methods that have been used to tackle it. Thereafter, we also classify the related work on the basis of the modalities of data used for sarcasm recognition, including unimodal data in the form of text and audio alone and multimodal samples in the form of visual-caption pairs and videos. We propose a novel model for multimodal sarcasm recognition on the MUSTARD dataset that employs bimodal multi-head attention as the mechanism for feature fusion. The empirical results indicate the efficacy of the introduced framework for sarcasm recognition. For future studies, we can consider experimenting with more complex feature fusion methods to obtain better results. In place of simply studying visual features for the entire frame in general, a specific focus on facial expressions and poses can enhance the accuracy of recognition. The MUSTARD dataset is still too small to employ complex architectures. More samples must be incorporated into the dataset to better learn a complex phenomenon like sarcasm.

## REFERENCES

- [1] D. Davidov, O. Tsur, and A. Rappoport, ‘Semi-Supervised Recognition of Sarcasm in Twitter and Amazon’, in *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, Uppsala, Sweden: Association for Computational Linguistics, Jul. 2010, pp. 107–116. Accessed: Apr. 15, 2023. [Online]. Available: <https://aclanthology.org/W10-2914>
- [2] R. González-Ibáñez, S. Muresan, and N. Wacholder, ‘Identifying Sarcasm in Twitter: A Closer Look’, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 581–586. Accessed:

- Apr. 15, 2023. [Online]. Available: <https://aclanthology.org/P11-2102>
- [3] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, 'Sarcasm as Contrast between a Positive Sentiment and Negative Situation', in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 704–714. Accessed: Apr. 12, 2023. [Online]. Available: <https://aclanthology.org/D13-1066>
  - [4] A. Kumar, V. T. Narapareddy, V. Aditya Srikanth, A. Malapati, and L. B. M. Neti, 'Sarcasm Detection Using Multi-Head Attention Based Bidirectional LSTM', *IEEE Access*, vol. 8, pp. 6388–6397, 2020, doi: 10.1109/ACCESS.2019.2963630.
  - [5] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, 'Towards Multimodal Sarcasm Detection (An Obviously Perfect Paper)', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4619–4629. doi: 10.18653/v1/P19-1455.
  - [6] J. Tepperman, D. Traum, and S. Narayanan, 'Yeah right: Sarcasm recognition for spoken dialogue systems', Sep. 2006. doi: 10.21437/Interspeech.2006-507.
  - [7] A. Pandey and D. K. Vishwakarma, 'Attention-based Model for Multi-modal sentiment recognition using Text-Image Pairs', in *2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT)*, Feb. 2023, pp. 1–5. doi: 10.1109/ICITIIT57246.2023.10068626.
  - [8] A. Pandey and D. K. Vishwakarma, 'VABDC-Net: A framework for Visual-Caption Sentiment Recognition via spatio-depth visual attention and bi-directional caption processing', *Knowl.-Based Syst.*, vol. 269, p. 110515, Jun. 2023, doi: 10.1016/j.knosys.2023.110515.
  - [9] R. Schifanella, P. de Juan, J. Tetreault, and L. Cao, 'Detecting Sarcasm in Multimodal Social Platforms', in *Proceedings of the 24th ACM international conference on Multimedia*, Oct. 2016, pp. 1136–1145. doi: 10.1145/2964284.2964321.
  - [10] S. Sangwan, M. S. Akhtar, P. Behera, and A. Ekbal, 'I didn't mean what I wrote! Exploring Multimodality for Sarcasm Detection', in *2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, United Kingdom: IEEE, Jul. 2020, pp. 1–8. doi: 10.1109/IJCNN48605.2020.9206905.
  - [11] H. Pan, Z. Lin, P. Fu, Y. Qi, and W. Wang, 'Modeling Intra and Inter-modality Incongruity for Multi-Modal Sarcasm Detection', in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online: Association for Computational Linguistics, Nov. 2020, pp. 1383–1392. doi: 10.18653/v1/2020.findings-emnlp.124.
  - [12] Y. Cai, H. Cai, and X. Wan, 'Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2506–2515. doi: 10.18653/v1/P19-1239.
  - [13] B. Liang *et al.*, 'Multi-Modal Sarcasm Detection via Cross-Modal Graph Convolutional Network', in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1767–1777. doi: 10.18653/v1/2022.acl-long.124.
  - [14] D. S. Chauhan, D. S. R., A. Ekbal, and P. Bhattacharyya, 'Sentiment and Emotion help Sarcasm? A Multi-task Learning Framework for Multi-Modal Sarcasm, Sentiment and Emotion Analysis', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, Jul. 2020, pp. 4351–4360. doi: 10.18653/v1/2020.acl-main.401.
  - [15] S. Pramanick, A. Roy, and V. M. Patel, 'Multimodal Learning using Optimal Transport for Sarcasm and Humor Detection'. arXiv, Oct. 21, 2021. doi: 10.48550/arXiv.2110.10949.
  - [16] '(PDF) Multimodal Sarcasm Detection via Hybrid Classifier with Optimistic Logic'. [https://www.researchgate.net/publication/364035312\\_Multimodal\\_Sarcasm\\_Detection\\_via\\_Hybrid\\_Classifier\\_with\\_Optimistic\\_Logic](https://www.researchgate.net/publication/364035312_Multimodal_Sarcasm_Detection_via_Hybrid_Classifier_with_Optimistic_Logic) (accessed Apr. 10, 2023).
  - [17] Y. Wu *et al.*, 'Modeling Incongruity between Modalities for Multimodal Sarcasm Detection', *IEEE Multimed.*, vol. 28, no. 2, pp. 86–95, Apr. 2021, doi: 10.1109/MMUL.2021.3069097.
  - [18] A. Ray, S. Mishra, A. Nunna, and P. Bhattacharyya, 'A Multimodal Corpus for Emotion Recognition in Sarcasm'. arXiv, Jun. 05, 2022. doi: 10.48550/arXiv.2206.02119.
  - [19] D. S. Chauhan, G. V. Singh, A. Arora, A. Ekbal, and P. Bhattacharyya, 'An emoji-aware multitask framework for multimodal sarcasm detection', *Knowl.-Based Syst.*, vol. 257, p. 109924, Dec. 2022, doi: 10.1016/j.knosys.2022.109924.
  - [20] Y. Liu *et al.*, 'RoBERTa: A Robustly Optimized BERT Pretraining Approach'. arXiv, Jul. 26, 2019. doi: 10.48550/arXiv.1907.11692.
  - [21] F. Eyben, M. Wöllmer, and B. Schuller, 'Opensmile: the munich versatile and fast open-source audio feature extractor', in *Proceedings of the 18th ACM international conference on Multimedia*, in MM '10. New York, NY, USA: Association for Computing Machinery, Oct. 2010, pp. 1459–1462. doi: 10.1145/1873951.1874246.
  - [22] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, 'Aggregated Residual Transformations for Deep Neural Networks'. arXiv, Apr. 10, 2017. doi: 10.48550/arXiv.1611.05431.
  - [23] N. Ding, S. Tian, and L. Yu, 'A multimodal fusion method for sarcasm detection based on late fusion', *Multimed. Tools Appl.*, vol. 81, no. 6, pp. 8597–8616, Mar. 2022, doi: 10.1007/s11042-022-12122-9.



# Multi-objective optimization of mechanical properties of chemically treated bio-based composites using response surface methodology

Ankit Manral<sup>a</sup>, Rakesh Singh<sup>a</sup>, Furkan Ahmad<sup>b</sup>, Partha Pratim Das<sup>c</sup>, Vijay Chaudhary<sup>d,\*</sup>, Rahul Joshi<sup>e</sup>, Pulkit Srivastava<sup>f</sup>

<sup>a</sup> Department of Mechanical Engineering, Noida Institute of Engineering & Technology, Greater Noida 201308, India

<sup>b</sup> Department of Mechanical Engineering, Delhi Technological University, Shahbad Daultpur, Rohini, New Delhi 110042, India

<sup>c</sup> Department of Materials Science and Metallurgical Engineering, IIT Hyderabad, Sangareddy, Telangana, 502285, India

<sup>d</sup> Department of Mechanical Engineering, Amity School of engineering and Technology, Amity University Uttar Pradesh, Noida, 201313 India

<sup>e</sup> ME Department, Netaji Subhas University of Technology, Sec-3, Dwarka, New Delhi, 110078, India

<sup>f</sup> Department of Mechanical Engineering, Sharda University, Greater Noida 201308, India

## ARTICLE INFO

### Keywords:

Kenaf fiber mats  
Sodium acetate  
Response  
RSM (CCD)  
Optimization

## ABSTRACT

Eco-friendly surface treatment of natural fibers using sodium acetate ( $\text{CH}_3\text{COONa}$ ) affects the mechanical properties of the developed composites in many ways. In present study, geometrically different kenaf fiber mats (bidirectional (BC), unidirectional (UD) and randomly oriented (RO)) were treated at different concentration (10, 15 and 20 percentage w/w) of sodium acetate aqueous solution for varying time (24, 48 and 72 hr.) at room temperature. PLA (Poly-Lactic Acid) was used for the fabrication of treated fiber reinforced bio-degradable composites. The influence of above parameters on mechanical properties were studied. Response surface methodology (RSM) module face centered central composite design was employed for the development of regression models. The relationship between chemical treatment parameters and mechanical responses were predicted by quadratic model. In this study, predicted model was developed for two numerical factors (chemical concentration (CC) and treatment time (TT)) and one categorical factor (type of mat (TOM)). Tensile strength (TS), flexural strength (FS) and impact strength (IS) are considered as response variables. The statistical analysis showed that chemical concentration, treatment time and kenaf mat type have individually and interactively influenced the response of experiments. Chemical concentration was found to be the most influencing factor among all for the changes in mechanical properties. Optimization of input variables was done based on predicted model within bounded reason of responses.

## 1. Introduction

From the last decade, natural fiber has been accepted as a reinforced material for thermoset and thermoplastic polymer because of their comparable properties such as low-density, high specific strength to weight ratio, availability, low cost etc. over synthetic fibers and it is available abundantly in nature. Using biopolymer and natural fibers, green composite materials are developed. As both the constituents of green composites are biodegradable in nature, it is being used to overcome the environmental problem that is being faced by the conventional polymer composites [1].

Natural fibers for green composites are obtained from various parts of plants and tress such as leaves, seeds and bast [2]. Fibers obtained

from different parts of the plants exhibits distinct properties. Some most popular natural fiber that are being used as reinforcement material with thermoset and thermoplastic based composites are Jute, hemp, flax, bamboo, pineapple, kenaf etc. Among all these natural fibers, kenaf is an important bast fiber extracted from plant via mechanical retting method. It has excellent acoustic properties and higher thermal stability compared to other natural fibers. Number of automotive companies are using kenaf fibers as reinforcement material for the development of automotive parts to enhance the acoustic behavior capabilities [3].

Cellulosic kenaf fiber has some drawbacks in extension to their advantages such as its hydrophilic nature and presence of unwanted constituents which influences the mechanical properties of developed composites. Furthermore, the properties of natural fibers depend upon

\* Corresponding author at: Amity University, India.

E-mail address: [vijaychaudhary111@gmail.com](mailto:vijaychaudhary111@gmail.com) (V. Chaudhary).

<https://doi.org/10.1016/j.jcomc.2022.100337>

the climate condition where it grow, environment or weather condition, soil characteristic etc. [2]. The foremost drawback with natural fibers is its lower compatibility with matrix materials. Lower compatibility results in lower interfacial adhesion between matrix and reinforcement which further affects stress transfer between the constituents negatively. Lower interfacial adhesion is due to presence of excess amount of non-cellulosic content lignin, hemicellulose, pectin and waxes. These constituents are bonded with fiber fibrils with hydrogen bonding [4]. All unwanted contents are bounded with fiber by hemicellulose matrix. But hemicellulose is hydrophilic in nature and can be easily hydrolysis by aqueous solution of acids and bases [5].

In order to achieve a good interfacial adhesion between the matrix and fiber, the surface of natural fibers is altered by chemical treatments (Silane, alkaline, mercerization and benzylation etc.). In some recent past studies, various authors have done the chemical treatment of natural fibers for enhancement of the various properties of developed composites. Oushabi et al. [6] studied the effect of alkaline treatment on mechanical properties of date palm fiber-polyurethane composites. The fibers were treated with 2 wt%, 5 wt% and 10 wt% NaOH (Sodium Hydroxide). From given result, authors concluded that at 5 % NaOH concentration, the TS of fiber increases 76 % as compares to untreated fiber. Moreover, the pull-out test show that at 5 % concentration the interface of fiber and matrix is higher. Moreover, the interface properties drastically decrease if further increase in chemical concentration for treatment.

Accordingly, it is required to find a chemical which do not have acidic nature, economical and must be ecofriendly for the treatment of fiber. In previous findings, various authors used sodium bicarbonate for ecofriendly treatment of natural fibers. Sodium bicarbonate called as baking soda used in cooking, baking, toothpastes and in various biopesticide [7, 8]. Recently, Chaitanya et al. [9] used sodium bicarbonate for the treatment of short aloe vera fiber and incorporated it with polylactic acid (PLA) with the help of injection molding process. Authors reported that at 10 percentage concentration of sodium bicarbonate, among the treatment time of 24, 48, 72, 120 and 168 hr., 72 hr. is the best treatment time and exhibited better mechanical properties compared to other treatment time. Fiore et al. [10] also used sodium bicarbonate for the treatment of sisal fibers. Authors reported that among different treatment time periods (24,120 and 240 hr.) of sisal fiber, the optimized time is 120 hr. At this treated time, fiber reinforced composites showed higher flexural properties of developed composites.

Considering all the reasoning related to chemical treatment of natural fibers, the current research intent to analyze a new greener way for the development of fully degradable biopolymer composites. In present study, different kenaf mats (UD, BD, and RO) treated with sodium acetate ( $\text{CH}_3\text{COONa}$ ) and incorporated with PLA by compression molding process. Sodium acetate is a type of salt used as a pickling agent, cleaning agent and as a food additive. The pH level of sodium acetate aqueous solution maintained between 7-8 that implies its non-acidic nature. Maintained pH level of aqueous solution give environmentally friendly treatment of natural fibers.

In present study, use two numeric factors: chemical concentration (CC) and treatment time (TT) and type of mat (TOM). Both numeric and categorical factor influenced the properties of developed composites. Whereas, from given experimental input factors, the optimized were existed between them that cannot be easily find from experiment or may be required a greater number of experiments. So, for minimizing the number of experiments and from economic point of view, introduce an optimizing software technique called design of experiment (DOE). In DOE central composite design module of response surface methodology (RSM) are employed for optimizing the factor. This technique develops regression model or governing equation for predicting the response for every range of input factor. The after effect of input variable (Chemical concentration, treatment time and type of kenaf mats) on mechanical properties (TS, FS and IS) of developed composites were studies based on contour and surface plots in DOE. Additionally, DOE optimized the

independent variables to get maximum output response for developed composites were also describe in this research.

## 2. Experimental description

### 2.1. Fiber and matrix

Raw mats (Untreated; UD and BD) form of kenaf fiber in 220 GSM were procured from Go Green products, Alwarthirunagar, Chennai, India. RO mat was prepared from raw kenaf fibers by using hand compression molding machine. Polylactic acid (Indego 3052D) were supplied from Nature Tech. India Pvt. Ltd, Nagalkeni, Chennai, and Tamilnadu, India in pellet form. As per supplier specification the density of PLA is  $1.46 \text{ cm}^3$  with glass transition and melting point temperature of  $55\text{--}60^\circ\text{C}$  and  $200^\circ\text{C}$  respectively. Sodium acetate ( $\text{CH}_3\text{COONa}$ ) was procured from Krishna chemicals New Delhi, India.

### 2.2. Sodium acetate treatment

Kenaf fibers were treated with sodium acetate ( $\text{CH}_3\text{COONa}$ ), with three different concentrations of 10%, 15% and 20%. The treatment was conducted at room temperature with three different treatment time periods of 24 hr., 48 hr. and 72 hr. As sodium acetate is not acidic, it requires more time and concentration for removal of unwanted content from the fiber surface. The reaction during treatment of natural fiber shown in Fig. 1.

After treatment, fiber mats were subsequently washed with running water to get relieve the sodium ions and dry in oven at  $70^\circ\text{C}$  for 8 hr. Aqueous solution of sodium acetate is mildly alkaline due to the formation of acetate and  $\text{OH}^-$ . The disintegration of  $\text{CH}_3\text{COONa}$  into acetate and hydroxyl ions as shown in Fig. 1. Therefor  $\text{CH}_3\text{COONa}$  aqueous solution react as in same manner as NaOH react with natural fiber as illustrate in Fig. 1. Mildly alkaline nature of  $\text{CH}_3\text{COONa}$  conventionally required more concentration and treatment time for treatment to get desired result. Hence, in present studies, chemical concentration and time for treatment are being investigated to get optimized results.

### 2.3. Composites fabrication

Composite samples were developed by hot compression molding process by using film stacking method. For every developed composite, kenaf fiber mats fraction was maintained of 30 %. Three different geometric oriented kenaf fiber mats (BD, UD, and RO) are used in this study. Four layers of kenaf mats were incorporated with polymer for every type of composites. To remove the moisture content from fiber, it can be preheated in oven at  $70^\circ\text{C}$  for 5 hr. before fabrication of composite. Polylactic acid (PLA) granules were converted in thin sheet of thickness 1 mm by compression molding machine at a temperature of  $150^\circ\text{C}$  initially and at low pressure for 2 min. Consequently, pressure was increase at constant temperature for next 3 min. Finally, the thin film of PLA allowed to cool inside a mold under pressure until the mold temperature is not equal to room temperature. Thereafter, every type of treated fiber reinforced composite (bidirectional fiber reinforced polymer composite (BDFRPC), unidirectional fiber reinforced polymer composite (BDFRPC) and randomly fiber reinforced polymer composite (ROFRPC)) kenaf fiber mats and polymer sheets are stacked alternatively inside a mold. All stacked layers are put inside the mold between

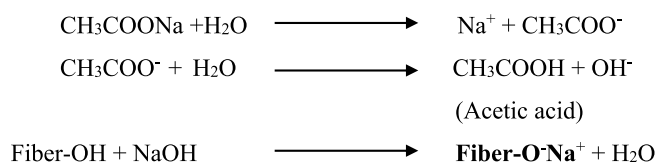


Fig. 1. Chemical reaction during treatment of fiber.



the Teflon sheets. Teflon sheets are introduced to overcome or avoid the sticking of polymer during fabrication of composites laminates.

During composite fabrication whole assembly or mold initially maintained at a temperature of 170 °C for 6 min. at low pressure. Thereafter, at a same temperature of 170 °C, pressure further increased and maintained the mold for given parameters for next 3 min. After that, composites sheets were removed from the mold when the temperature of mold reached 70 °C. The average samples thickness was 3 mm. The compaction pressure range during fabrication of composites laminates were 3-5 MPa After fabrication, all composite laminates were put in a desiccator to prevent it for moisture until further use.

#### 2.4. Tensile and flexural testing

Tensile and flexural testing of untreated and treated fiber reinforced polymer composites were carried out on INSTRON 5952 with constant cross head speed of 2 mm/min. All specimen for tensile and flexural testing were prepared according to ASTM D3039 and ASTM D790 respectively [11]. Every value reported represents average value of five specimen. The effect of different geometries of treated kenaf fiber mat reinforced composites on tensile and flexural mechanical properties of the developed composites has been studied and published [12]. This work is novel extension of the author's previous work as this work mainly focuses on the application of RSM in order to find the optimized results without spending much on the experimental work in future.

#### 2.5. Impact test

Notched Charpy test was conducted on the impact tester (Tinius Olsen model IT-503) at IIT Delhi. The impact tester has maximum free impact energy of 50 J and has a striking velocity of 3 m/sec. All specimen for impact testing were prepared as per the standard. The effect of different geometries of treated kenaf fiber mat reinforced composites on impact properties of the developed composites was previously studied by the same authors and has been published [13].

**Table 1**  
RSM randomized experimental table with experimental values.

Std	Run	Chemical concentration %	Treatment time (hr.)	Fiber mat (Type)	Response 1 Tensile strength (MPa)	Response 2 Flexural strength (MPa)	Response 3 Impact strength (J/m)
18	1	15	72	UD	83.57	73.56	67.8
24	2	20	72	RO	26.87	42.313	43.01
17	3	15	24	UD	77.29	68.46	93.8
9	4	15	48	BD	39.08	105.56	43.26
22	5	20	24	RO	45.56	74.36	62.5
23	6	10	72	RO	34.06	64.16	98.54
4	7	20	72	BD	45.9	88.4	28.75
11	8	10	24	UD	65.341	60.08	85.9
1	9	10	24	BD	32.15	89.94	54.54
3	10	10	72	BD	37.63	101.01	62.15
25	11	10	48	RO	32.46	63.16	103.9
15	12	10	48	UD	70.16	66.99	90.1
16	13	20	48	UD	90.46	81.24	40.9
30	14	15	48	RO	38.42	69.35	82.5
6	15	20	48	BD	55.37	112.45	30.65
20	16	15	48	UD	79.86	69.56	75
13	17	10	72	UD	75.56	67.5	92
29	18	15	48	RO	37.1	69.88	81.1
8	19	15	72	BD	45.02	106.06	38.95
21	20	10	24	RO	31.62	55.08	96.64
26	21	20	48	RO	52.4	76.78	52.96
2	22	20	24	BD	50.18	108.79	37.57
12	23	20	24	UD	85.709	76.832	48.4
10	24	15	48	BD	38.041	104.32	42.19
28	25	15	72	RO	40.16	70.13	75.2
5	26	10	48	BD	35.03	95.56	61.73
14	27	20	72	UD	74.326	46.53	35.5
7	28	15	24	BD	38.56	103.96	52.2
27	29	15	24	RO	35.56	66.76	97.5
19	30	15	48	UD	78.5	68	74.32

#### 2.6. Optimization process using RSM

Central composite module of RSM with full factorial were employed for optimization of two numerical factor (Chemical concentration and treatment time) and one categorical factor (type of kenaf mat). Addition of categorical factor in optimization, the existing runs is multiplied by categorical factor making the analysis have thirty runs.

Both numeric and categorical factor have three levels. Chemical concentration has three levels as 10, 15 and 20 percentage whereas, for treatment time it was 24, 48 and 72 hr. The three level of categorical factor are BD, UD, and RO mat types. Numerical and categorical factors have factorial, axial and center points. The developed experiments are in randomized form to overcome the unknown noise or error distorted the result of experiment. There is certain repetition in some experiment to overcome the formation of noise or error affecting during experiment. Based on the CCD, analysis was performed to develop regression models for TS, FS and IS in term of three input factors: CC, TT and TOM. Every input factor has three levels:  $\pm 1$  and 0. Table 2 shows the link between the input parameters with their corresponding selected levels. The experimental design matrix used to perform the experiment by combination of varying input variables. Experimental matrix with combination of different input variables is shown in Table 1. All experiments were strictly performed according to design Table 1.

**Table 2**  
Input factors and their corresponding varying level.

Symbols	Input parameters	Units	Levels		
			-1	0	+1
1.	Chemical concentration. (CC)	%	10	15	20
2.	Treatment time (TT)	hr.	24	48	72
3.	Kenaf mats	type	BD	UD	RO

### 3. Result and discussion

#### 3.1. Development of regression model

CCD methodology developed thirty experiments including few repetitions to minimize the noise arising during actual experiments. All experiments sequence is in randomized form, experimental conditions and their outcomes are shown in Table 1. Results are infused by “Design Expert (DX)” software (student version, Stat-Ease Inc., USA). Quadratic model was perfect to fit the results, whereas cubic model is aliased for developing the regression model. In built ANOVA in DOE generated a coded and actual equation (regression model) for prediction of the response at ever bounded inputs values are given by following equation in Table 3 and Table 4.

In above coded regression, all coded values A, B and C represent chemical concentration, treatment time and type of mat respectively. The outcomes obtained from analysis of variance by ANOVA for developed models are shown in Tables 5, 6 and 7.

Table 5 shows the ANOVA variance analysis of proposed model for TS. This table include sum of squares (SS), degree of freedom (df), mean square (MS), F and P- values. Table 5 implies that fitted model are significant ( $p < 0.05$ ) and have 95 % confidence level. The P-values of input variable, lower than .05 indicates its maximum significance in output response. For TS ANOVA regression model factors, A, C and interaction of AC are more significantly affect the response output for this model. Although the P-value of variable higher than 0.1 are not significantly influenced the response. So, for TS chemical concentration and mat type are more influencing input factors. For FS the variance analysis by ANOVA shown in Table 6, same as TS the fitted model for FS is significance as its P- value is less than 0.05. The input variable which are most significance in regression model of FS is C, AC and  $B^2$ , this implies that almost all input variables influence the FS obtained by regression model. The proposed model by the variance analysis done by ANOVA for IS shown in Table 7. Model P-values indicates the proposed model is significant and most significant variable that highly influenced the response are A, B, C, AB, AC and  $A^2$  but from the following A, B and C are highly influence the response. Predicted Vs actual values for TS, FS and IS shown in Fig. 2a, b and c respectively. Fig. 3a, b and c exhibit residual Vs predicted values for response. Actual values are the observed value during experiments, whereas predicted value are the values that generated by the regression model. Predicted values are based on semi-empirical model (Correlation) that was not equal but near by the actual value that observed during experiments. Although, Residual value is the difference between predicted and actual values. The minimum residual value shows the higher significance of model. R-squared values for the model developed for TS, FS and IS are 0.96, 0.94 and 0.97 respectively. The more the value of  $R^2$  approaches near to unity, there are more chances to better fit of model in experimental value.  $R^2$  value is statistical measure and explained how much independent variables influence dependent variables.

#### 3.2. The effect of input parameters on response

All three inputs parameters were influenced the output response. The

Table 3

Coded equation.

<b>Tensile strength</b>	$53.86 + 6.26 \times A + 0.0626 \times B - 10.70 \times C [1] + 25.68 \times C [2] - 4.37 \times AB + 1.51 \times AC [1] + .3077 \times AC [2] + 1.21 \times BC [1] + 0.7901 \times BC [2] + .0907 \times A^2 - 2.53 \times B^2$
<b>Flexural strength</b>	$84.01 + 2.46 \times A - 2.48 \times B + 23.38 \times C [1] - 10.35 \times C [2] - 9.19 \times AB + 1.40 \times AC [1] - .7844 \times AC [2] + 1.27 \times BC [1] - .4859 \times BC [2] - 4.22 \times A^2 - 5.43 \times B^2$
<b>Impact strength</b>	$67.56 - 19.60 \times A - 4.85 \times B - 19.49 \times C [1] + 4.66 \times C [2] - 3.52 \times AB + 7.00 \times AC [1] - 3.45 \times AC [2] + 1.48 \times BC [1] + 0.4806 \times BC [2] - 5.58A^2 - 0.1914 \times B^2$

Table 4

Actual equation.

<b>Tensile strength</b>	For Bidirectional mat	$-18.26109 + 3.19557 \times CC + 1.021548 \times TT - 0.036453 \times CC \times TT + 0.003628 \times CC^2 - 0.004391 \times TT^2$
	For Unidirectional mat	$22.57141 + 2.95537 \times CC + 1.00387 \times TT - 0.036453 \times CC \times TT + 0.003628 \times CC^2 - 0.004391 \times TT^2$
	For Randomly oriented mat	$-6.12452 + 2.53057 \times CC + 0.887440 \times TT - 0.004391 \times TT^2$
<b>Flexural strength</b>	For Bidirectional mat	$-16.60093 + 9.51082 \times CC + 2.00348 \times TT - 0.036453 \times CC \times TT - 0.168762 \times CC^2 - 0.009423 \times TT^2$
	For Unidirectional mat	$-40.26106 + 9.07422 \times CC + 1.93013 \times TT - 0.076603 \times CC \times TT - 0.168762 \times CC^2 - 0.009423 \times TT^2$
	For Randomly oriented mat	$-42.8446 + 9.10826 \times CC + 1.91753 \times TT - 0.076603 \times CC \times TT - 0.168762 \times CC^2 - 0.009423 \times TT^2$
<b>Impact Strength</b>	For Bidirectional mat	$20.53314 + 5.58205 \times CC + 0.331176 \times TT - 0.029313 \times CC \times TT - 0.223190 \times CC^2 - 0.000332 \times TT^2$
	For Unidirectional mat	$78.02981 + 3.49205 \times CC + 0.289578 \times TT - 0.029313 \times CC \times TT - 0.223190 \times CC^2 - 0.000332 \times TT^2$
	For Randomly oriented mat	$93.36681 + 3.47238 \times CC + 0.187912 \times TT - 0.029313 \times CC \times TT - 0.223190 \times CC^2 - 0.000332 \times TT^2$

contribution of input parameters according to p-value for generated regression model may vary for output response. Some response influenced by two variable or interaction of two and some of them were influenced by all input variables. Effect of input variable on response are shown in Figs. 4, 5 and 6. These plots of individual input variable were drawn at mean value of others.

##### 3.2.1. Effect of CC, TT and fiber mats on response

CC is a numeric factor, and it were enhanced the tensile and flexural response. The significance of CC was also shown in ANOVA Table 5 and 6. The variation of response respective to CC and mat type are linearly or curved. TS response increase with increase in CC for both BDFRPC and UDFRPC. But for ROFRPC after mean value of CC the tensile strength were almost constant. All values were examined according to the statics data at 48 hr. treatment time. Increment in tensile response after treatment is due to improve in interfacial adhesion between reinforced and matrix material. Treatment of kenaf fiber mats with sodium acetate remove the non-cellulosic content from the surface of fiber, which give higher interlocking of reinforced fiber with polymer. Mats type is a categorical factor, ANOVA table also show the significance of this categorical factor on all response. Orientation of fiber also a governing factor to handle the mechanical properties of developed composites. Manral et al. [13] study the effect of fiber orientation on mechanical properties of developed composites, tensile strength is maximum in longitudinal reinforced mats kenaf/PLA composites. In FS response the CC individually not significantly influencing the flexural response as shown in Fig 3b, it almost linear with small curvature at different chemical concentration. But the combine effect or interaction of CC and TT significant influenced the flexural response as per the ANOVA Table 6. The most dominating independent variable that boost up the flexural strength is type of fiber mat reinforced. Fig. 3b shows that BDFRPC achieved higher FS as compared to other oriented treated mat reinforced composites. Compared to CC and TT fiber mat show more significant effect for maximizing the output responses. All input variables and its interaction are significant for IS response according to the ANOVA Table 7. At mean treatment time the IS of developed composites were decrease with increase chemical concentration. ROFRPC achieved higher impact strength as shown in Fig. 4c. TT and mats type are also significant for IS response, with increase treatment time the response



**Table 5**

ANOVA analysis using input variable for TS.

Source	Sum of squares	df	Mean Square	F- values	P-values	
<b>Model</b>	11035.23	11	1003.20	46.07	< 0.0001	<b>significant</b>
A-Chemical concentration	706.43	1	706.43	32.44	< 0.0001	
B-Treatment time	0.0704	1	0.0704	0.0032	0.9553	
C-Fiber Mat	9982.83	2	4991.42	229.20	< 0.0001	
AB	229.62	1	229.62	10.54	0.0045	
AC	34.02	2	17.01	0.7810	0.4728	
BC	36.69	2	18.35	0.8424	0.4470	
A <sup>2</sup>	0.0576	1	0.0576	0.0026	0.9596	
B <sup>2</sup>	44.78	1	44.78	2.06	0.1687	
<b>Residual</b>	392.00	18	21.78			
Lack of Fit	389.67	15	25.98	33.37	0.0073	<b>significant</b>
Pure Error	2.34	3	0.7786			
<b>Cor Total</b>	11427.23	29				

**Table 6**

ANOVA analysis using input variable for FS.

Source	Sum of squares	df	Mean Square	F- values	P-values	
<b>Model</b>	9895.54	11	899.59	25.66	< 0.0001	<b>significant</b>
A-Chemical concentration	108.61	1	108.61	3.10	0.0954	
B-Treatment time	110.50	1	110.50	3.15	0.0928	
C-Fiber Mat	8234.64	2	4117.32	117.44	< 0.0001	
AB	1014.01	1	1014.01	28.92	< 0.0001	
AC	17.69	2	8.85	0.2523	0.7797	
BC	14.89	2	7.45	0.2124	0.8107	
A <sup>2</sup>	124.60	1	124.60	3.55	0.0757	
B <sup>2</sup>	206.20	1	206.20	5.88	0.0260	
<b>Residual</b>	631.08	18	35.06			
Lack of Fit	628.95	15	41.93	59.17	0.0031	<b>significant</b>
Pure Error	2.13	3	0.7087			
<b>Cor Total</b>	10526.62	29				

**Table 7**

ANOVA analysis using input variable for IS.

Source	Sum of squares	df	Mean Square	F- values	P-values	
<b>Model</b>	14406.69	11	1309.70	73.88	< 0.0001	<b>significant</b>
A-Chemical concentration	6916.84	1	6916.84	390.19	< 0.0001	
B-Treatment time	423.21	1	423.21	23.87	0.0001	
C-Fiber Mat	6212.67	2	3106.34	175.23	< 0.0001	
AB	148.47	1	148.47	8.38	0.0097	
AC	440.96	2	220.48	12.44	0.0004	
BC	37.54	2	18.77	1.06	0.3674	
A <sup>2</sup>	217.94	1	217.94	12.29	0.0025	
B <sup>2</sup>	.2565	1	0.2565	0.0145	0.9056	
<b>Residual</b>	319.08	18	17.73			
Lack of Fit	317.12	15	21.14	32.40	0.0076	<b>significant</b>
Pure Error	1.96	3	0.6525			
<b>Cor Total</b>	14725.77	29				

decreases linearly for all type of mats. Two factors' interactions of CC and TT or CC and mat types were also show have some significant effect on IS. Significance contribution of every input factor on response can be understood by R<sup>2</sup> values. If R<sup>2</sup> values of any factor is greater than .1 indicates that model term (input variables) is not significant.

Incorporation of different type of kenaf mat with polymer also influenced the response. Fig. 6 show the effect of mat type in composites on response generated by regression model. Geometric of mat influenced the output response, bidirectional mat enhanced the flexural properties, unidirectional mat influenced the tensile strength and for impact strength randomly oriented contribute more.

### 3.2.2. Effect of factor interaction on response

The individual and interacted effect of independent variable on response depicted in Figs. 7, 8 and 9. The interaction curved show how the responses changes with independent variables. Based on response curves tensile and flexural properties were increase with increased in CC

and with TT response were slightly increase and almost constant after 48 hr. of TT. Although, impact response was slightly increased with CC and then after response start to decrease, TT does not affect efficiently on impact response. For tensile response unidirectional mat is predominating for higher tensile strength, bi-directional mat composites show high flexural properties and for impact response randomly oriented mat contributing more compared to other kenaf mat reinforced composites. Response curves clearly indicate the importance of mat type in composites for enhancing its performance.

**3.2.2.1. Tensile strength.** 3D Response and contour surface plots for tensile strength at varying mat type are shown in Fig. 7 which show the effect of chemical concentration and treatment time on response. The curvilinear profile of response curve is because of quadratic model fitted. From response tensile response is first increase up to a certain limit then decrease. At every CC with varying TT the tensile response was increased significantly. Chemical treatment conditions improve the

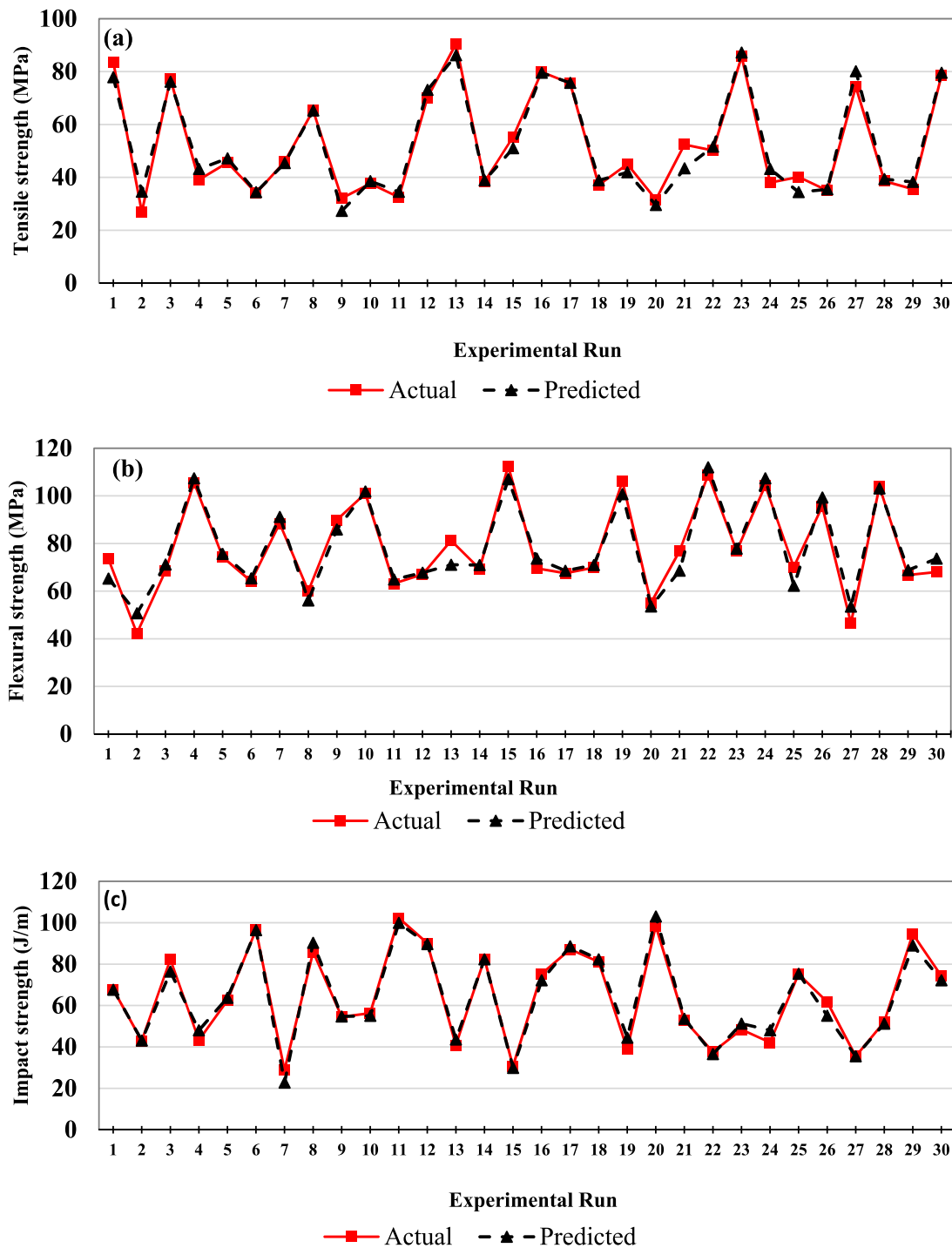


Fig. 2. Predicted vs. actual values of (a) Tensile strength, (b) Flexural strength and (c) Impact strength.

interfacial adhesion between reinforced and matrix that enhanced the response. In this study the chemical used for treatment of fiber is alkaline in nature, so it is required more time to remove the non-cellulosic constituent from the surface of fiber. Increment in response value increase subsequently with escalation in CC and TT. From the experimental observed value, the optimized condition for higher tensile response is at 20 % chemical concentration with 48hr. of treatment time. If TT is further increasing the tensile properties start to deteriorate, this is due to the damage of fiber surface. This trend of decreasing tensile response were seen in all type of kenaf mat reinforced composites. But it is interesting to note that tensile response is simply proportional to the CC and some for extent on TT if the individual effects of input variable

were studies.

Apart from two numerical factor, tensile response is also dependent on third input factor (mat type) that is categorical factor. For all three different level of mat type tensile response were vary according to it. Unidirectional kenaf mat reinforced composites achieved higher tensile response than other kenaf mat reinforced composites. Response clearly shows the importance of kenaf mats geometry on tensile strength. In unidirectional mat, fiber reinforced along the direction of load which makes them cable to convey higher tensile load. But these alignments of fibers are absent in other kenaf fiber mat reinforced composites, little bit alignment of fiber along the load were seen in BDFRPC but as compared to UDFRPC the fiber in warp direction practically just half. Additionally,

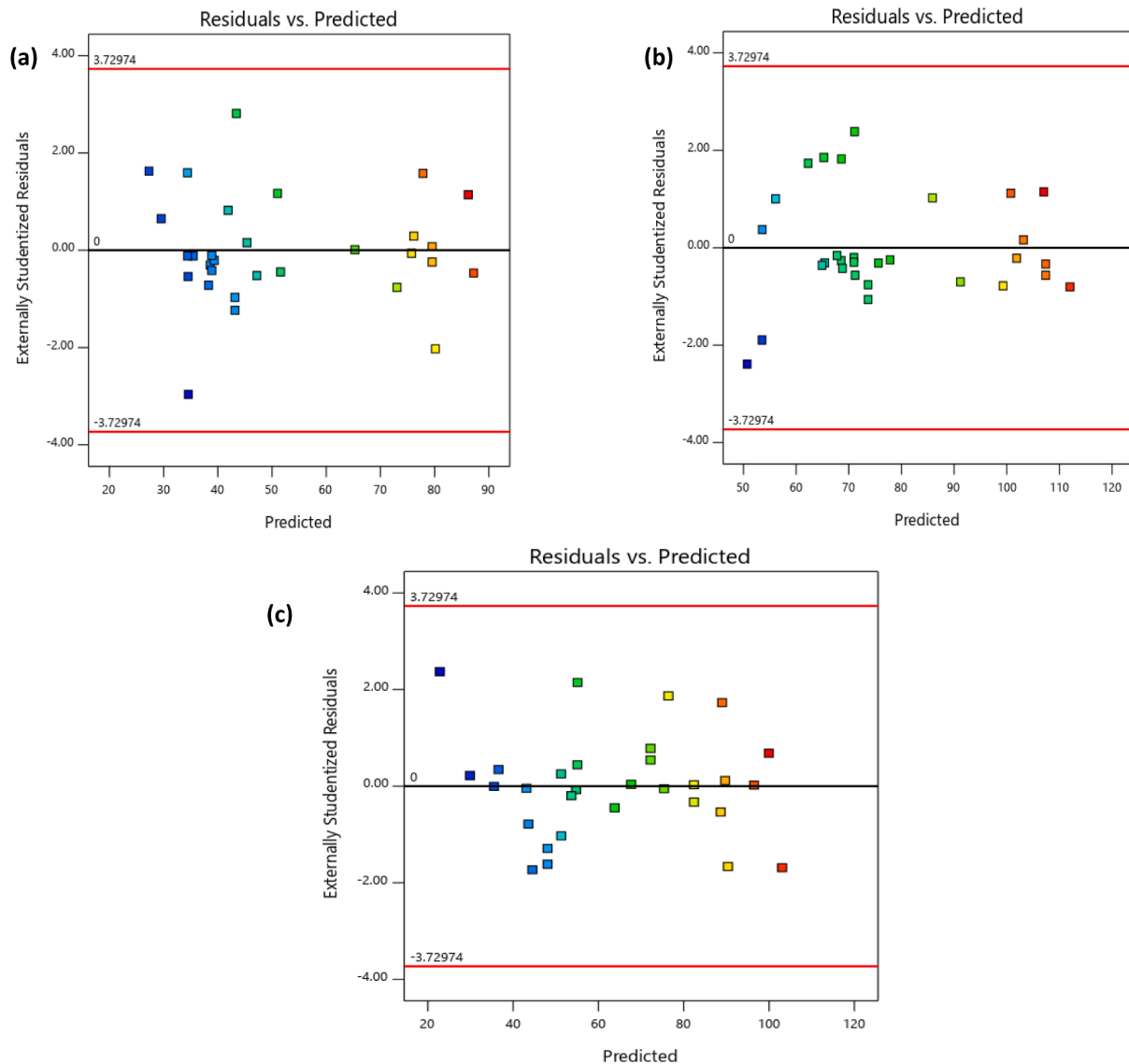


Fig. 3. Residual vs. predicted values of (a) Tensile strength, (b) Flexural strength and (c) Impact strength.

treatment of fiber with sodium acetate boosts up the tensile response. In Fig. 7 the individual effects of input variables on response can also be predict. For all type of kenaf mat reinforced composites TT show the same effect on response. Response first increases with increase in TT up to 48 hr. then further increment in TT value response remains constant and this nature behavior of tensile response were almost same for all type of kenaf mat composites. This variation of response with respect to TT is non-linearly. Although, the effect of CC on response are predominating, CC is the highly influencing factor for tensile response. The tensile response is directly proportional on CC, it was increased with CC ranges from lower to higher. This incremental variation of tensile response with respect to CC were same for all type of kenaf mat reinforced composites.

The interaction effect of CC and TT on tensile strength is shown in response contours in Fig. 7. The response curves ascertain that the increase in magnitude of CC and TT the tensile response was also increase. Increase CC with TT remove the non-cellulosic content from fiber surface sub sequentially. This interaction nature of numeric independent variables was almost same for all type of kenaf mat composites. But due to alignment of fiber in UDFRPC it has achieved higher tensile strength. At maximum CC if TT were further increased beyond approx. 48 hr. the tensile strength of developed composites was started to decrease due to

damage of fiber surface. At maximum CC if fibers were further treated from optimized value fiber surface start to damage, this damage fiber surface directly influenced the interfacial adhesion between fiber and matrix material. Lower interfacial interaction reduces the tensile strength of developed composites. Response curves in Fig. 7 show the importance of all three-input factor on tensile response.

**3.2.2.2. Flexural strength.** Flexural strength is the bending ability of any material to resist bending load. The response plots exhibiting the flexural strength of different kenaf fiber mat reinforced PLA composites at varying CC ranging from 10 % to 20 % and TT ranging from 24 hr. to 72 hr. are shown in Fig. 8. Response curve and contour plots indicated that CC, TT, and geometry of fiber mat influenced the flexural strength of developed composites. Based on response surface curve, it is feasible that the flexural strength increases with increase the concentration of sodium acetate and treatment time. The flexural strength value is differed for every kenaf mat composites. This is the evident that the geometric of kenaf mat is also contributed to enhancing the flexural properties of developed composites. According to the response value generated from regression model, the increment in response value is up to the 20 % CC and approx. 48 hr. of TT. Further if TT were increased the flexural properties start to deteriorate that are clearly envision in

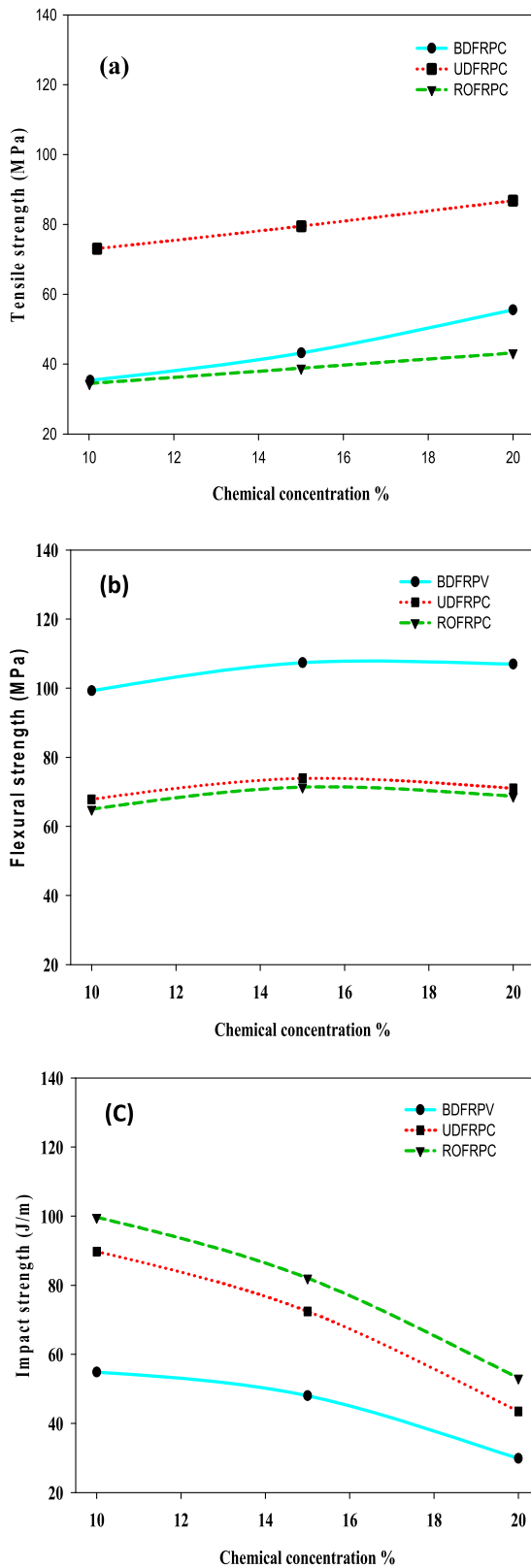


Fig. 4. Effect of chemical concentration on response (a) TS, (b) FS and (c) IS.

response curves. The apprehension behind is further increment in TT damage the surface of fiber which reduces the interfacial interaction between fiber and matrix material.

Bidirectional treated kenaf fiber mat composites show maximum flexural properties compared to other developed composites. In

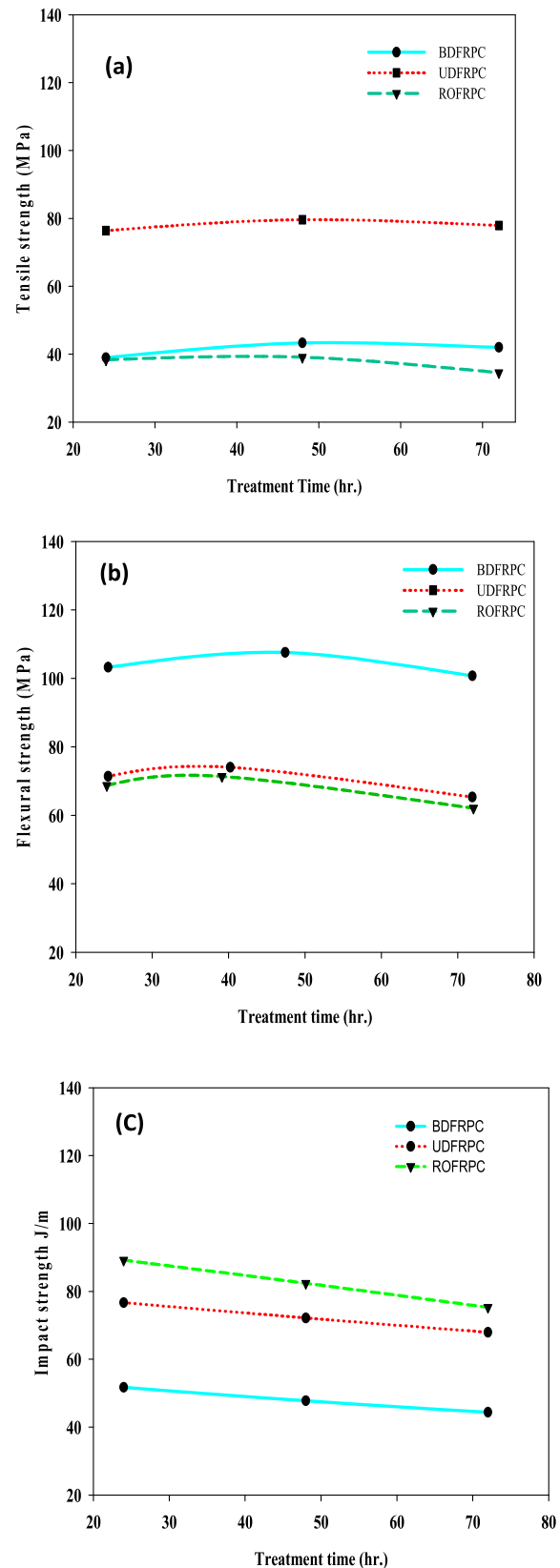


Fig. 5. Effect of treatment time on response (a) TS, (b) FS and (c) IS.

bidirectional mat fiber are aligned in warp and weft direction which develop a mat of inter-linking cross points. This inter-linking cross point mat developed composite resist the higher bending load. Whereas in other kenaf mats reinforced composites these inter-linking cross points

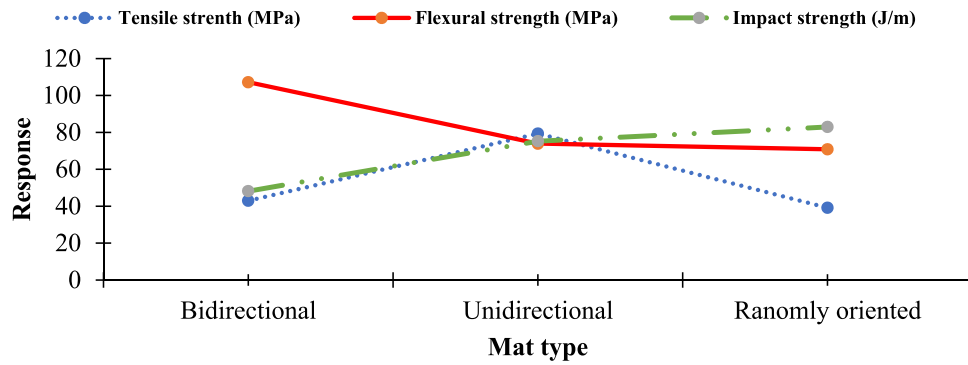


Fig. 6. Effect of mat types on response.

are absent which result drop in flexural properties. Effect of input numerical factors (CC and TT) on flexural response curve were almost the same. But the flexural response was highly affected by type of kenaf fiber mat used in reinforcement of composites. The parabolic profile of response curve is because of quadratic model fitted. As same tensile response if further TT increases the flexural properties of all developed composites start to deteriorate due to damage of fiber surface.

**3.2.2.3. Impact strength.** Impact response curves shown in Fig. 9, impact strength of any material tells about the energy absorbing capability of material during impact loading. Fig. 9 clearly understand that impact strength decreases with increase in CC and TT. In previous study various authors reported that, impact strength decreases with increase in CC or TT. Chaitanya et al. 2016 [9] done the experimental study on chemical treatment of Aloe Vera fiber at 10 % w/w of sodium bicarbonate with different treatment time of (24, 48, 72, 120, and 168 hr.). Authors concluded that impact strength of Aloe Vera /PLA composites is maximum up to 48 hr. TT after further increased in TT impact strength of developed composites starts to deteriorate. Fiber pullout, fiber fracture and crack propagation during impact loading are the main causes of composites failure [13, 14]. Improved bonding strength between fiber and matrix exhibited more fiber fracture instead of fiber pullout during impact loading. Fiber fracture instead of fiber pull out absorb lower energy during impact load [15]. Similar findings were observed in this study, after increase in CC with TT the impact strength of developed composites start to decrease. Response curves of impact strength with respect to input factors shown in Fig. 9. For Every impact response CC is highly influenced the impact strength, whereas effect of TT on impact strength is almost constant. But the interaction of these two factors CC and TT may influenced the impact strength of developed composites. Response contour curves indicates that ROFRPC achieved higher impact strength, the coordination of numeric input factor highly responsible for decreasing impact strength of developed composites. In ROFRPC, the arrangement of fiber in randomly form, that help to lock or arrested the crack formation arise during impact loading. Crack arresting capability of material improved its energy absorbing capability. Arresting of crack during impact load enhanced the material energy absorbing capability of material during sudden load. But this arrestment of cracks was absent or minimum in UDFRPC and BDFRPC, reason for lower impact strength compared to ROFRPC.

#### 4. Optimization of the conditions for tensile, flexural and impact response

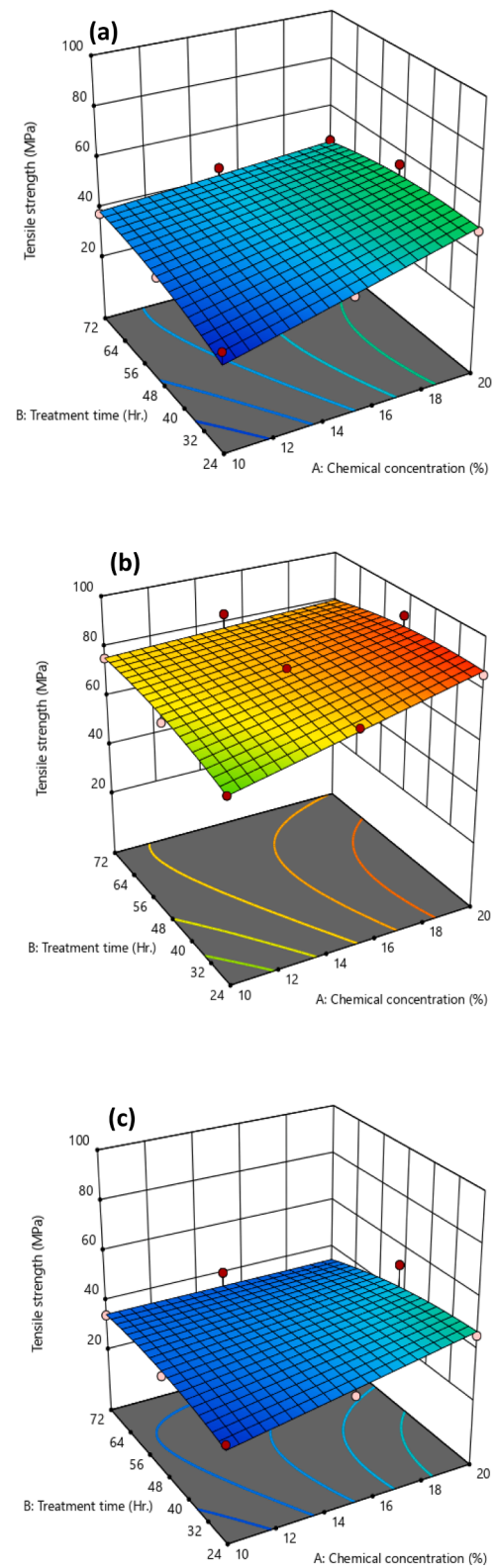
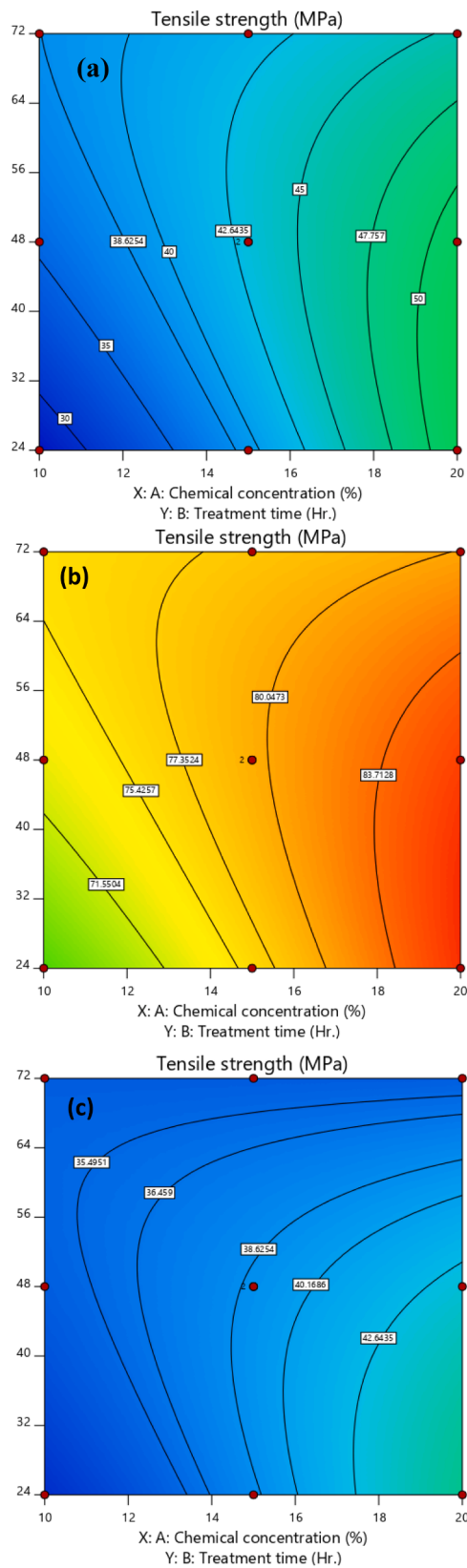
Regression model equation together solve to find the optimum input chemical treatment parameters. Design of experiment was used for maximization of tensile, flexural and impact response. Optimization of input parameters done based on response for individual mat and based on all comparable properties for individual mat type. The optimized output response is obtained by modeled equation by iterating several

runs whenever the optimal solution was not obtained. Optimization help to reduce the number of trails to get a better result that exist between the ranging values of input parameters. Additionally, optimization helps to minimize the magnitude of input factor that is very important for economical point of view and it reduce the unnecessary used of entity for optimization during experimental study.

Fig. 10 shows the ramp response for tensile strength after optimizing the treatment parameters for all type kenaf mat reinforced composites. The obtained response value according to the regression model that may be different from experimental value due to error and noise. For BDFRPC, the optimized chemical treatment condition for tensile response are TT of 32.53 hr. and CC of 18.99 %. These optimized values give a tensile response of value 61.89 MPa, this value is according to the regression model it was less than experimental value with desirability of 0.685. Similarly, the optimization of input factors by iterating its value in regression model to get maximum output response for other kenaf mat reinforced composites. UDFRPC achieved higher tensile response of value 87.43 MPa after optimizing the input parameters, CC of 24 % and TT of 31.28 hr. The optimum condition for ROFRPC are CC of 20 hr. and TT of 31.998 give tensile response of 46.51 MPa. It is evidence from optimized parameters value all value are approximate near to each other, for CC value near to 20 % and for TT it is 32 hr. As all optimized value are near to each other but after that the tensile response are different for each mat. This is the evident not only the numeric factor, but categorical factor is also highly influenced the tensile response.

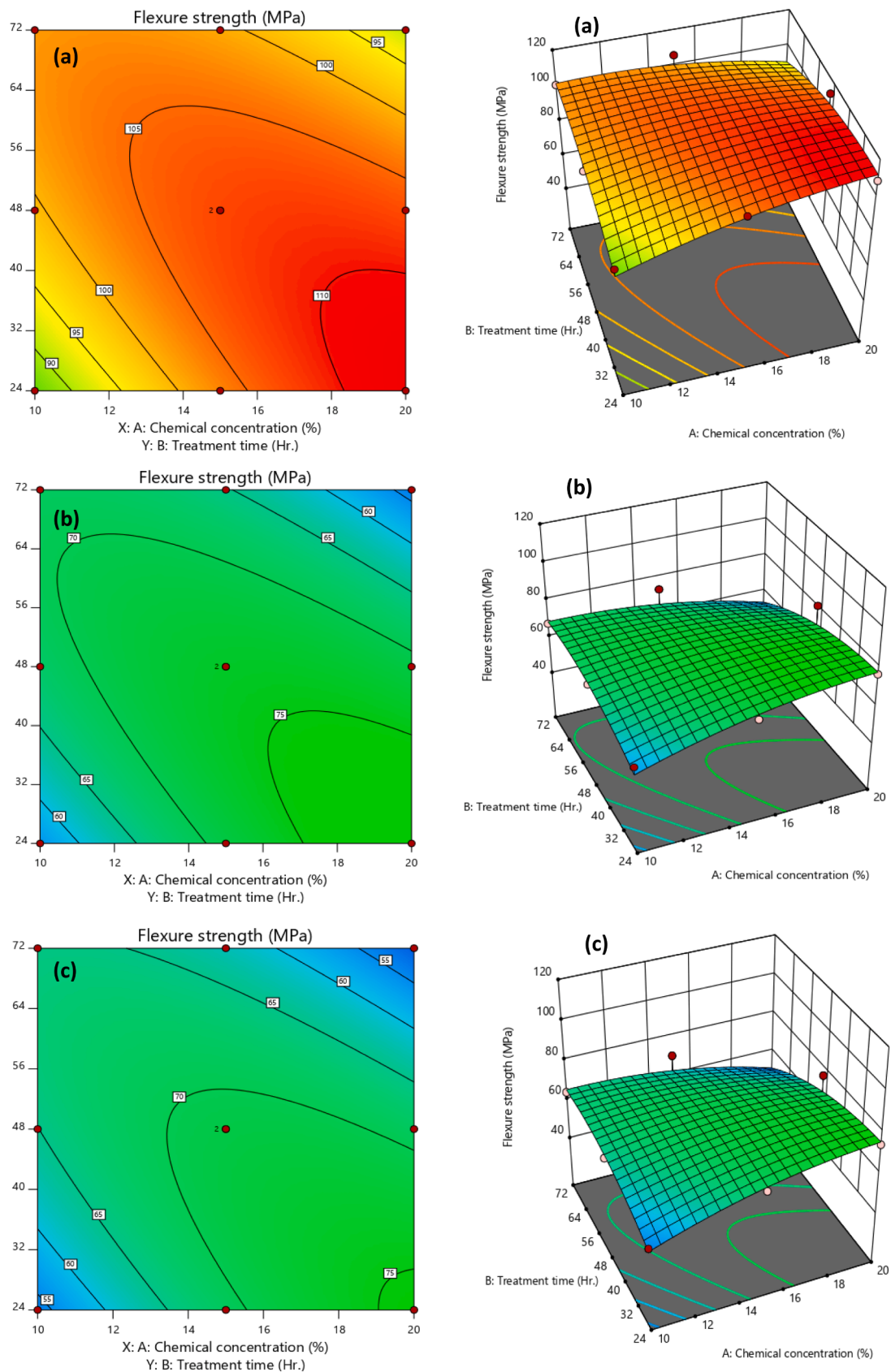
Similarly, the iteration of model equations was performed by design of experiment to find the optimum conditions for flexural response of chemical treated kenaf mats composites. Fig. 11 show the ramp response for flexural response after optimizing input numeric factor for individual kenaf mat type composites. BDFRPC achieved higher flexural strength of 112.45 MPa was experimentally observed at CC of 20 % and TT of 48 hr. But optimization according to design experiment after number of iterations of regression model. The optimized outcomes parameters are CC of 19.99 % with treatment time of 26.95 hr. give flexural response of 111.971 MPa. The difference in experimental and modeled value is due to noise and unwanted factors that are not considered in regression model design. UDFRPC and ROFRPC achieved lower flexural response value of 77.39 MPa and 75.60 MPa respectively according to regression model than BDFRPC. The optimized condition for UDFRPC is CC of 19.25 % and TT of 24 hr. whereas for ROFRPC CC of 19.93 % and TT of 24 hr. As these optimized factor values correlated with experimental input factor values the intensity of optimized factors are low. It means that optimization reduce the unwanted quantity of input factor that are unusable and have no effect on response. As same as tensile response all input optimized factor for flexural response are approx. same but have distinct flexure response value for developed kenaf mat reinforced composites. It implies that not only the input numeric but geometry of mat also contribution in enhanced the flexural response of developed composites.

Impact testing were performed to check the energy absorbing

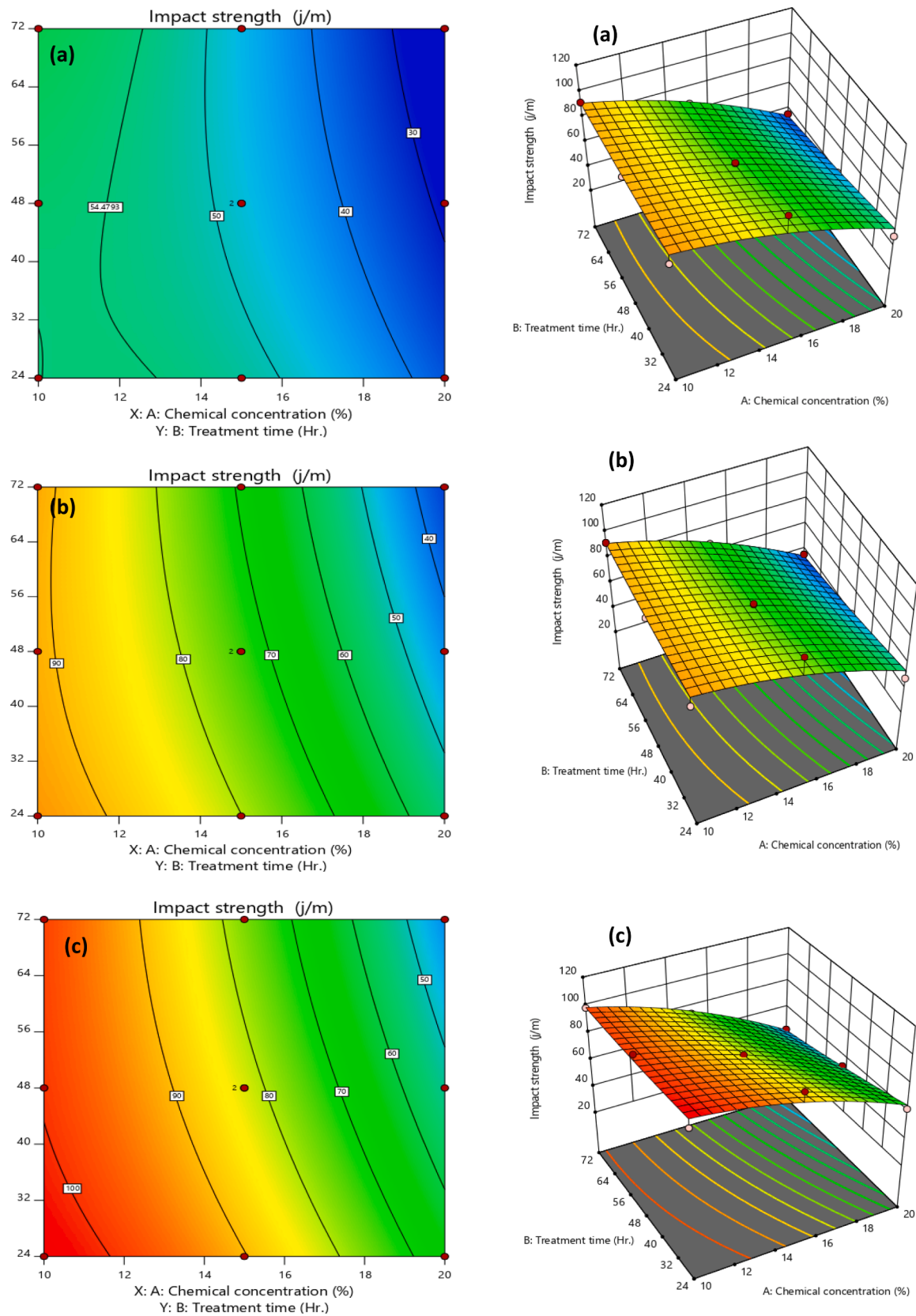


**Fig. 7.** Tensile Response surface of the effects two independent variables for individual mat type. (a) Bidirectional mat, (b) Unidirectional mat, (c) Randomly oriented mat.



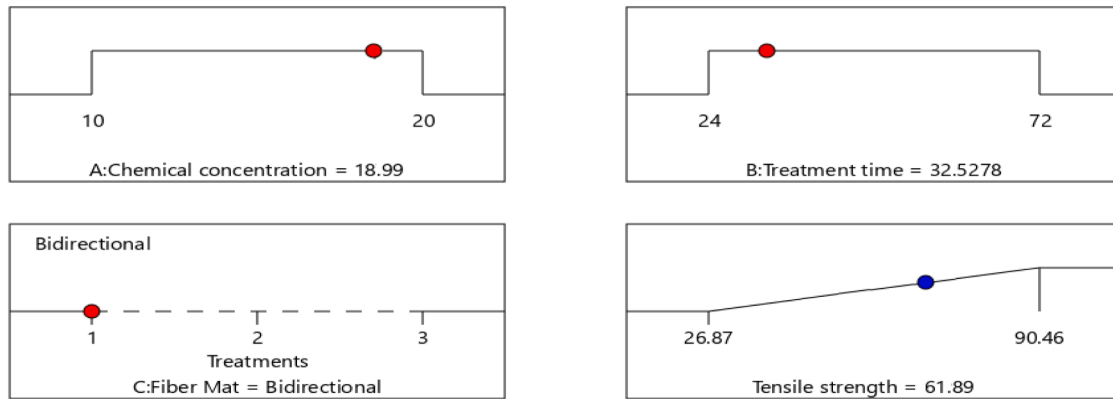


**Fig. 8.** Flexural Response surface of the effects two independent variables for individual mat type. (a) Bidirectional mat, (b) Unidirectional mat, (c) Randomly oriented mat.

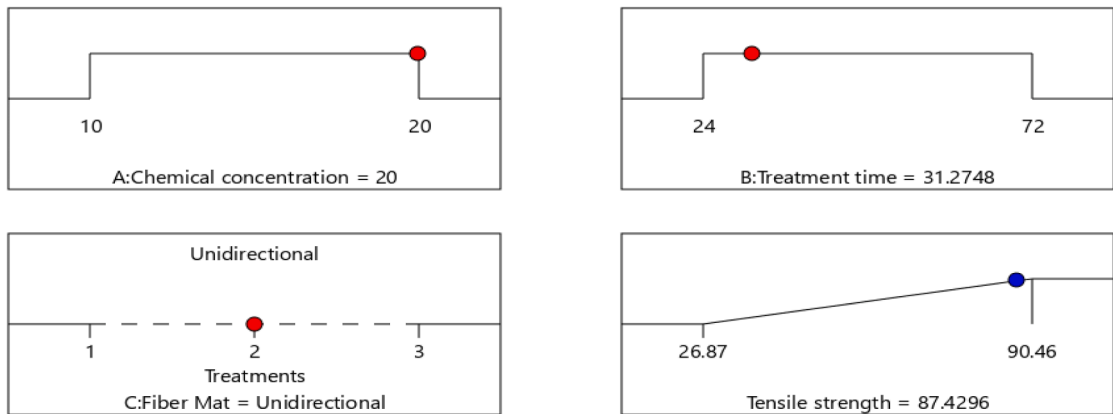


**Fig. 9.** Impact Response surface of the effects two independent variables for individual mat type. (a) Bidirectional mat, (b) Unidirectional mat, (c) Randomly oriented mat.

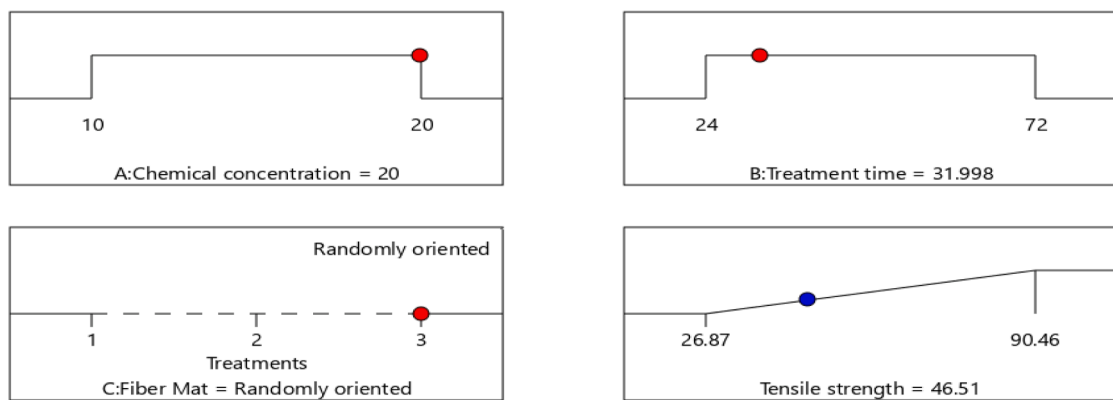
(a)

Desirability = **0.685**

(b)

Desirability = **0.952**

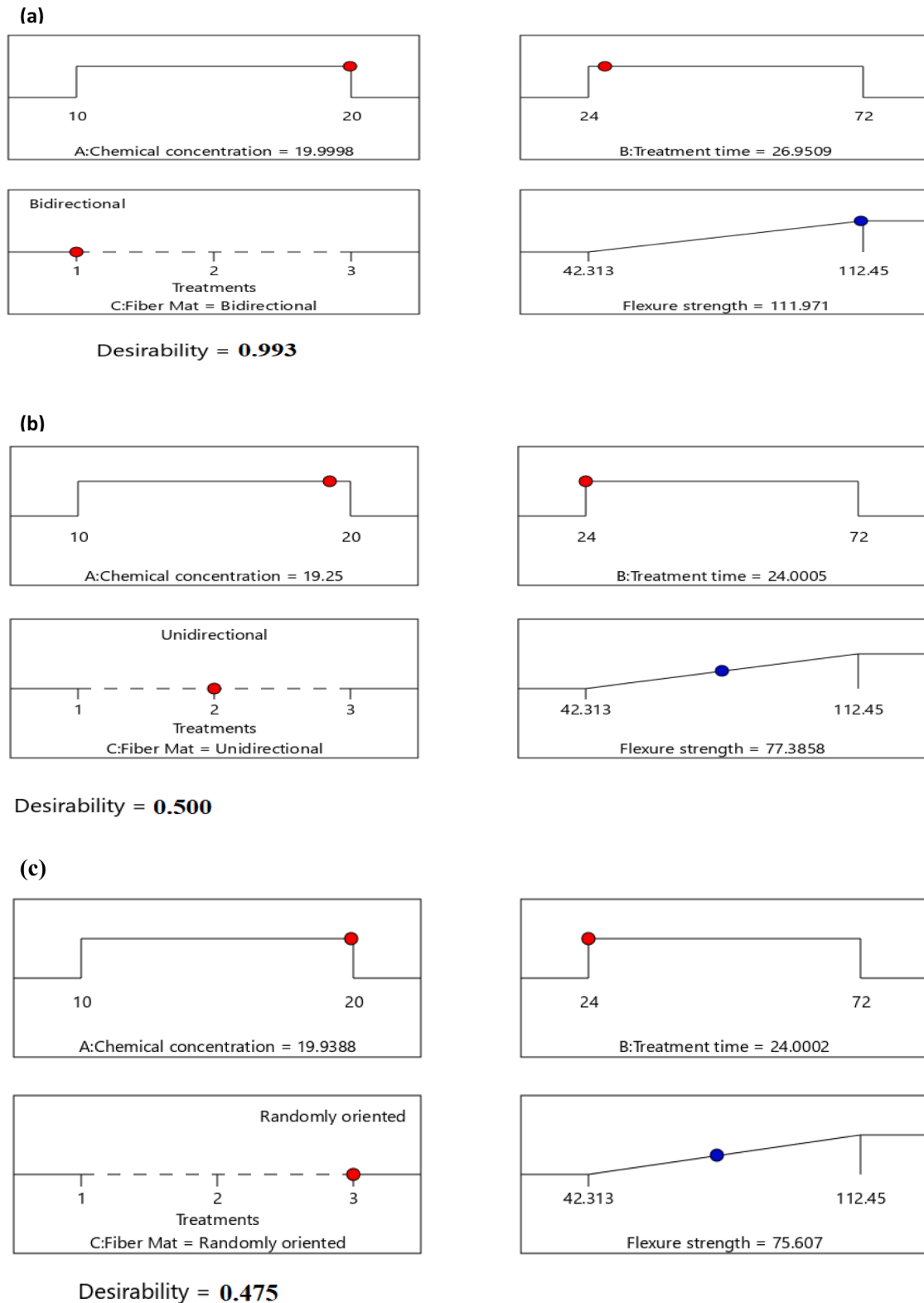
(c)

Desirability = **0.5141**

**Fig. 10.** Ramp function graph for chemical treatment condition and response as tensile strength for all different kenaf mat composites (a) BDFRPC, (b) UDFRPC and (c) ROFRPC.

capability of material during impact. Fig. 12 shows the effect of optimized parameters on impact strength of developed composites. ROFRPC composites achieved higher impact strength value of 102.3 J/m at CC of 20 % with treatment time of 48 hr. as experimentally observed. As per

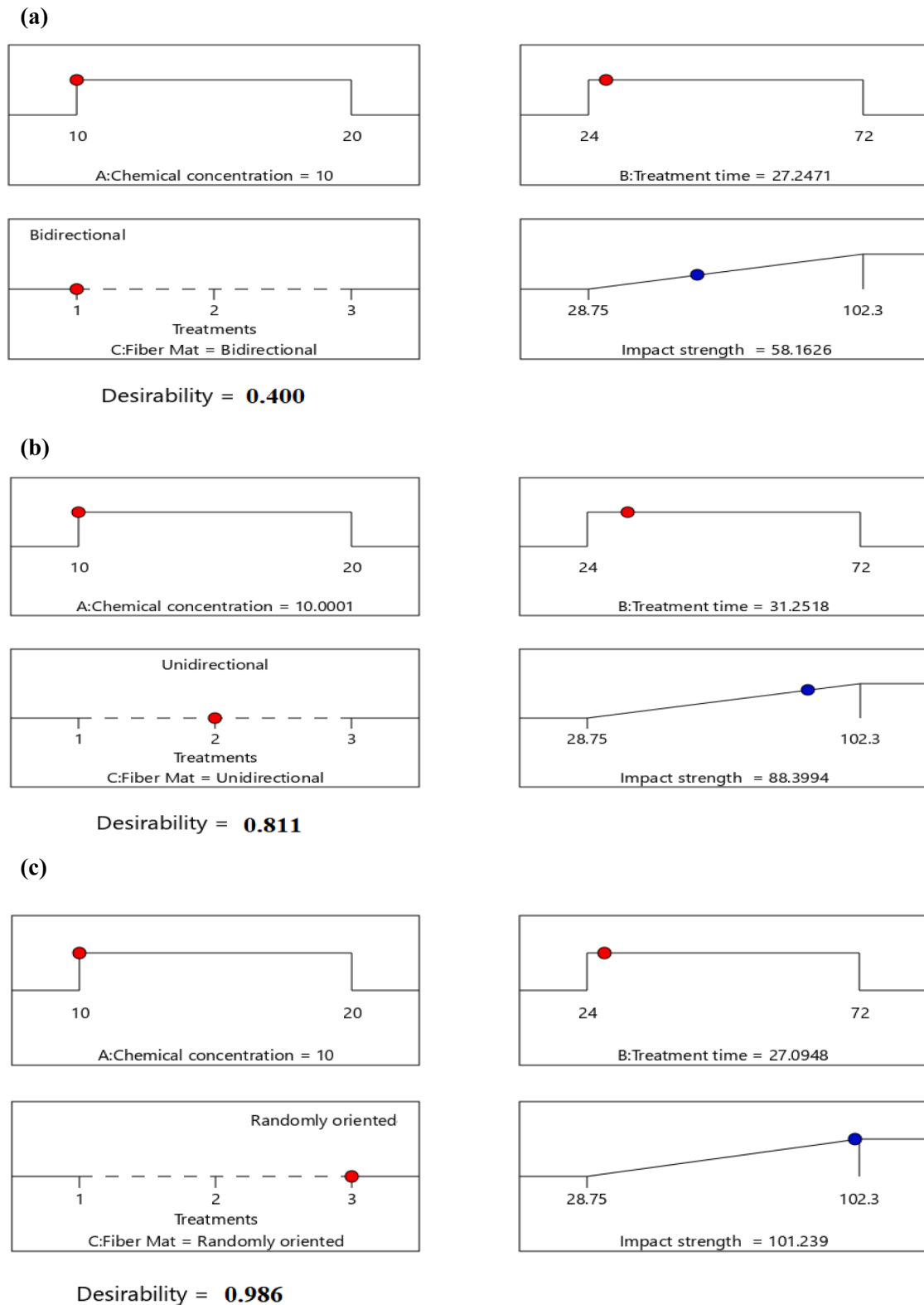
design experiments, the highest impact strength value is 101.239 J/m under the optimized conditions at 10 % of CC and 27.09 hr. TT. Difference in optimized impact response value compared to experimental value due not consideration of unwanted parameters during chemical



**Fig. 11.** Ramp function graph for chemical treatment condition and response as Flexural strength for all different kenaf mat composites (a) BDFRPC, (b) UDFRPC and (c) ROFRPC.

treatment, optimization reduced the unnecessary TT during experimental study. BDFRPC and UDFRPC achieved lower impact response as compared to ROFRPC. As shown in Fig. 12 the optimized parameters are approx. near to each other, but the impact response is different for each type of kenaf mat reinforced composites. It is the evident not only the

input parameter, but fiber geometry may influence the impact response. Increase in chemical removed non-cellulosic content from surface of fiber and improved the interfacial adhesion between fiber and matrix material. Higher interfacial adhesion minimized the fiber pullout chances during impact loading instead fiber fracture is generally seen in



**Fig. 12.** Ramp function graph for chemical treatment condition and response as Impact strength for all different kenaf mat composites (a) BDFRPC, (b) UDFRPC and (c) ROFRPC.

that condition. Fiber pullout in composites during impact loading absorb higher impact energy instead of fiber fracture. This is the evident that composite materials have higher impact strength at lower CC and treatment time. Higher CC and TT increase the interfacial adhesion, results low energy absorption due to fiber fracture. This trend of

decreasing impact strength with respect to increasing CC were seen in all type of kenaf mats reinforced composites. The fiber geometry is also a judging factor to decide its impact response.

Randomly arrangement of fiber in ROFRPC arrest the crack propagation during impact loading that enhanced the energy absorbing

capability of material. But the arrangement of fiber in UDFRPC and BDFRPC are different from ROFRPC, and they are not cabled to arrest the crack during impact loading. This is the reason for Unidirectional and bidirectional mats achieving least impact value than ROFRPC. The concentration of fiber in warp direction of UDFRPC is very high that helps to arrest the fiber fracture for some extent and absorb high impact energy hence UDFRPC exist impact strength between BDFRPC and

ROFRPC. Generated regression model equation helps to find out the response values at every range of response values.

#### 4.1. Optimization of parameters for individual kenaf mat composite

Individual composites have distinct mechanical properties. UDFRPC have higher tensile properties and for flexural and impact

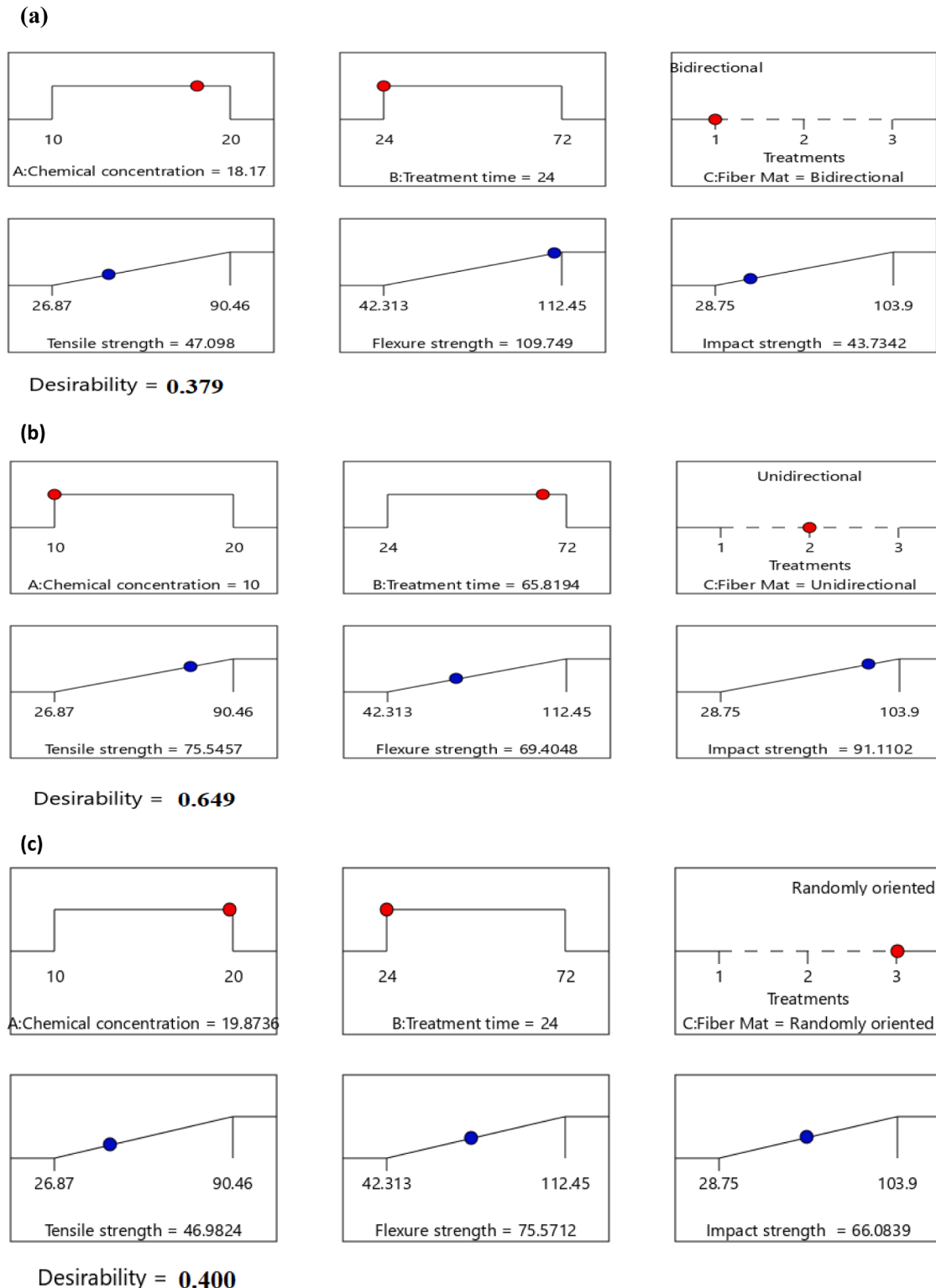


Fig. 13. Ramp function graph for chemical treatment condition and responses for all different kenaf mat composites (a) BDFRPC, (b) UDFRPC and (c) ROFRPC.



strength BDFRPC and ROFRPC are better options respectively. It means that every geometrically different kenaf mat composites is better option for individual application according to the type of load acting on it. Practically, materials are considered under varying type of load during application. So, it is required for every material that it can bear all type of loading condition. Design experimentation provide an environment to give an optimized condition for individual mat have permissible in all loading conditions. Fig. 13-a,b,c shows the optimized condition for higher mechanical properties for all kenaf mats composites. Fig. 13a. shows that at 18.17 CC with TT of 24 hr. BDFRPC achieved optimum mechanical properties (Tensile strength- 47.09 MPa, Flexural strength- 109.74 MPa and impact strength- 43.73 J/m). At these optimum parameters the BDFRPC achieved higher mechanical properties. As same as BDFRPC other kenaf mats reinforced composites optimized conditions are shown in Fig. 13b. and 13c. In Fig. 13b the optimized parameters for UDFRPC are CC of 10 % and TT of 65.81 hr. give optimum mechanical properties (Tensile strength- 75.5457 MPa, Flexural strength- 69.40 MPa and impact strength- 91.11 J/m). Similarly, for ROFRPC the optimized parameters are shown in Fig. 13c. at CC of 19.87 % with TT of 24 hr. the composite achieved higher mechanical properties (Tensile strength- 46.98 MPa, Flexural strength- 75.57 MPa and impact strength- 66.08 J/m). For given optimized input parameters maximum output responses were obtained for individual kenaf mat reinforced composites.

## 5. Conclusion

In order to optimize the chemical treatment conditions for maximized output responses, a set of experiments based on RSM central composite module was conducted. The response results recommended that this technique for optimization is compelling to minimize chemical concentration and treatment time without immolate the output response (Tensile, flexural and impact strength). Chemical treatment of kenaf fiber with sodium acetate enhanced the properties of developed composites. Experimental results indicated all three independent factor chemical concentration, treatment time and type of kenaf mat (categorical factor) contributed to enhance the properties of developed composites. For tensile, flexural and impact responses, the predominating numerical factor to enhanced the response is chemical concentration as compared to treatment time. The responses were successfully concluded from second order polynomial equation generated by RSM in built ANNOVA. This technique was also used to interrogate the effect of interaction of factors such as chemical concentration and treatment time for individual mat on response. Optimization of input parameters for maximizing the output responses were also predicted by RSM. The optimum conditions were obtained for individual kenaf mat composites as follows. For UDFRPC, the optimized condition for maximum tensile response of 87.43 MPa have chemical concentration of 20 % and treatment time of 31.27 hr. BDFRPC achieving high flexural response of 111.97 MPa have chemical concentration of 19.99 % and treatment time of 26.95 hr. and for maximum impact response of 101.239 J/m have chemical concentration of 10 % and treatment time of 27.09 hr. All values are obtained by number of iteration of regression equation. These values are mathematically generated and it have some error/difference in comparison to experimental values. For maximizing all properties of individual kenaf mat reinforced composites, the optimized condition are as follows: At

18.17 % of chemical concentration with 24 hr. of treatment time BDFRPC achieved maximum tensile, flexural and impact response of value 47.09 MPa, 109.75 MPa and 43.73 J/m respectively. Although, UDFRPC optimum conditions are at chemical concentration of 10% with treatment time of 65.81 hr. has corresponding output response tensile, flexural and impact are 75.55 MPa, 69.40 MPa and 91.11 J/m respectively. Similarly, for ROFRPC the optimized conditions for maximum response are chemical concentration of 19.87 % and treatment time of 24 hr. give tensile, flexural and impact response of 46.98 MPa, 75.57 MPa and 66.08 J/m respectively. This approach necessarily helped to minimize the number of experimental trail for optimizing the responses.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## References

- [1] A. Chamas, H. Moon, et al., Degradation rated of plastics in the environment, *ACS Sustain. Chem. Eng* 8 (2020) 3494–3511.
- [2] N Chand, M Fahim, Natural fibers and their composites, tribology of natural fiber polymer composites. Woodhead Publishing Series in Composites Science and Engineering, Woodhead Publishing, 2008, <https://doi.org/10.1533/9781845695057.1>.
- [3] F. Hassan, R. Zulkifli, et al., Kenaf fiber composite in automotive industry: an overview, *Int. J. Adv. Sci. Eng. Inform. Technol.* 7 (1) (2017) 315.
- [4] M.M Kabir, H. Wang, et al., Chemical treatments on plant-based natural fibre reinforced polymer composites: An overview, *Compos. Part B Eng.* 43 (2012) 2883–2892.
- [5] F Delbecq, Y Wang, et al., Hydrolysis of hemicellulose and derivatives—a review of recent advances in the production of furfural, *Front. Chem.* (2018), <https://doi.org/10.3389/fchem.2018.00146>.
- [6] A Oushabi, S Sair et al. (2017) The effect of alkali treatment on mechanical, morphological and thermal properties of date palm fibers (DPFs): Study of the interface of DPF–Polyurethane composite. 23; 116–123.
- [7] T. Lakhani, SIDS Initial Assessment Report for SIAM 15 (2022). <http://www.inchem.org/documents/sids/sids/sodbicarb.pdf> (Accessed 20 April).
- [8] Potassium bicarbonate (073508) and sodium bicarbonate (073505) fact sheet, [http://www3.epa.gov/pesticides/chem\\_search/reg\\_actions/registration/fs\\_G-135\\_01-Oct-04.pdf](http://www3.epa.gov/pesticides/chem_search/reg_actions/registration/fs_G-135_01-Oct-04.pdf) (Accessed 20 April 2022).
- [9] S Chaitanya, I Singh, Ecofriendly treatment of aloe vera fibers for PLA based green composites, *Int. J. Precis. Eng. Manufact. Green Technol.* 5 (2018) 143–150.
- [10] V. Fiore, T. Scalici, et al., A new eco-friendly chemical treatment of natural fibres: Effect of sodium bicarbonate on properties of sisal fibre and its epoxy composites, *Compos. Part B Eng.* 85 (2016) 150–160.
- [11] J.K. Wells, P.W.R. Beaumont, Crack-tip energy absorption processes in fibre composites, *J. Mater. Sci.* 20 (1985) 2735–2749.
- [12] A Manral, P Bajpai. Effect of non-acidic chemical treatment of kenaf fiber on physico mechanical properties of PLA based composites. *J Natural Fiber*.
- [13] Manral A, P.K Bajpai, Static and dynamic mechanical analysis of geometrically different kenaf/PLA green composite laminates, *Polym. Compos.* (2019), <https://doi.org/10.1002/pc.25399>.
- [14] M.J.A. Van den Oever, H.L. Bos, K. Molenveld, *Appl. Macromol. Chem.* 272 (1999) 71.
- [15] S chaitanya, I singh, Sisal fiber-reinforced green composites: effect of ecofriendly fiber treatment, *Polym. Compos.* (2017), <https://doi.org/10.1002/pc.24511>.

# Novel Band-Subtraction Technique to Differentiate Screws for Microwave Cavity Filter Tuning

Even Sekhri\*

Department of Electrical Power  
Engineering and Mechatronics  
Tallinn University of  
Technology  
Tallinn, Estonia  
ORCID: 0000-0002-2578-1910

Mart Tamre

Department of Electrical Power  
Engineering and Mechatronics  
Tallinn University of  
Technology  
Tallinn, Estonia  
ORCID: 0000-0002-7489-9683

Rajiv Kapoor

Department of Electronics and  
Communication  
Delhi Technological  
University  
New Delhi, India  
ORCID: 0000-0003-3020-1455

Dhanushka Chamara Liyanage

School of Engineering –  
Smallcraft Competence Center  
Tallinn University of  
Technology  
Tallinn, Estonia  
ORCID: 0000-0003-4526-0837

**Abstract**— Frequency response of a Microwave (MW) cavity filter is changed by rotating the tuning screws installed on the filter surface. Numerous screws are present on the surface of the filter, not all of which contribute to the alteration of tuning state as some of the screws are just the plate mounting screws. This paper presents a vision-based method for distinguishing the tuning screws of a cavity filter from the mounting screws. The tuning screws used in industry are coated with a conducting material to avoid losses. In this work, through hyperspectral imaging, characteristic image bands of screws of a commercial cavity filter were analyzed. From this analysis, the tuning screws were identified using their material properties since every material or compound has its own reflectance to Electromagnetic (EM) waves. The novel band subtraction technique proposed in this work distinguished all the tuning screws from the mounting screws. This proposed technique was then validated using a monochrome industrial camera attached with suitable optical bandpass filters. Achieving the classification accuracy of 100% with a monochrome camera proved the effectiveness of the proposed method. The results obtained can be used to identify and locate the tuning screws especially for the case when the technical drawing of the filter is not available. These extracted positional coordinates of tuning screws can assist in Fully Automated Tuning (FAT) of the cavity filters.

**Keywords**— *Bandpass filters, Band Subtraction, Cavity Filter, Feature Extraction, Filter Tuning, Hyperspectral Imaging, Microwave filter, Monochrome camera, Screw Detection.*

## I. INTRODUCTION

Microwave (MW) cavity filters are used in Radio Base Stations (RBS) for separating the desired frequencies from tensed communication spectrum. To compensate for the mistakes like manufacturing defects, design errors, variations in material properties, mechanical tolerances etc., the assembled filters require tuning. Mechanical tuning, that uses tuning screws, is the most common way to tune the cavity filters. The frequency response of the filter is determined by the depth at which the tuning elements are inserted within the cavity. Now since, the filter tuning process is stochastic in nature, it is time consuming, laborious and requires skilled technicians to tune the filter to the desired frequency range [1].

A cavity filter is usually made up of a metallic block with screws on its top plate. Fig. 1 shows a commonly used cavity filter. One can observe the presence of several screws on the assembled filter. Among these, the ‘mounting screws’ serve the purpose of holding the top metallic plate over the whole structure. The remaining screws are the ‘tuning screws’ which are used to alter the performance of the filter. The present research focuses on the filter type shown in Fig. 1



Fig. 1. A commercial MW filter

Fig. 2 presents the magnified portion of a small region of the cavity filter presented in Fig. 1. Noticeably the screws shown in Fig. 2 have different shapes (a mounting screw on bottom left corner and a tuning screw on the top right corner).



Fig. 2. Different types of screws of a cavity filter

The screws shown in Fig. 2 have different shapes and hence image processing techniques like shape detection/pattern matching/contour matching etc. could be used to differentiate them. However, the difference between the shapes of tuning screws and mounting screws is not guaranteed. Rather, commercial filters sometimes have same screw head for all the screws (tuning screws as well as mounting screws). Hence, a robust technique is needed to identify and classify the screws present on the filter structure.

This research work presents a novel band-subtraction technique for distinguishing the tuning screws and mounting

screws of a MW cavity filter. The information about relevant bands was extracted using the datacube of a hyperspectral camera. The suggested methodology can assist in discriminating the screws on the basis of their material composition. The methodology plays crucial role for the case when no technical information about the filter and its screws is available. An in-depth analysis of the existing research reveals that no such methodology has been used or presented in the literature yet.

The remainder of the paper is structured in five sections. Section II presents the literature review of various screw detection techniques; Section III presents hyperspectral imaging (HSI) along with its application in metal detection. In section IV authors discuss the novel band-subtraction methodology for differentiating the screws. Experimentation results and their analysis is discussed in Section V. Section VI concludes the paper and provides future recommendations.

## II. LITERATURE REVIEW: SCREW DETECTION

In [2], authors used Canny operator-based edge detection technique to extract the contours and then template matching step was used to classify the screws. [3] presents a multi-template matching algorithm for the detection of screws and their semi-autonomous removal from the ceiling panel. The technique didn't hold firm grounds because of the following reasons – any change in color or illumination yielded inaccurate results; there was always a reliance on a fixed template and; every application would require a new template which is both tedious and time consuming. Hence, the major drawback of this technique was its lack of generalizability.

A combination of grayscale, color depth and HSV characteristics were used to achieve high accuracy in screw detection [4]. The algorithm was invariant to scale, translation and rotation but relied on Harris corner detection and HSV analysis, both getting influenced by the lighting conditions. Also, RGB depth sensor (Kinect) was needed to remove holes which demanded extra computations.

In 2018, a screw detection technique utilizing Hough circle detection was introduced [5]. However, its effectiveness was limited when dealing with multiple circular components. The commercial application of this method was further constrained due to the requirement of adjusting multiple parameters such as the camera's brightness settings. In [6], a fusion technique which combined the features extracted from Hough transform, and from Deep Learning (DL)-based classifier was presented. The work in [6] was then extended where the algorithm could additionally do the classification of 12 different types of screwheads [7]. While this work successfully eliminated the need for a depth sensor, the detection setup requirements were unsuitable for the production lines.

Some researchers used CNN-based techniques for screw detection applications. A combination of Faster R-CNN and RES (Rotation Edge Similarity) was used for classifying the screws [8] but the technique's commercialization was hindered by its paltry computational speed. Another Faster-RCNN-based model used general screw features for detecting the screws [9] using a DSLR camera. The authors combined image pre-processing and object detection steps with visual reasoning to

achieve accurate results. The model's performance was enhanced by retraining it with true-negative results. Nevertheless, the image processing steps executed in this study had a substantial impact on the proposed model's performance.

The work presented in [10] is closely related to the one discussed in the current research. The authors of [10] used basic image processing techniques for detecting the tuning screws on the basis of its geometry. However, their work was reliant on a particular type of screw.

The dependencies of various screw detection techniques discussed above are listed in Table I. Each technique is constrained by one or more of these specific issues.

TABLE I. DEPENDENCIES OF SCREW DETECTION TECHNIQUES AVAILABLE IN LITERATURE

Specification	Related Example
Device Specific	Electric motor screws, battery screws, etc.
Screw Specific	Shape and/or size of its head
Environment Specific	Illumination state, shadows, shiny objects
Methodology Specific	Stickers or other round objects detected as screws, damaged screws are not detected etc.

From Table I, it can be inferred that all the aforementioned methods of detecting the screws lack generalizability. Therefore, a more robust screw detection technique is needed to differentiate between the tuning and mounting screws.

Since the tuning screws are plated with silver and mounting screws are made of steel alloy, analyzing their reflectance characteristics by HSI can help in differentiating them. To date, this approach of tuning screw detection has not been reported in scholarly research. In the next section authors discuss research in the field of HSI and its application in detecting the coatings, metals and compounds.

## III. HYPERSPECTRAL IMAGING-BASED DETECTION

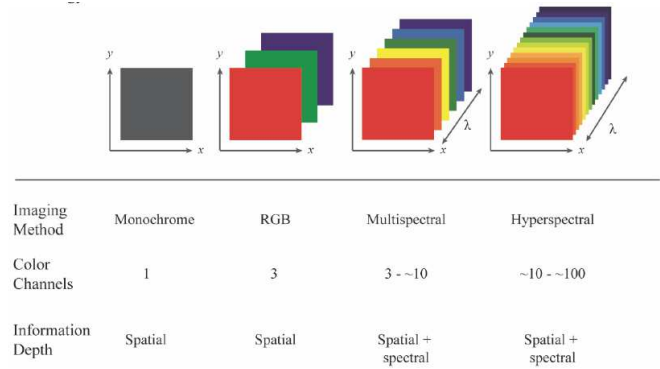


Fig. 3. Different image methods and their characteristics [11]

Fig. 3 displays various imaging techniques and their features. A monochrome camera considers the whole visible spectrum and only measures its integral intensity. Only the visible portion of EM spectrum (400-700 nm) is covered by RGB, which primarily contains spatial information. The RGB band is not appropriate for all applications, hence multispectral or hyperspectral cameras are employed instead. Multispectral



images are a collection of images captured in a continuous spectrum at several wavelength bands that form a datacube. Datacube is a 3D representation of spatial data ( $x$  and  $y$  coordinates) and spectral data (in the  $z$ -axis). A hyperspectral image is a multispectral image with hundreds of bands [12] that contains the entire wavelength spectrum for each pixel. Additionally, the wavelength range of hyperspectral images encompasses the ultraviolet to mid-infrared spectrum in addition to visible range. HSI combines spectroscopy and digital imaging. With this imaging technique, the scene is captured in narrow bandwidths, and necessary image bands can be selected from the datacube for further processing [13].

The reflectance characteristics of different materials produce a unique signature at different wavelength bands. Considering the scope of the current research work, the reflectance behavior of Silver (Ag) and Carbon Steel from 0.2  $\mu\text{m}$  (200 nm) to 20  $\mu\text{m}$  (20000 nm) wavelength can be seen in Fig. 4. Significant variations between the reflectance curves for two metals are visible. On the basis of this spectral signature, similar materials can be identified.

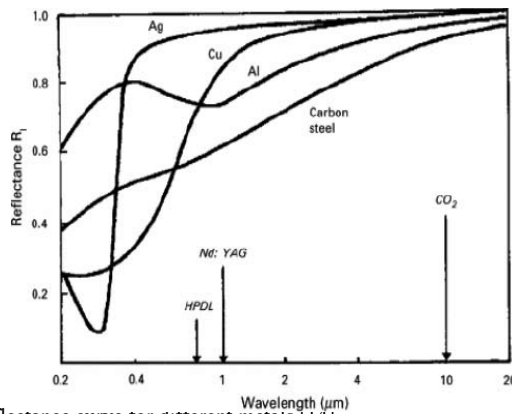


Fig. 4. Reflectance curve for different metals [14]

Since HSI contains hundreds of narrow bandwidth spectral bands, the majority of them are correlated and provide redundant information. Processing a large amount of data decreases the computational efficiency [15] due to the Hughes phenomenon [16]. Hence, it makes sense to choose the spectral bands that provide distinct characteristic information [17]. The main advantages of utilizing a band selection technique are the increase in classification accuracy [18] while preserving the intrinsic information of the original pixel [19], in addition to increase in computation efficiency.

In related literature, band selection technique has been used in determining the aluminium oxide thickness [20]. Several research groups have employed the HSI technique for detection of corrosion. Using HSI, corrosion on carbon steel samples [21][22], mild steel used in the aeronautical industry [23], copper [24] etc. has been detected. Researchers have also provided an SVM classifier-based Metal Object Detection (MOD) approach for identifying ferromagnetic, non-ferromagnetic, and non-metallic items [25].

A hypothesis that the tuning and mounting screws of a cavity filter can be differentiated using HSI was developed after reviewing the research into HSI-based methods for detecting the metals, metal coatings, and alloys. Since mounting screws

are typically made of an alloy of steel and tuning screws used in industry are usually coated with a 3  $\mu\text{m}$  silver layer (to increase the conductivity), their reflectance signature can be utilized to differentiate them from each other.

#### IV. METHODOLOGY

##### A. Imaging Setup

The initial attempts to detect the tuning screws were made using a Specim IQ Mobile Hyperspectral Camera [26]. In total, this hyperspectral camera has 204 image bands (for each pixel), a spectral resolution of 7 nm, and operates in the 400–1000 nm wavelength range. The reflectance values for all the bands can be displayed in the spectral distribution of each pixel. Using a calibrated tile with 99% reflectance, this camera is calibrated for white reference. After initial calibration, no further image processing steps are performed on HSI images. However, unlike RGB, HSI is unable to detect the geometry of the objects. Instead, the required bands are selected and forwarded to be processed further.

The overall configuration utilized to capture the image from the hyperspectral camera is shown in Fig. 5. A reference plate used for calibrating the Specim IQ camera can also be seen in Fig. 5. Two 400W halogen projectors were used for lighting, along with a light diffusing sheet. The halogen projectors were chosen as the light source since they cover a wider spectrum of wavelengths. The wavelength covered by halogen light spans from the UV region to the IR region in the EM spectrum. To ensure homogeneous lighting, the diffusing sheet was used.

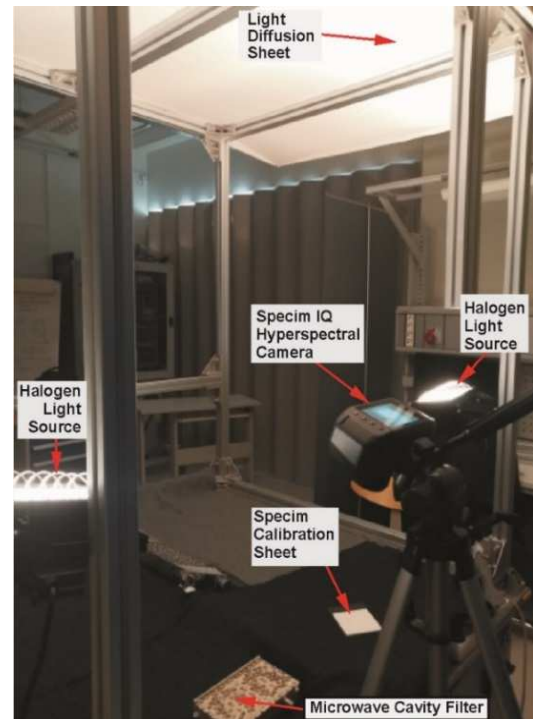


Fig. 5. Image acquisition setup

Even though a built-in RGB camera exists in the Specim IQ hyperspectral camera, the RGB image it produces has a different spatial resolution. Additionally, there can be misalignments between the RGB image and hyperspectral image in the vertical and/or horizontal direction. The RGB

scene of an image taken with a hyperspectral camera is displayed in Fig. 6, but only the marked area was used for HSI processing, as is the case with all only the marked area was used as shown in all other images in the following sections of this paper.

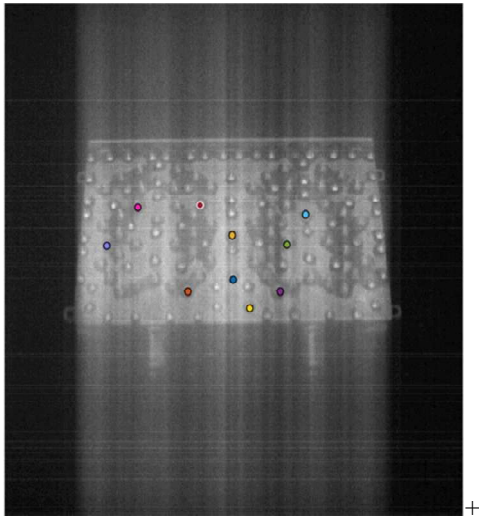


Fig. 6. RGB scene captured by a hyperspectral camera

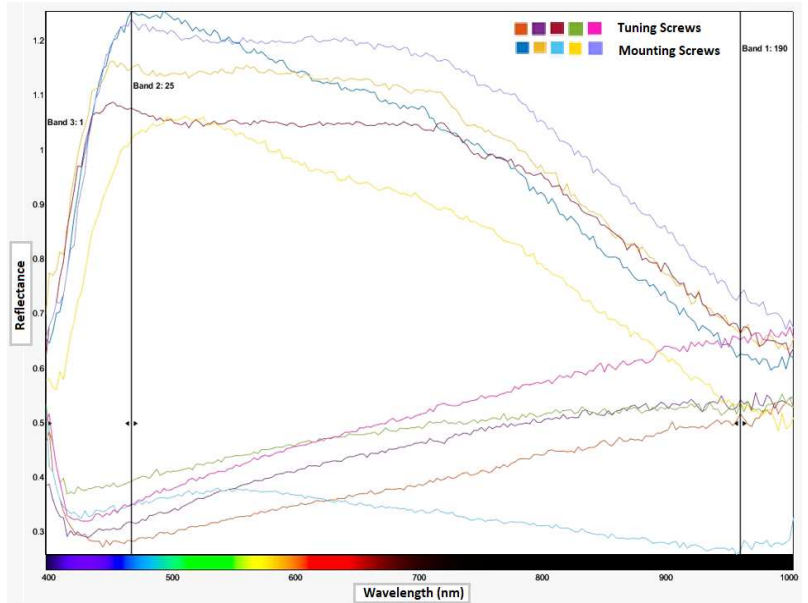
### B. Reflectance Trends

Tuning screws and mounting screws of a MW filter have different material composition. Since each material has its unique spectral characteristics, different band alternatives have to be tested in order to identify the effective spectral image bands to distinguish between the two types of screw.

Five screws from each category (05 tuning screws and 05 mounting screws) were chosen in order to retrieve the reflectance data from the screws that were installed on the filter under consideration. The purpose of considering many screws from each category was to precisely choose the appropriate bands using the mean reflectance value. Fig. 7a shows the locations of selected screws in one of the bands. The corresponding reflectance plots for all these screws are presented in Fig. 7b for the wavelength range of 400 nm to 1000 nm. The choice of bands was then made using the average reflectance value, as is covered in the next subsection.



7 (a): Screw Locations on a Band Image



7 (b): Reflectance-Wavelength Plots

Fig. 7. Sampled 05 screws from each screw category and their reflectance response

### C. Selection of Bands

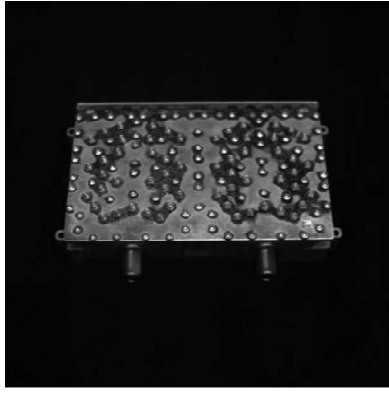
After averaging the reflectance values of the chosen tuning and mounting screws were averaged, band 25 and band 190 were selected from the datacube. Band 25 has been chosen because it has the most distinguishing features for both types of screws (see a significant variation in reflectance between the two categories of screws in Fig. 7b). Band 25 has an approximate wavelength of 467 nm. Band 190, having wavelength of roughly 930 nm, was chosen from the opposite end of the datacube because it exhibits some similarities in features between the two screws (see Fig. 7b for a modest variation in reflectance characteristics in this band). It is significant to note that the authors chose band number 190 even though the difference between the two curves was minimal around band 197. This band was chosen in order to get rid of any spectral noise that might have been present in the last few bands near the end of the image. Images of bands 25 and 190 are shown in Figs. 8a and 8b, respectively.

### D. Band Subtraction

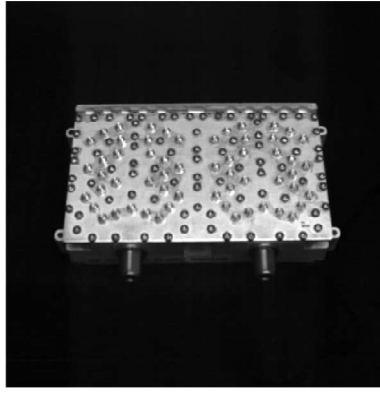
Empirically, it was found that subtracting one band from the blue region (band number 25) and another band from the infrared (IR) region (band number 190) of the datacube could clearly differentiate tuning screws and mounting screws. As presented by (1),  $I_{result}$  image was obtained when a 467 nm image ( $I_{467}$ ) was subtracted from a 930 nm image ( $I_{930}$ ).

$$I_{result} = I_{930} - I_{467} \quad (1)$$

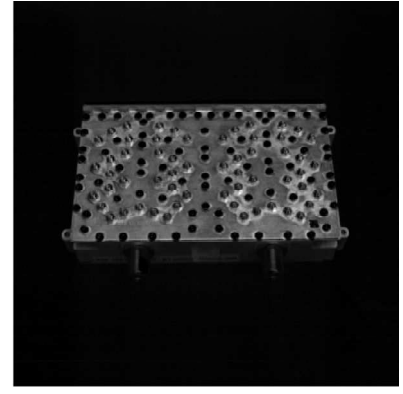
Figure 8c, which shows the image produced adhering to band subtraction, makes it evident that the mounting screws looked noticeably darker when compared to the silver-plated tuning screws. With the aid of this knowledge, one can determine the coordinates for each tuning screw's location and tune a filter autonomously.



8 (a). Band 25 Image



8 (b). Band 190 Image



8 (c). Resultant Image after Band Subtraction

Fig. 8. Band selection and band subtraction results

## V. EXPERIMENTATION RESULTS AND ANALYSIS

### A. Specim Hyperspectral Camera

Fig. 9 displays the outcome of the band subtraction method applied to the image captured with the Specim hyperspectral camera. Both the RGB and the processed image, which is the result of band subtraction, are displayed next to each other for convenience.

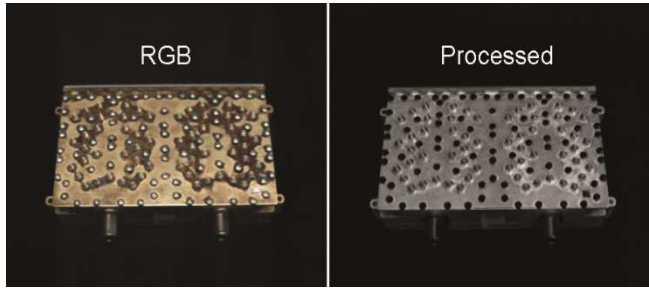


Fig. 9. Classification of screws using Specim hyperspectral camera

Fig. 9 demonstrates that using the suggested methods led to encouraging findings. The mounting screws stand out from the tuning screws owing to their darker appearance. As detailed in the following subsection, the efficiency of the postulated band subtraction methodology was then assessed on an industrial monochrome camera.

### B. Basler Monochrome Camera

The analysis done on the hyperspectral images showed that the most effective spectral bands for the application considered in this work are band 25 and band 190, and they must be subtracted. This methodology was tested on the images acquired by an industrial monochrome camera.

To apply the suggested band subtraction process, the Basler camera with a progressive scan CCD-sensor for capturing VGA-640 x 480 images was used. The camera's connection with a compatible PC was made via an IEEE 1394 firewire interface. Two optical bandpass filters i.e., blue bandpass filter (associated to band 25) and IR bandpass filter (corresponding to band 190), were attached to the camera. Fig. 10 shows a Basler monochrome camera used in this work on which the light (optical) filters were mounted.



Fig. 10. Basler monochrome camera mounted with optical bandpass filters

It is evident from the findings displayed in Fig. 11 that the classification accuracy of 100% was reached by using the light bandpass filters (decided by our suggested methods) with a monochrome camera. As shown in Fig. 11, the mounting screws appear to be darker than the tuning screws.

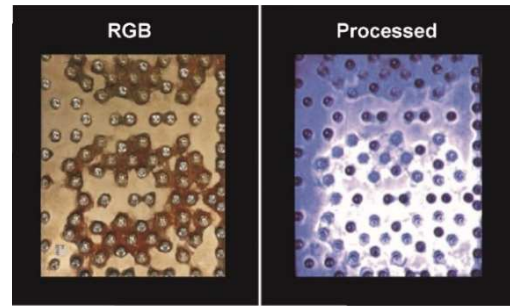


Fig 11. The results screws classified by the Basler monochrome camera

A monochrome camera, which is significantly less expensive and computationally efficient than a hyperspectral camera, can attain the same classification accuracy as a hyperspectral camera, as shown in Figures 10 and 11. The technique of automating cavity filter tuning can benefit from this study's conclusions. Once the tuning screws have been differentiated, a robotic manipulator can be instructed to tune a filter based on their spatial coordinates.



## VI. CONCLUSION

In this research, the tuning screws, that are used to tune a MW cavity filter, were distinguished using a novel band selection method. The analysis made on the images acquired from a hyperspectral camera helped in selecting the most efficient spectral bands from the hyperspectral datacube. Empirically, it was found that the suggested band subtraction methodology could distinguish between the screws according to their material composition. The process was subsequently tested on an industrial monochrome camera fitted with the appropriate optical bandpass filters. The conclusions were validated by the results, which showed that a monochrome camera—which is less expensive and computationally more effective than a hyperspectral camera—was able to detect and classify screws with 100% accuracy. Based on a comprehensive analysis, it can be concluded that the methodology utilized in this study is highly effective in distinguishing and categorizing various types of screws. The proposed methodology plays a pivotal role for the case when the technical drawings of the filter is not available. The camera can be set in perspective view or can be mounted overhead. The position coordinates of tuning screws thus determined can be utilized in a FAT system for MW filters.

## REFERENCES

- [1] R. V. Snyder, "Practical aspects of microwave filter development," *IEEE Microwave Magazine* 8, no. April, pp. 42–54, 2007.
- [2] P. Gil, J. Pomares, S. V. T. Puente, C. Diaz, F. Candelas, and F. Torres, "Flexible multi-sensorial system for automatic disassembly using cooperative robots," *Int J Comput Integr Manuf*, vol. 20, no. 8, pp. 757–772, Dec. 2007, doi: 10.1080/09511920601143169.
- [3] S. R. Cruz-Ramirez, Y. Mae, T. Takubo, and T. Arai, "Detection of screws on metal-ceiling structures for dismantling tasks in buildings," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, Nice - France*, 2008.
- [4] M. Bdiwi, A. Rashid, and M. Putz, "Autonomous disassembly of electric vehicle motors based on robot cognition," in *IEEE International Conference on Robotics and Automation, Stockholm - Sweden*, 2016.
- [5] N. M. Difilippo and M. K. Jouaneh, "A system Combining Force and Vision Sensing for Automated Screw Removal on Laptops," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 2, pp. 887–895, Apr. 2018, doi: 10.1109/TASE.2017.2679720.
- [6] E. Yildiz and F. Worgötter, "DCNN-Based screw detection for automated disassembly processes," in *Proceedings - 15th International Conference on Signal Image Technology and Internet Based Systems, SISITS 2019*, Institute of Electrical and Electronics Engineers Inc., Nov. 2019, pp. 187–192, doi: 10.1109/SITIS.2019.00040.
- [7] E. Yildiz and F. Wörgötter, "DCNN-based screw classification in automated disassembly processes," in *ROBOVIS 2020 - Proceedings of the International Conference on Robotics, Computer Vision and Intelligent Systems*, SciTePress, 2020, pp. 61–68, doi: 10.5220/0009979900610068.
- [8] X. Li *et al.*, "Accurate screw detection method based on faster R-CNN and rotation edge similarity for automatic screw disassembly," *Int J Comput Integr Manuf*, vol. 34, no. 11, pp. 1177–1195, 2021, doi: 10.1080/0951192X.2021.1963476.
- [9] G. Foo, S. Kara, and M. Pagnucco, "Screw detection for disassembly of electronic waste using reasoning and re-training of a deep learning model," in *Procedia CIRP*, Elsevier B.V., 2021, pp. 666–671, doi: 10.1016/j.procir.2021.01.172.
- [10] C. Yao, Y. Yuan, J. Li, and L. Bi, "High Precision Tuning Device of Microwave Cavity Filter based on Hand-Eye Coordination," *Chinese Control Conference, CCC*, vol. 2019-July, pp. 7063–7068, 2019, doi: 10.23919/ChiCC.2019.8866244.
- [11] D. C. Liyanage, M. Tamre, and R. Hudjakov, "Hyperspectral/Multispectral Imaging Methods for Quality Control," in *Handbook of Research on New Investigations in Artificial Life, AI, and Machine Learning*, M. K. Habib, Ed., Hershey, PA, USA: IGI Global, 2022, pp. 438–461, doi: 10.4018/978-1-7998-8686-0.ch017.
- [12] J. M. Amigo, "Hyperspectral and multispectral imaging: setting the scene," in *Data Handling in Science and Technology*, Elsevier Ltd, 2020, pp. 3–16, doi: 10.1016/B978-0-444-63977-6.00001-8.
- [13] S. Jia, L. Shen, J. Zhu, and Q. Li, "A 3-D gabor phase-based coding and matching framework for hyperspectral imagery classification," *IEEE Trans Cybern*, vol. 48, no. 4, pp. 1176–1188, Apr. 2018, doi: 10.1109/TCYB.2017.2682846.
- [14] M. M. Quazi, M. A. Fazal, A. S. M. A. Haseeb, F. Yusof, H. H. Masjuki, and A. Arslan, "Laser-based surface modifications of aluminum and its alloys," *Critical Reviews in Solid State and Materials Sciences*, vol. 41, no. 2, pp. 106–131, Mar. 2016, doi: 10.1080/10408436.2015.1076716.
- [15] F. Luo, B. Du, L. Zhang, L. Zhang, and D. Tao, "Feature learning using spatial-spectral hypergraph discriminant analysis for hyperspectral image," *IEEE Trans Cybern*, vol. 49, no. 7, pp. 2406–2419, Jul. 2019, doi: 10.1109/TCYB.2018.2810806.
- [16] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans Inf Theory*, vol. IT-14, no. 1, pp. 55–63, 1968.
- [17] W. Sun and Q. Du, "Hyperspectral band selection: A review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 2, Institute of Electrical and Electronics Engineers Inc., pp. 118–139, Jun. 01, 2019, doi: 10.1109/MGRS.2019.2911100.
- [18] Q. Wang, F. Zhang, and X. Li, "Optimal clustering framework for hyperspectral band selection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 10, pp. 5910–5922, Oct. 2018, doi: 10.1109/TGRS.2018.2828161.
- [19] Y. Zhan, D. Hu, H. Xing, and X. Yu, "Hyperspectral band selection based on deep convolutional neural network and distance density," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 12, pp. 2365–2369, Dec. 2017, doi: 10.1109/LGRS.2017.2765339.
- [20] F. Gruber, P. Wollmann, B. Schumm, W. Grähler, and S. Kaskel, "Quality control of slot-die coated aluminum oxide layers for battery applications using hyperspectral imaging," *J Imaging*, vol. 2, no. 2, Jun. 2016, doi: 10.3390/jimaging2020012.
- [21] T. De Kerf, G. Pipintakos, Z. Zahiri, S. Vanlanduit, and P. Scheunders, "Identification of corrosion minerals using shortwave infrared hyperspectral imaging," *Sensors*, vol. 22, no. 1, Jan. 2022, doi: 10.3390/s22010407.
- [22] T. De Kerf, A. Gestels, K. Janssens, P. Scheunders, G. Steenackers, and S. Vanlanduit, "Quantitative detection of corrosion minerals in carbon steel using shortwave infrared hyperspectral imaging," *RSC Adv*, vol. 12, no. 50, pp. 32775–32783, Nov. 2022, doi: 10.1039/d2ra05267a.
- [23] M. M. Antony, C. S. S. Sandeep, and M. V. Matham, "High resolution probe for corrosion monitoring using hyper spectral imaging," in *AIP Conference Proceedings*, American Institute of Physics Inc., Feb. 2021, doi: 10.1063/5.0036100.
- [24] M. Al Ktash *et al.*, "UV hyperspectral imaging as process analytical tool for the characterization of oxide layers and copper states on direct bonded copper," *Sensors*, vol. 21, no. 21, Nov. 2021, doi: 10.3390/s21217332.
- [25] Y. Tian *et al.*, "Metal object detection for electric vehicle inductive power transfer systems based on hyperspectral imaging," *Measurement (Lond)*, vol. 168, Jan. 2021, doi: 10.1016/j.measurement.2020.108493.
- [26] Specim Spectral Imaging Ltd, "Specim IQ Technical Specifications," Apr. 14, 2019. <https://www.specim.com/iq/tech-specs/> (accessed Apr. 10, 2023).

# Optimal Harmonic Current Extractor using Digital Warped Filter for a Single –Phase PV Integrated Grid-Tied System with 5-Level DSTATCOM

Praveen Bansal

Department of Electrical Engineering  
Madhav Institute of Technology & Science  
Gola ka Mandir-474005  
pbansal444@mitsgwalior.in

Alka Singh

Department of Electrical Engineering  
Delhi Technological University  
Shahabad Road, Rohini-110042  
alkasingh@dce.ac.in

**Abstract** – A new control method for PV integrated single-phase grid-connected 5-level distribution static compensator (DSTATCOM) system is presented and investigated in this paper. It uses a warped digital filter control technique to estimate the fundamental component of load current from the non-linear load. DSTATCOM has been developed using a cascaded multilevel inverter and is used to improve power quality. The proposed system is single-stage, single-phase grid-tied and can operate in two modes viz. During the night, acting as a DSTATCOM unit and performing the necessary shunt compensation while providing additional power to the grid during the day. A conventional Proportional and Integral (PI) controller has been used to regulate the DC link voltages of both capacitors during the sudden variation in solar irradiance and load. Phase shifted PWM scheme has been implemented to generate firing pulses of DSTATCOM. The system is tested under normal and distorted grid conditions. Simulation and experimental results are presented and obtained within stipulated IEEE-519 and IEEE-1547 standards.

**Index Terms** – SAPF, MLI, Non-Linear Load, PWM

## I. INTRODUCTION

Power quality has always been a major concern in the operation of the distribution system, and numerous initiatives have been put in place to address this problem [1]. Many advantages come with better power quality, such as increased loading capacity, optimal use of various electrical devices, zero voltage regulation, etc. The distribution system, including its subsystems, and the consumer loads are the two main broad categories on which the sources of poor power quality. The main culprit is the extensive use of nonlinear loads [2]. The problems with the power quality of the electrical distribution system are gradually getting worse nowadays as a result of the steadily growing use of these nonlinear loads in distribution systems, such as rectifiers, computers, and switched mode power supplies (SMPS). Furthermore, inadequate grid conditions found in the electric distribution system in underdeveloped countries exacerbate the power quality issues [3]. These power quality issues are mostly caused by voltage and current harmonics that are generated within the system. Voltage distortion, nonlinear loads, unbalanced loads, voltage sag/swell circumstances, adding or removing loads, and nonlinear loads, cause power quality issues of concern.

For assessing and adhering to the appropriate degree of power quality, international standards have also been developed, such as the IEEE-519 and IEEE-1547 standards [3-4].

Power electronics-based shunt custom power devices are mostly used to strengthen the reliability of the distribution networks by reducing the injected harmonics and enhancing power quality. Conventional 2-level multilevel inverters suffer from higher switching losses especially in medium and high voltage systems and have high PIV rating of switches and the switches also suffer from high dv/dt stress. Therefore, nowadays MLI has gained keen interest though it was introduced lately. These converters are widely used for medium and high power distribution system [5] because of several advantages viz. the capability to handle more power with reduced PIV rating of switches and low stress; reduced the size of filter; reduced THD in output voltage and current. Due to numerous advantage of MLI over conventional 2-level inverter, a 5-level cascaded multilevel inverter (CHB-MLI) has been implemented as DSTATCOM in the proposed system and controlled using digital warped filter technique. The proposed system is capable of balancing reactive power burden, controlling the voltage at PCC, and enhancing power quality.

Many control strategies, including d-q based Reference Frame Theory, Instantaneous Reactive Power Theory (IRPT), Power Balance Theory, and currently controllers employing advanced neural networks and adaptive filters have been utilised for shunt power compensation [6-7]. The neural network still faces difficulties with imbalanced datasets and saturation issues and the large amount of training data it needs to operate well. Numerous adaptive filter control algorithms have been implemented [8-9] to control DSTATCOM for single phase single stage PV integrated grid connected system but these algorithms are suffer from weak convergence and high computational burden to be implemented properly using digital signal processor. In this research work, reference grid currents and CHB-MLI gating pulses are generated using warped filter which is further used to produce switching pulses for 5-level MLI. Thus with the help of this filter, DSTATCOM is operated efficiently, enhancing the system's power quality. Warped filters are frequently employed in a wide range of audio applications, including linear prediction, echo

audio applications, including linear prediction, echo cancellation, and the identification of band pass signals in broad band transmissions, and others. [10]. Warped filters can be configured using changeable low-pass, high-pass, band pass, or band stop responses according to the required application.

The major contribution of work is as follows:

1. Developed 5-level closed-loop control operation of DSTATCOM designed using a cascaded multilevel inverter.
2. Estimation of fundamental component of current using designed warped digital filter
3. Testing of proposed system in simulation and also testing experimentally using scaled down prototype model developed in the laboratory

## II. SYSTEM CONFIGURATION

The schematic diagram of the single phase 5-level CHB-MLI SAPF coupled to the non-linear loads and single-phase grid AC mains is shown in Fig. 1. Voltage at the point of common coupling (PCC), load current ( $i_L$ ), source current ( $i_s$ ), and dc bus voltages  $V_{DC1}$ ,  $V_{DC2}$  are the sensing input variables used to regulate the SAPF. To reduce ac output ripples, a connecting or interface inductor ( $L_f$ ) is used in the CHB-MLI. In order to reduce the harmonics produced by the load current, the SAPF unit is current-controlled using the warped filter to inject suitable compensation current in phase opposition. Both of the dc-link voltages must be kept constant and well regulated for the CHB-MLI to function properly, and the standard PI controller performs this duty. Using MATLAB/SIMULINK, the system is simulated and prototype model is developed in the laboratory. In the proposed single phase single stage PV array grid tied system, 1kW each capacity of PV array is

connected to DC link side of 5-level CHB-MLI used as DSTATCOM unit. To suppress the current harmonics generated by the PV arrays connected on DC side of CHB-MLI an interfacing inductor is used. The system operates in two modes viz day (mode-1) and night (mode-2). During day the PV arrays supplied active power to the grid and during night the system acting as SAPF unit and do harmonic compensation. A single phase grid source supplies power to non-linear load and to fulfil power required to charge the DC link capacitors during mode-2. The control block diagram of proposed system is shown in Fig. 2. The system design and related calculations are discussed.

### A. Designing of DC link voltages

A single phase, 110V (AC rms) supply is connected at PCC. The reference DC link voltage required is calculated by using given relation [14]

$$E_{DC} = \sum_{j=1}^N E_{DC,j} = \frac{\sqrt{2} \times V_g}{m_i} \quad (1)$$

where  $m_i$  is the modulation index, grid voltage is  $V_g$ , for  $j^{\text{th}}$  H-bridges are connected in cascaded fashion. The calculated DC link reference voltage is 172.82V and in the simulation, it is approximated as 200V.

### B. Design of Interfacing Inductor

The value of interfacing inductor is calculated as

$$L_{inf} = \frac{E_{o,rms}}{8 \times g \times f_r \times \Delta I_g} \quad (2)$$

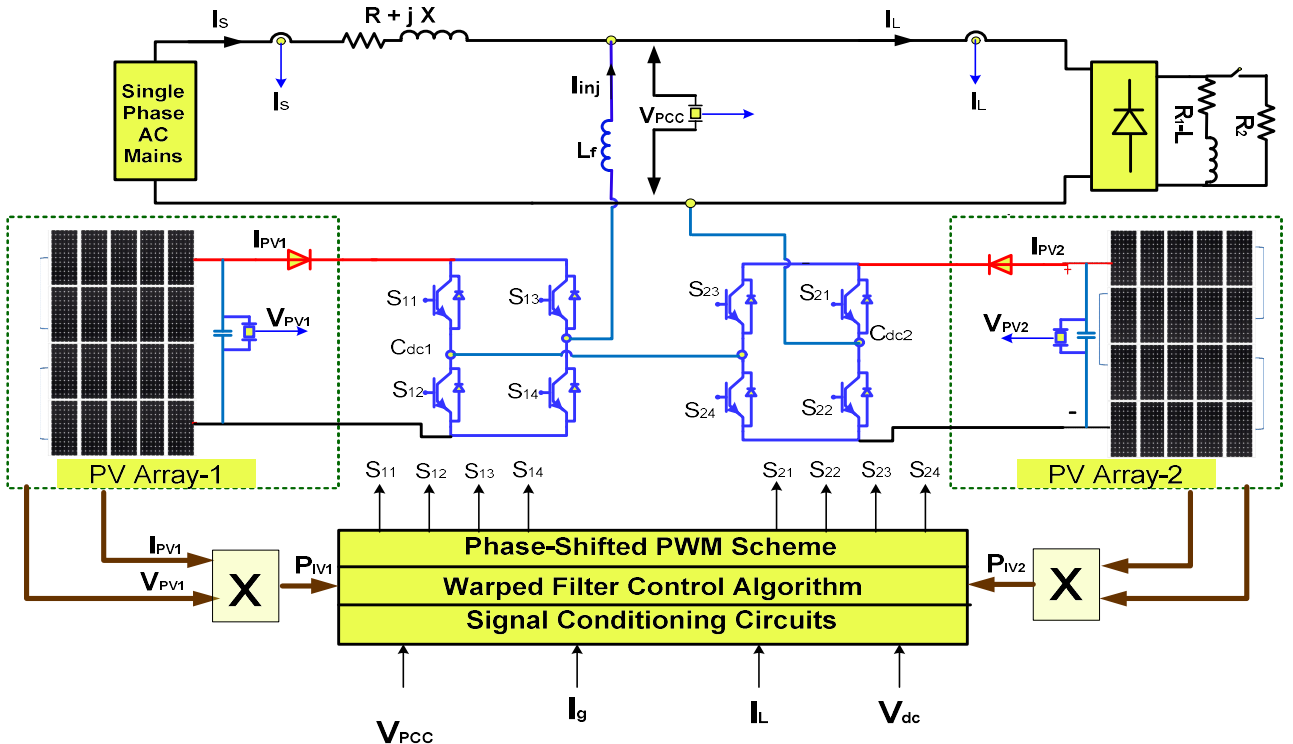


Fig. 1 Schematic diagram of the proposed system

$$I_g = \frac{P_{PV(\max)}}{V_g} = \frac{2000}{110} = 18.18 A \quad (3)$$

In the proposed system, there are two DC link capacitors whose voltages are regulated to 100V each so as to equally share the total DC link voltage of 200V. The DC link capacitance value can be calculated as

where for  $j^{\text{th}}$  PV array the DC power is  $P_{\text{DC}j}$  and DC link voltage is  $E_{\text{DC}j}$ ,  $\omega_r$  is the angular frequency and the DC voltage ripple is  $E_{\text{DC-ripple}}$  and it is considered as 5% of each DC link voltage

The overall implementation of Warped filter is depicted in Fig. 2. The fundamental estimated component of load current ( $I_f$ ) is extradited from this filter as shown in this figure. The overall controller design involves various calculations to achieve stable closed loop operation. These control functions include unit vector generation, DC link voltage control under varying load conditions, extraction of real component, determining reference current and finally the generation of PWM pulses for firing of insulated Gate bipolar junction transistors (IGBT).

$$\frac{i_{L1}}{i_{L\beta}} = \frac{\gamma_1 + \beta_0 z^{-1}}{1 - \gamma_1 z^{-1}} \quad (5)$$

The diagram illustrates the control system for a VSC-based HVDC system. It is divided into two main sections: the power stage (top) and the control system (bottom).

**Power Stage (Top):**

- DC Link:** The DC voltage  $e_{DC}$  is filtered by a Low Pass Filter (LPF) and compared with a reference voltage  $E_{ref}$  to generate a voltage error signal.
- PI Controller:** The voltage error signal is processed by a PI controller to produce a reference current  $I_{Loss}$ .
- Sample and Hold Circuit:** The reference current  $I_{Loss}$  is sampled and held to provide a reference for the current estimator.
- Current Estimation:** The reference current  $I_{Loss}$  is compared with the actual DC link current  $i_L$  to produce an error signal  $I_{est}$ . This error signal is used to estimate the DC link current  $i_L$  using a current estimator.
- Current Estimator:** The current estimator uses the reference current  $I_{Loss}$  and the actual DC link current  $i_L$  to produce the estimated DC link current  $i_L$ .
- Current Estimation Error:** The estimated DC link current  $i_L$  is compared with the actual DC link current  $i_L$  to produce a current estimation error  $I_{est}$ .
- Current Estimation Error:** The current estimation error  $I_{est}$  is used to generate a reference current  $I_{ref}$  for the VSC bridge.
- VSC Bridge:** The reference current  $I_{ref}$  is used to generate the reference current  $I_{ref}$  for the VSC bridge.
- VSC Bridge:** The reference current  $I_{ref}$  is used to generate the reference current  $I_{ref}$  for the VSC bridge.

**Control System (Bottom):**

- Current Estimation:** The current estimation system uses the reference current  $I_{ref}$  and the actual DC link current  $i_L$  to produce the estimated DC link current  $i_L$ .
- Current Estimation Error:** The estimated DC link current  $i_L$  is compared with the actual DC link current  $i_L$  to produce a current estimation error  $I_{est}$ .
- Current Estimation Error:** The current estimation error  $I_{est}$  is used to generate a reference current  $I_{ref}$  for the VSC bridge.
- VSC Bridge:** The reference current  $I_{ref}$  is used to generate the reference current  $I_{ref}$  for the VSC bridge.
- VSC Bridge:** The reference current  $I_{ref}$  is used to generate the reference current  $I_{ref}$  for the VSC bridge.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 03, 2023 at 09:48:16 UTC from IEEE Xplore. Restrictions apply.

$$\frac{i_{L2}}{i_{L\beta}} = \left( \frac{\gamma_1 + \beta_0 z^{-1}}{1 - \gamma_1 z^{-1}} \right) (-z^{-1}) - \beta_1 \quad (6)$$

where,  $\beta_0, \beta_1$  represent the filter gains as shown in Fig. 2. The fundamental load component of current ( $i_{Lfa}$ ) can be extracted by implementing the warped filter the output transfer function for  $i_{Lfa}$  and secondary intermediate signal ( $i_L$ ) can be represented as

$$\frac{i_{Lfa}}{i_L} = \frac{i_{Lfa}}{i_{L2}} \times \frac{i_{L2}}{i_L} = \left[ \left( \frac{\gamma_2 + z^{-1}}{1 - \gamma_2 z^{-1}} \right) (-h_1 z^{-1}) \right] \times \left[ \left( \frac{\gamma_1 + \beta_0 z^{-1}}{1 - \gamma_1 z^{-1}} \right) (-z^{-1}) - \beta_1 \right] \quad (7)$$

The overall control loop implementation of digital warped is depicted in Fig. 2. As shown is figure the fundamental estimated component of load current ( $I_{Lfa}$ ) is extradited using warped filter.

The overall controller is designed to perform various calculations to achieve stable closed loop operations such as unit vector generation, DC link voltage control under varying load conditions, Extraction of real component, determining reference current and then finally generation of PWM pulses for firing of insulated Gate bipolar junction transistors (IGBT). The proposed system also employs sample & hold circuit(S&H) and zero current detector circuit (ZCD). The extracted fundamental load current is fed to S&H circuit and it is synchronized with the ZCD, when the unit synchronizing template crosses the zero then ZCD generates the triggering the signal which is further fed to S&H circuit. The S&H logic circuit captures the samples of the sensed load current once the signal is received from ZCD circuit; as a result an accurate and fast estimation of signal is achieved.

### E. Generation of unit vector template

The unit vector template or synchronizing templates are generated from grid voltage  $V_g$ . The grid voltage is passed through a delay of  $90^\circ$  as shown in figure 3. The in-phase component is considered as  $V_{gp}$  and the quadrature component is  $V_{gq}$ , these voltage vectors are further used to generate  $V_t$ . Now the synchronizing template ( $u_p$ ) and quadrature synchronizing template ( $u_q$ ) is calculated as

$$u_p = \frac{V_{gp}}{V_t}; \quad u_q = \frac{V_{gq}}{V_t}; \quad V_t = \sqrt{V_{gp}^2 + V_{gq}^2} \quad (8)$$

### F. DC Loss calculation and generation of reference current

The proposed structure serves two purposes (i) providing active power to the AC grid (ii) compensation of the harmonics generated by the non-linear load so as to achieve a unity power factor operation on supply side. For effective operation of the proposed system under both the control modes, there is a necessity to control the fluctuations in DC link voltages obtained across PV arrays. Therefore a

proportional –integral (PI) controlled is used to control the DC link voltages. The DC link error can be estimated as

$$E_{DCE} = E_{DC-ref} - E_{DC} \quad (9)$$

The error signal is fed to PI controller and  $I_{loss}$  is calculated as shown in fig. 2. Mathematically can be represented as

$$I_{loss}(n+1) = I_{loss}(n) + k_p \{E_{DCE}(n+1) - E_{DCE}(n)\} + k_i E_{DCE}(n+1) \quad (10)$$

where  $k_p$  and  $k_i$  denote the proportional and integral gains respectively and the system dynamic performance of current and voltage can be improved by employing the feed forward term estimated using  $P_{PV}$  as

$$I_{PV} = \frac{2(P_{PV1} + P_{PV2})}{V_t} \quad (11)$$

The reference current is generated by multiplying the unit-synchronizing template with estimated load current, represented as

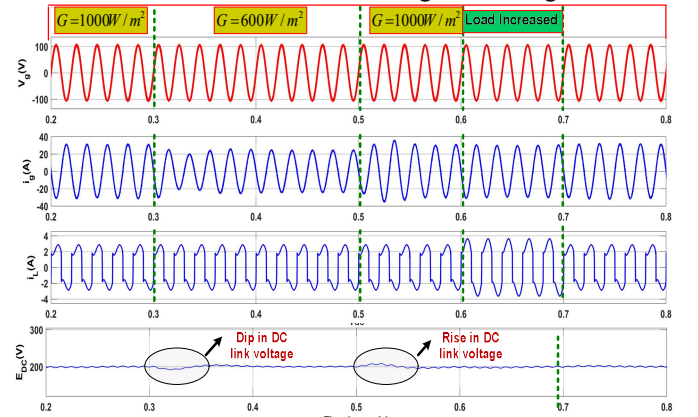
$$i_{gr}^* = u_p I_{est} \quad (12)$$

$$I_{est} = I_{loss} + I_f - I_{PV} \quad (13)$$

The fundamental estimated load current is  $I_{est}$ , fundamental load current is  $I_f$  and PV feed forward current is  $I_{PV}$ . The generated reference current is subtracted from actual grid current and further the signal is compared with phase-shifted PWM technique to generate the firing pulses for 5-Level CHB-MLI.

## III SIMULATION RESULTS AND DISCUSSION

The simulation results of grid voltage  $v_g(V)$ , grid current  $i_g(A)$ , load current  $i_L(A)$  and the total DC link voltage  $V_{dc}(V)$  has been shown in Fig. 3 under varying solar irradiation and dynamic change in load during  $t=0.3s$  to  $0.5s$  and  $t=0.5s$  to  $0.7s$  respectively. PV arrays feed power to the grid and since the PV power extracted is more as compared to the load therefore the grid current is observed to be out of phase with respect to grid voltage as shown in figure. Moreover during transition in solar radiations from  $1000W/m^2$  to  $600W/m^2$  or during load changes, the DC link voltage experiences small variation. However, total DC link voltage is well regulated.



**Fig. 3** Simulation of single-phase grid connected system under varying solar irradiation and load changes

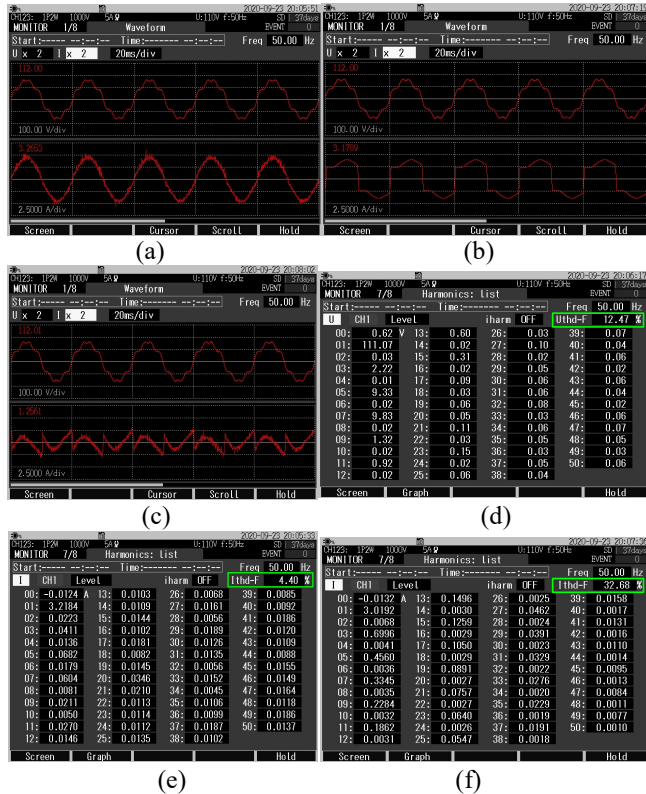


## IV EXPERIMENTAL RESULTS AND DISCUSSION

The proposed model developed in MATLAB is tested experimentally in the laboratory. The hardware system uses two LEM (LA-55P) current sensors to sense the supply current  $i_s(A)$  and load current  $i_L(A)$ . Three voltage sensors LEM (LA-25P) are used to sense the PCC voltage  $V_{pcc}(V)$ , DC link voltage across capacitor-1  $V_{DC1}(V)$  and DC link voltage across capacitor-2  $V_{DC2}(V)$ . The sensed signals are sent to ADC pins of Digital Signal processor and further processed to Warped filter. The reference current is generated and compared with actual source current in order to generate PWM pulses to trigger the IGBTs. Using DAC pins of DSP these generated PWM signals are fed to IGBT driver (SKYPER 32 Pro) and finally passed to eight IGBTs of CHB-MLI for triggering. The signals are recorded using HIOKI Power analyzer and dynamic results are captured using KEYSIGHT digital storage oscilloscope. The experimental results have been recorded without PV integration and discussed next.

### A. Steady State Experimental Results

Fig.4 (a-c) shows the steady state waveforms obtained using warped filter. The steady waveforms show the source voltage  $V_s(V)$  w.r.t to source current  $i_s(A)$ , load current  $i_L(A)$ , compensating current  $i_c(A)$ . Fig. 4(d) shows the experimental results of THD obtained under the distorted source voltage found to be 12.47%. The load current shows THD of 32.68% as depicted in Fig. 4(f) and the source current THD is observed to be 4.40% after compensation as shown in Fig 4 (e). The compensator injects current, which significantly



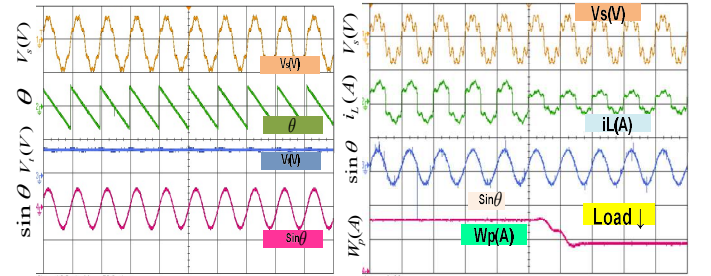
**Fig.4 (a-f) Harmonic spectrum (THD) analysis (a)  $V_s(V)$  with  $i_s(A)$  (b)  $V_s(V)$  with  $i_L(A)$  (c)  $V_s(V)$  with  $i_c(A)$  (d) THD of  $V_s(V)$  (e) THD of  $i_L(A)$  and (f) THD of  $i_s(A)$**

reduces the THD of supply current as per IEEE-519 stipulated standards.

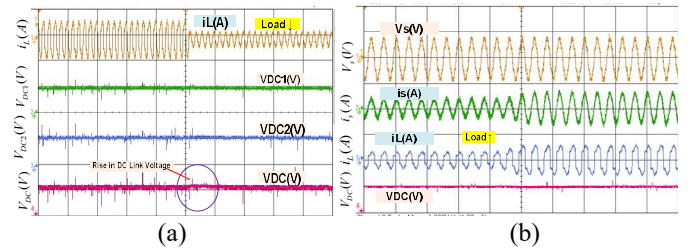
### B. Dynamic Experimental Results

The system is tested under dynamic load variations and well as in distorted voltage grid conditions. The grid voltage distortion of magnitude  $V_s(t) = 110 \sin \omega t + 49.50 \sin 3\omega t + 14 \sin 5\omega t$  is added at  $t=0.7s$  till  $t=0.8s$  as depicted in Fig. 5(a-b). The second order generalized integrator (SOGI) block has been implemented to generate the filtered unit template. The estimated phase angle  $\hat{\theta}$  and amplitude of voltage ( $v_i$ ) and unit sine template under distorted grid conditions are shown in Fig.5a. During distorted grid condition as shown on Fig 5(b), when sudden load is increased then the estimated fundamental weights from warped filter converge quickly within 1~2 cycles and SOGI filter correctly estimate the unit templates.

The experimental closed loop performances of proposed system under distorted grid conditions are presented in Figure 6(a-b). It is seen from the fig .6(a) that the during sudden decrease in the load the both the DC link voltages stabilizes to steady state value within few cycles with minimum variations that means the PI controller quickly response to the proposed system. Likewise, from fig. 6(b) the load current is varied under distorted grid condition.



**Fig. 5 (a) Estimate parameters under distorted grid conditions (b)  $V_s(V)$ ,  $i_L(A)$ ,  $\sin \theta$  and  $W_p(A)$  under dynamic load conditions**



**Fig. 6 Experimental dynamics results of (a)  $i_L(A)$ ,  $V_{DC1}(V)$ ,  $V_{DC2}(V)$ ,  $W_p(A)$  (b)  $V_s(V)$ ,  $i_s(A)$ ,  $i_L(A)$  and  $V_{DC}(V)$**

The system is tested under dynamic load conditions. It is observed from the figures that during sudden load dynamics the DC link voltage quickly stabilizes to steady state value within 1~2 cycles and source current in phase with the load current. The obtained results are satisfactory.



## V CONCLUSION

In the proposed study, warped digital filter has been implemented for a single-phase single-stage PV integrated grid connected system for harmonic elimination and shunt compensation. The 5-level CHB-MLI is used as DSTATCOM unit, which is controlled using the proposed algorithm and serves multiple objectives. Power quality is improved with features such as reactive power compensation, power factor improvement and harmonic reduction. A SOGI based synchronization technique has been used for the generation of unit sine templates under distorted grid conditions. In the laboratory, a prototype model has been developed to test the effectiveness of the proposed configuration. Extensive simulation and experimental results have been demonstrated and THD obtained in utility grid voltage and current is <5% as required by IEEE-519 standard.

## Acknowledgement

The authors are thankful to the Director of **Madhav Institute of Science & Technology, Gwalior** for granting permission to enroll in Ph.D programme and pursue research at Delhi Technological University, Delhi

## REFERENCES

- [1] G. Benysek and M. Pasko, *Power Theories for Improved Power Quality*, Springer-Verlag, London, 2012.
- [2] B. Singh, Chandra, and K. Al-Hadad, 'Power Quality: Problems and Mitigation Techniques', John Wiley & Sons Ltd., U. K., 2015.
- [3] B. Singh, K. Al-Haddad, A. Chandra, "A review of active filters for power quality improvement", IEEE Trans. Ind. Electron., vol. 46, no. 5, pp. 960- 971, 1999.
- [4] Leon, J. I., Sergio Vazquez, L. G., & Franquelo, L. G. (2017). Franquelo "Multilevel converters: Control and modulation techniques for their operation and industrial applications", Proc. IEEE, 105(11), 2066–2081
- [5] A. Makur and S. Mitra, "Warped discrete-Fourier transform: Theory and applications," IEEE Trans. Circuits Syst. I, Fundam. Theory Appl., vol. 48, no. 9, pp. 1086–1093, Sep. 2001.
- [6] H. Akagi, E. H. Watanabe, and M. Aredes, *Instantaneous Power Theory and Applications to Power Conditioning*. Piscataway, NJ: IEEE Press, 2007.
- [7] S. A. Gonzalez, R. Garcia-Retegui, and M. Benedetti, "Harmonic computation technique suitable for active power filters," IEEE Trans. Ind. Electron., vol. 54, no. 5, pp. 2791–2796, Oct. 2007
- [8] M. Badoni, A. Singh and B. Singh, "Comparative Performance of Wiener Filter and Adaptive Least Mean Square-Based Control for Power Quality Improvement," in IEEE Transactions on Industrial Electronics, vol. 63, no. 5, pp. 3028-3037, May 2016, doi: 10.1109/TIE.2016.2515558
- [9] B. Singh and J. Solanki, "An Implementation of an Adaptive Control Algorithm for a Three-Phase Shunt Active Filter," in IEEE Transactions on Industrial Electronics, vol. 56, no. 8, pp. 2811-2820, Aug. 2009, doi: 10.1109/TIE.2009.2014367
- [10] C. Asavathiratham, P. E. Beckmann, and A. V. Oppenheim, "Frequency warping in the design and implementation of fixed-point audio equalizers," in Proc. IEEE Workshop Appl. Signal Process. Audio Acoust., Oct. 1999, pp. 55–58

# Performance Evaluation of Machine Learning Methods for Detecting Credit Card Fraud

Anuj Yadav, Arpajit Adhikary, Aryan Kainth, and Rohit Kumar  
Department of Electronics and Communication, Delhi Technological University,  
Shahbad Daulatpur Village, Rohini, New Delhi, Delhi 110042  
aryankainth\_2k19ec034@dtu.ac.in

**Abstract** – Fraud regarding Credit card transactions is a growing concern for both consumers and financial institutions. Traditional methods of detection, such as rule-based systems and manual review, are often time-consuming and ineffective. Machine learning algorithms offer a potential technique for identifying fraudulent transactions regarding credit cards. These algorithms can learn from past transactions and detect patterns that indicate fraudulent activity. In this paper, we review different machine learning methods that have been used to identify credit card frauds, including decision trees, neural networks, and clustering algorithms. We also discuss the challenges associated with using these techniques, such as handling imbalanced data and ensuring robustness to changing fraud patterns. Ultimately, this article reflects the need for more research in this field and shows how machine learning approaches might be used to predict fraud using credit cards.

**Keywords** – Naïve Baye algorithm, KNN (k-nearest neighbour) algorithm, Random Forest algorithm, and Regression algorithm.

## I. INTRODUCTION

The Term "Fraud" refers to the purposeful deceit and dishonesty committed for a personal advantage. Due to the extensive usage of the internet, many people and organisations have been made the subject of fraud. According to studies, efforts at commercial fraud have increased in 2018 compared to 2016. Additionally, the E-commerce Fraud Index reports that retail fraud rates climbed from 0.06% in 2016 to 0.23% in 2017. Additionally, 10% of all frauds are thought to be credit card-related, costing businesses a lot of money. The number of active cards and the accompanying transaction data are growing along with the number of digital transactions. As a result, researchers have started using numerous tools such machine learning techniques, categorization, and clustering approaches as the amount of data to be analysed throughout the detection phase has expanded. A lot of people are also concentrating on creating early-detection tools for credit card theft. Other academics are looking towards ways to identify fraud that are more accurate and efficient.

Machine learning and other similar approaches, such as Decision Trees (DT), Logistic Regression (LR) are commonly used in the identification of fraud. These AI techniques may be used to a variety of problems across various fields and specialties, many of which usually involve enormous amounts of data. The usefulness of machine learning techniques in these applications has to be evaluated and studied despite the fact that several ways to avoid and identify fraud have been offered.

## A. Clustering

The process of dividing a large set of data into smaller, similar groups based on their common characteristics is known as clustering. Items within a cluster will share similar properties, while items in different clusters may have distinct characteristics. This is illustrated in the Fig. 1.

## B. Classification and Methods

Certain algorithms can be used to find patterns or make inferences when input values are supplied from a sizable dataset. These techniques make an effort to separate multiple outputs as per input. Algorithms for machine learning are frequently utilised to carry out these tasks. The many categories used in machine learning are shown in Fig. 2 below.

## C. Construction of references

The output of this method of supervised learning is directly determined by the user's input data. This is frequently used with dataset that doesn't have any particular discrete values. These forecasts employ ML methods for Performance Evaluation of ML Methods for Detecting Credit Card Fraud. Main objective of this study is to evaluate the efficacy, precision of various ML classifiers as they are applied to prediction analysis and preventative analysis of particular algorithms. Additive techniques like oversampling and binary classification are significant components.

The remainder of the article is arranged and segmented into research-related sections. A summary of the works that are connected to the problem is provided in Section 2. The architecture and methodology investigation are provided in Section 3 with specifics using inference from several classifiers. The analysis of the test findings and their values are shown in Section 4 together with the results and graphical representations. In section 5, which comes to a conclusion, we highlight the future dimensions of the issue and its constraints for future research.

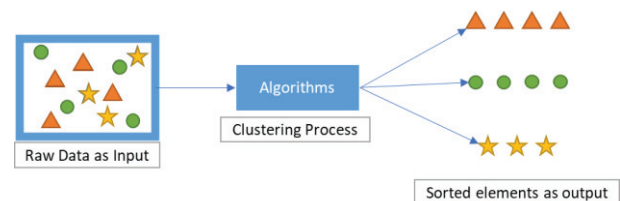


Fig. 1. Clustering technique

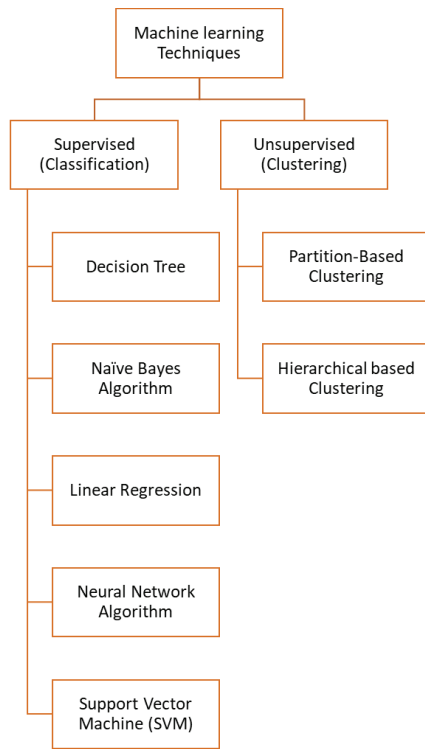


Fig. 2. Classification methods of Machine learning

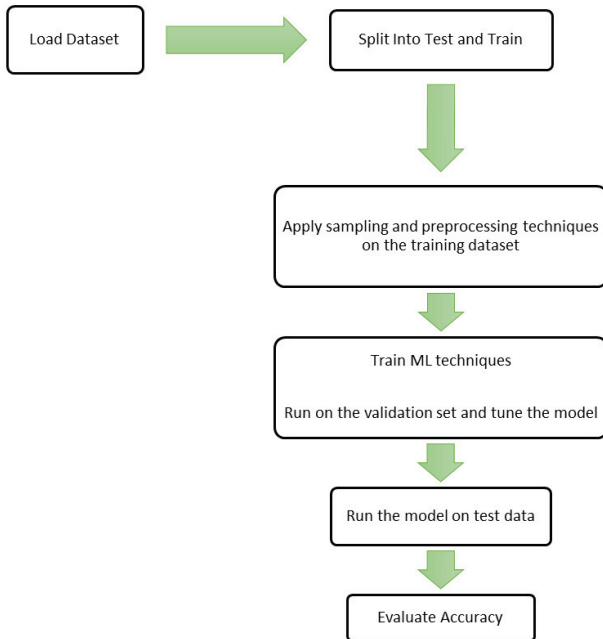


Fig. 3. The flow process diagram for developing a machine

## II. RELATED WORK

The categorization of credit card transactions—as either genuine or fraudulent—is largely a dichotomous problem. In essence, topics like detecting fraud, which evaluates if credit card purchases are legitimate or fraudulent, are included in data mining categorization. Webservices-based cohesive schemes, which enable private entities like financial firms to share information about the configurations and recurrence of fraud in

order to boost fraud detection functionality and minimise financial loss, are another technique and factor method used during fraud detection in addition to data mining. The key processes in developing any Machine Learning model are shown in Fig. 3 below [1]. To tackle the problem of fraud detection, various experimental investigations have used a variety of Machine Learning methods.

A data-driven method for configuring fraud warnings has been put out in [1] and relies on a few aspects including Oversampling under sizes and SMOTE methodology. A few writers [2] have compared several models and their analyses in their works, including XGBOOST, Random Forest, Decision Trees, etc. New methods like Adaboost and Majority Voting approaches that add to or improve the performance of the ML algorithm have also been studied. [3, 4]

While the algorithms are being implemented, Feature Selection (FS) with optimization is utilised in Artificial Neural Networks (ANN) to choose the necessary features [5]. When more than one valid parameter is provided, choosing the most useful characteristic becomes crucial. A novel structured sequenced learning ensemble classifier that enhances performance is also demonstrated [6, 7]. Numerous categories have had their measurements and performance examined. This provides a general notion of the range of metrics that may be taken into account when choosing an algorithm. By comparing the results of 5 various ways, it is possible to analyse the adaptive features selection procedures [8], which make it simpler to separate and weed out the irrelevant traits. The precision and accuracy found in [10, 11] are depicted in the picture below.

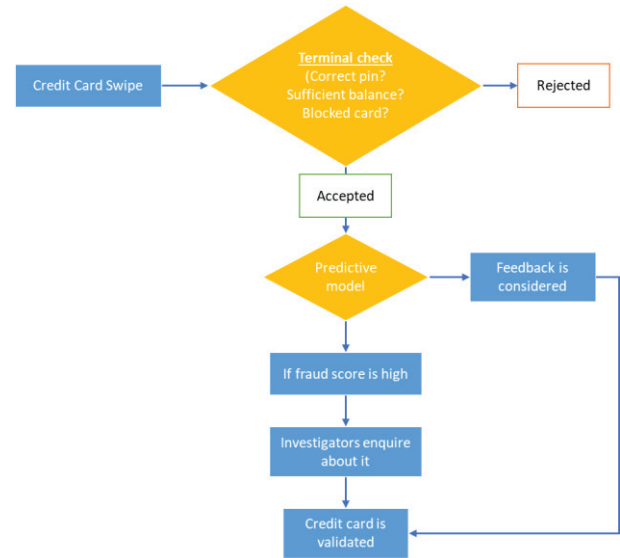


Fig. 4. Working of a credit card fraud detection model

Priya Gupta 's research that in-depth explores the major aspects of RF and its limitations discusses the structured comprehensive investigation on the use of Random Forest [12]. In the study linked [13] which compares and contrasts different strategies, the idea of real-time deep learning and binary data categorization by multiple methods has also been covered. Hassan Najadat and others who have also used other six tactics

like Ada Boost etc. [14] are looking at a powerful and new tool of bidirectional Long short-term memory (BiLSTM) and bidirectional Gated recurrent unit (BiGRU) to help boost the performance [15]. Figure 4 depicts a typical credit card transaction. To determine if a transaction is fraudulent or authentic, novel solutions that successfully address the skew distribution of data are evaluated using ML methodology and the creation of the API (Application Programming Interface) module [16, 17].

A particular ML criterion has been employed on the dataset and F1-score of those are produced to analyse it in the work by Aadhy Kaul and others, [18] and Naive Bayes is used in arbitrary classification. The study examines the various strategies in search of the most effective approach. Machine learning approaches and their capabilities are contrasted with those of data processing techniques, and predictions are made as a result. [19] also describes the 95% accurate supervised based dataset categorization utilising normalisation and principal element analysis.

GridSearchCV for HyperParameters Optimization (HPO), Recursive Feature Elimination (RFE) for the selection of useful predictive features, and Synthetic Minority Oversampling (SMOTE) to solve the imbalanced or disproportionate data problem are three sub-methods that make up the hybrid approach that is also proposed [19]. Apapan Pumsiratin and colleagues in their study introduced an auto-encoder based deep learning approach and restricted Boltzmann Machine (RBM) are applied in hidden layers to uncover patterns [22] and anomalies in the massive quantity of data. The findings show the mean squared error as well as the area underneath the curve. The comparison of several machine learning classifiers and algorithms and their performance accuracy was the exclusive subject of this article.

### III. PROPOSED WORK

Fraud detection is like a Boolean classification issue where each given interaction or exchange is classified as either illegitimate or genuine. Naive Bayes, Decision Trees, Random Forest, and K-Nearest Neighbour methods were a few of the popular classification techniques employed in this study.

Effective use of these algorithms necessitates a number of phases, including data collection, cleaning, research, visualisation, training of the algorithms, and evaluation of the outcomes.

#### A. Naïve Bayes

It's a theory supported by two pillars. Each feature in an entry that has to be split into two halves initially contributes equally. Second, the values they show do not indicate anything about the features because all of the aforementioned traits are statistically unrelated to one another. That may not always be the case, and in these situations, the Bayes rule is used to assess the validity of the assertion. For instance, the projected class is the one with the highest chance.

$$\frac{P[C(i)|f(k)] \times P[(i)]}{P[f(k)](i)P[f(k)|C(i)]} = \prod_{i=1}^n P[f(k)|C(i)]k = 1 \dots, n : i = 1, 2, etc. \quad (1)$$

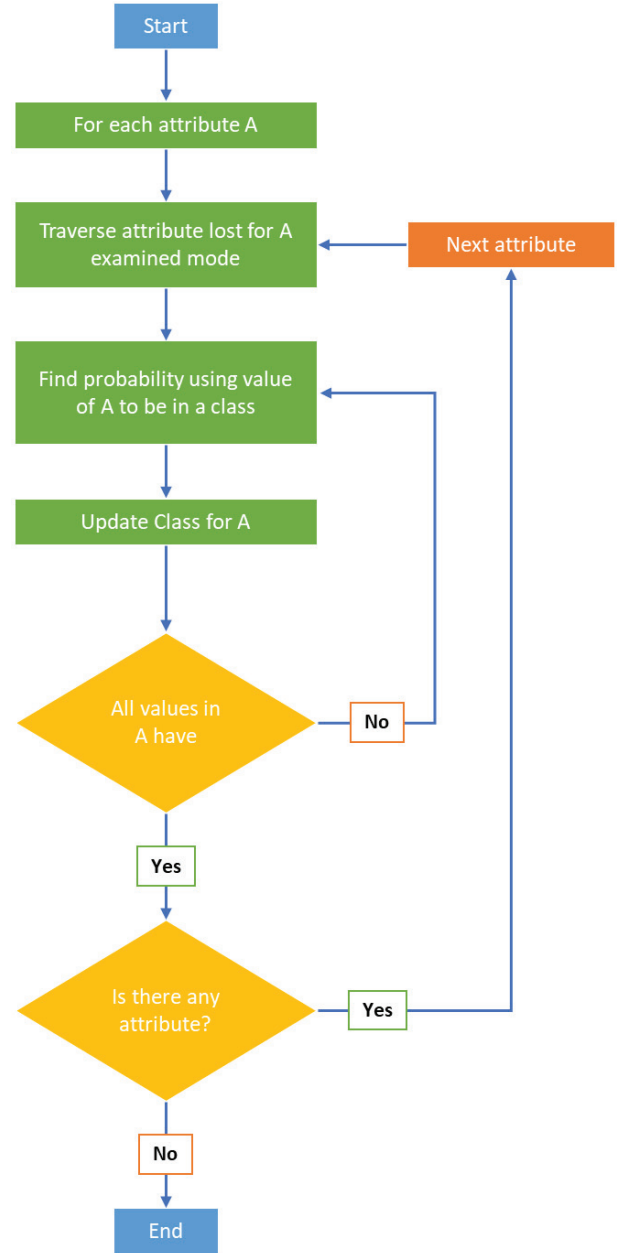


Fig. 5. Naïve Bayes algorithm

Uncertain probabilities are only compared in this case to the previously existing values obtained utilising Bayesian concepts, which rely on prior knowledge to construct logics. The flowchart with multiple predictions in Fig. 5 [21] below depicts the structure of how this algorithm functions.

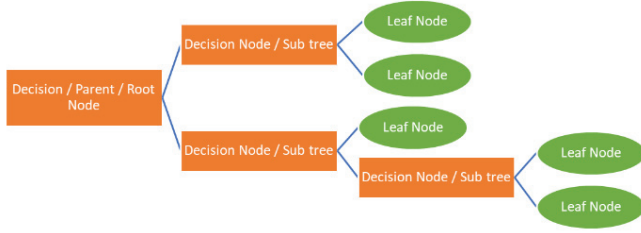


Fig. 6. Structure of a decision tree

### B. Decision Tree

Classification trees and Regression trees are the main types of trees used in this procedure. In this scenario, a decision tree is constructed using a training set. This decision tree's structure is made up of many nodes, the uppermost of which is referred to simply as the root of the tree. The other nodes on the tree stand in for the feature testing, each branch that follows for the test result, but every leaf node for a class label.

The classes that the model would yield if it used that prediction as its final conclusion are likewise shown in these leaf nodes in Fig. 6. Therefore, a proper traversal of the decision tree will reveal the prediction. The C4.5, CART, and ID3 are some examples of decision tree methods. By repeatedly using the divide-and-solve method to break the primary problem into smaller problems, this algorithm handles the fixed set of data. Figure 6 [12] depicts the structure of a decision tree.

### C. KNN Algorithm

This employs a straightforward logic in that it plots all of the training instances already in existence before classifying the instances with no labels dependent on who their closest neighbours are. Here, the instances are used directly for analysis, in contrast to decision trees. However, it is also recognised that since the algorithms in this area were created using training models, they are all already instance-based. In this instance, the unlabelled instance is broken into categories using the metric and the separation between each instance. The class with the largest percentage is designated as the unlabelled class.

### D. Random Forest classifier

This method is really an easy implementation of a Random Forest Bayes classifier, which uses Decision trees [12]. It might also be considered a stage of the Logistic Regression procedure [6]. It is a training algorithm that approaches ensemble-tree optimization techniques in a novel way. The training data set that was used in the experiment was primarily chosen from a range of randomly chosen portions. Thus, as the path continues, the other trees that are a part of it cast ballots for the objects that belong to the class.

- Before importing the appropriate Python libraries for any of the test processes, the relevant data set is uploaded in the frame.
- In agreement, this data has now been separated into a train dataset and a test dataset.

- With the provided training dataset Random Forest Regression model is utilised.
- These actions are followed by an evaluation of the test results, the creation of a forecast, and the creation of the relevant confusion matrix.

## IV. RESULTS AND DISCUSSIONS

Since the goal of this study is to develop an appropriate method to manage the enormous quantities of data that are employed as source for a con-detecting model, a real assessment of M.L. algorithms has been undertaken on a credit card data.

### A. F1 Strategy

To estimate the following models and evaluate their output, the dataset is split into training data and testing data. Let's look into a particular inference made based on the confusion matrix-based F1-score and hold of accuracy. To better comprehend the algorithms' effectiveness, let's evaluate them based on the other three characteristics.

### B. Data analysis and Pre-Processing

The raw data in the dataset utilised for the research was sorted and pre-processed in order to boost the classifiers' efficiency and reduce their learning and execution times else, it would take a while to sort the data according to its basic qualities.

Investigating the dataset feature space and dealing with the dataset's imbalance are additional tasks included in the pre-processing [6].

### C. Performance Matrix

The effectiveness of each strategy may be evaluated using a variety of metrics, including the Matthews Correlation coefficient, sensitivity, specificity, confusion set matrix, balanced classification rate, and even false positive rate. The amount of samples that properly or incorrectly fit each of the identified kinds is listed in a table called a confusion matrix. Positive indicates honest transactions in the fraud detection challenge, whereas negative indicates dishonest transactions.

The three factors that are examined here are as follows:

- *Specificity*: Given the total number of fraud occurrences as determined by equation, this is the amount of frauds that may be anticipated to occur (2).

$$Specificity = \frac{TN}{(TN + FP)} \quad (2)$$

- *Sensitivity*: This is the proportion between the total amount of valid transactions and the number of valid projections. But the most crucial element in fraud is the fraud detection rate or specificity.



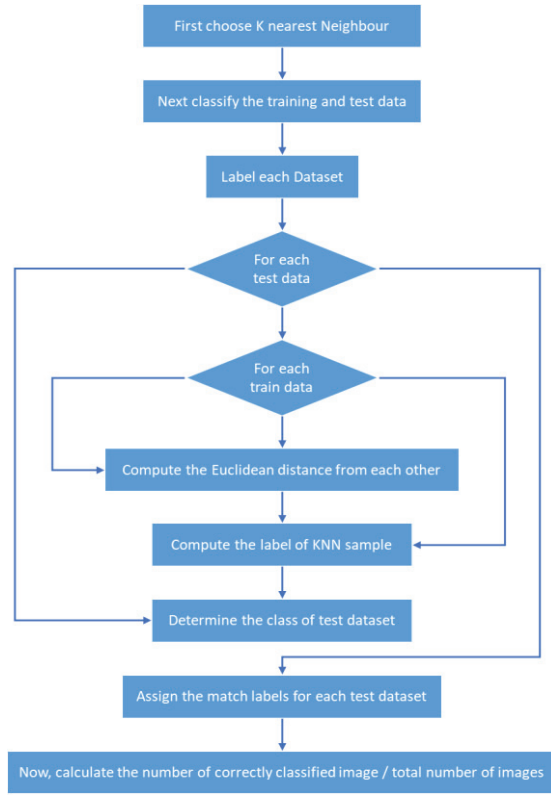


Fig. 7. KNN Flowchart

TABLE I. LOGISTIC REGRESSION OUTPUTS

Logistic Regression	
Accuracy	0.9789583699074237
Specificity	0.9794958731945226
Sensitivity	0.8870748299319728
F1 score	0.06941580756013745

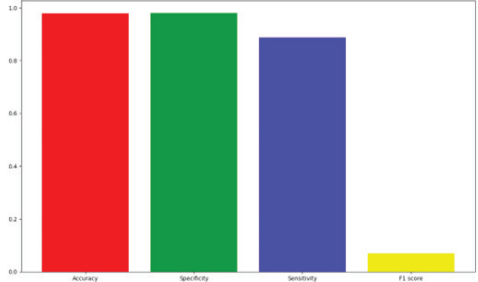


Fig. 8. Performance validation scores

According to equation (3), Recall value is thought to imply a company will sustain the minimum degree of financial harm.

$$Sensitivity = \frac{TP}{(TP + FN)} \quad (3)$$

- *Accuracy*: This parameter offers the overall correctness of the suggested system [6]. Equation (4)

demonstrates that it gives a total forecast to each instance taken.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (4)$$

Whenever it relates to fraudulent transactions, for instance, Occasionally, if the fraction of suspicious transactions to total transactions is relatively tiny, the precision of the model can be seriously deceptive. As a result, the dataset is totally skewed. The objective we're working toward affects the statistics we choose as well. In some cases, one approach could aid in customer satisfaction while the other one would be more suited to stop financial loss.

TABLE II. KNN OUTPUTS

KNN	
Accuracy	0.9383380733354401
Specificity	0.9699882761208029
Sensitivity	0.7408163265306122
F1 score	0.2779220779220779

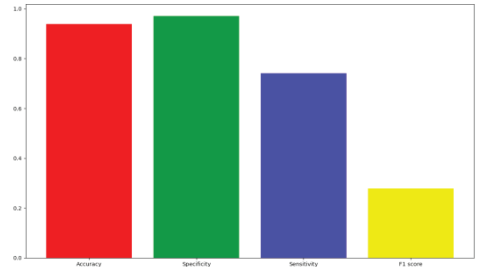


Fig. 9. Performance validation scores

TABLE III. DECISION TREE OUTPUTS

Decision Tree	
Accuracy	0.9491924440855306
Specificity	0.9496248358656912
Sensitivity	0.7482993197278912
F1 score	0.3612456747404844

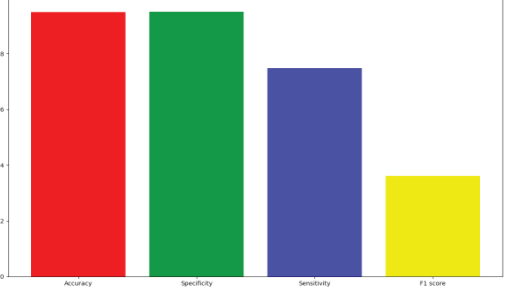


Fig. 10. Performance validation scores



TABLE IV. NAIVE BAYES OUTPUTS

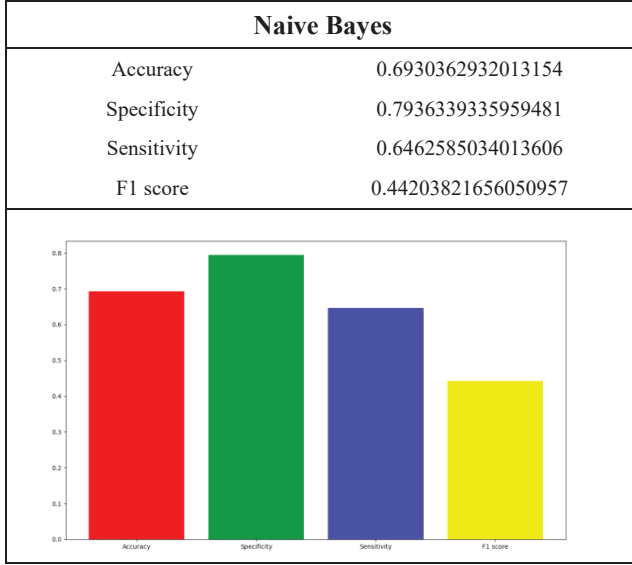


Fig. 11. Performance validation scores

TABLE V. RANDOM FOREST OUTPUTS

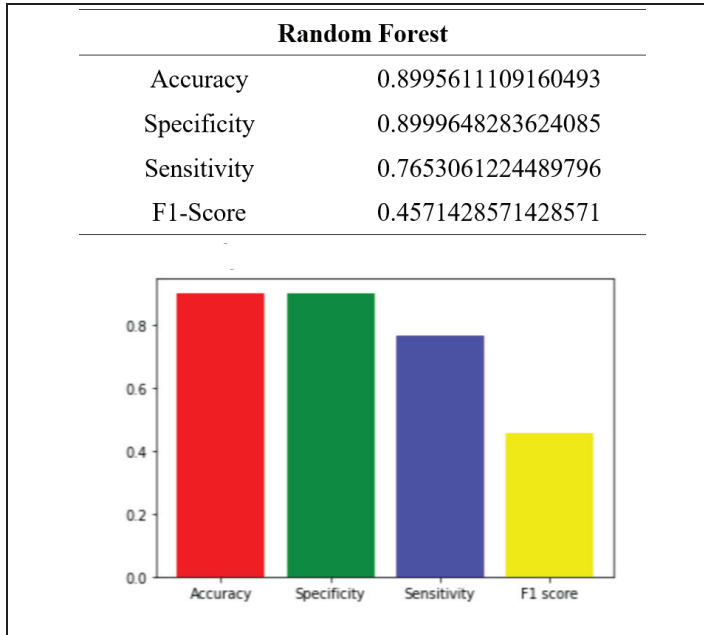


Fig. 12. Performance Validation scores

## V. CONCLUSION AND FUTURE WORK

In conclusion, credit card fraud detection is an important task that requires efficient and accurate methods. Machine learning algorithms have proven to be a valuable tool in this regard, and among these, Naive Bayes, Random Forest, Logistic Regression, and Decision Trees have been widely used. By studying previous transactions and spotting patterns that point to fraudulent behaviour, these algorithms have proven successful in recognising credit card fraud. However, there are also challenges associated with using these techniques, such as handling imbalanced data and ensuring

robustness to changing fraud patterns. Future research should thus concentrate on overcoming these difficulties and creating more sophisticated supervised ML methods for fraud detection with credit cards. Additionally, the integration of other technologies such as biometrics and blockchain can be researched to help make detection of credit card fraud systems more secure and effective. More research on machine learning techniques and technologies is required in order to develop a system that can identify credit card fraud with greater accuracy and dependability.

## REFERENCES

- [1] V. Jain, M. Agrawal and A. Kumar, "Performance Analysis of Machine Learning Algorithms in Credit Cards Fraud Detection," in 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2020.
- [2] R. Akula, "Fraud identification of credit card using ML techniques," in Int. J. Comput. Artif. Intell., 2020.
- [3] A. Krishnaiah and P. B. Divakarachari, "Automatic Music Mood Classification using Multi-class Support Vector Machine based on Hybrid Spectral Features," 2022.
- [4] G. K. Arun and K. Venkatachalapathy, "Intelligent feature selection with social spider optimisation based Artificial Neural Network Model for Credit card Fraud detection," in IIOABJ, 2020.
- [5] T. G. Nguyen, T. V. Phan, D. T. Hoang, T. N. Nguyen and C. So-In, "Efficient SDN-based traffic monitoring in IoT networks with double deep Q-network," in International conference on computational data and social networks, Springer, Cham, 2020.
- [6] X. Li, W. Yu, T. L. J. Q. X. Zheng, J. Zhao and L. X. Y. Li, "Transaction Fraud detection using GRU-Centered Sandwich-structured Model," in Proceedings of the 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design, 2022.
- [7] K. Yu, L. Lin, M. Alazab, L. Tan and B. Gu, "Deep learning-based traffic safety solution for a mixture of autonomous and manual vehicles in a 5G-enabled intelligent transportation system," in IEEE Trans. Intell. Transp. Syst. 22, 2020.
- [8] S. V. Suryanarayana, G. N. Balaji and G. V. Rao, "Machine learning approaches for credit card fraud detection," in Int. J. Eng. Technol., 2018.
- [9] A. Singh and A. Jain, "Adaptive Credit Card Fraud Detection Techniques Based on Feature Selection Method," in University Grants Commission (UGC), Delhi, India, 2022.
- [10] V. Jonnalagadda, P. Gupta and E. Sen, "Credit card fraud detection using Random Forest Algorithm," in Jonnalagadda Vaishnave et al.; International Journal of Advance Research, Ideas and Innovations in Technology, Volume 5, Issue 2.
- [11] B. D. Parameshachari, K. M. Keerthi, T. R. Kruthika, A. Melvina, R. Pallavi and K. S. Poonam, "Intelligent Human Free Sewage Alerting and Monitoring System," in 3rd International Conference on Integrated Intelligent Computing Communication & Security (ICIIC 2021), Atlantis Press, 2021.
- [12] Y. Abakarim, M. Lahby and A. Attioui, "An Efficient Real Time Model For Credit Card Fraud Detection Based On Deep Learning," in SITA'18, Morocco, 2018.
- [13] H. Najadat and e. al, "Credit Card Fraud Detection Based on Machine and Deep Learning," in 2020 11th International Conference on Information and Communication Systems (ICICS), 2022.
- [14] R. K. Dash, T. N. Nguyen, K. Cengiz and A. Sharma, "Fine-tuned support vector regression model for stock predictions," in Neural Comput. Appl., 2021.
- [15] A. Thennakoon, C. Bhagyan, S. Premadasa, S. Mihiranga and N. Kuruwitaarachchi, "Real-time Credit Card Fraud Detection Using Machine Learning," in 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2022.
- [16] Z. Guo, K. Yu, A. Jolfaci, A. K. Bashir, A. O. Almagrabi and N. Kumar, "Fuzzy detection system for rumors through explainable adaptive learning," in IEEE Trans. Fuzzy Syst. 29, 2021.

- [17] A. Kaula, M. Chhabraa, P. Sachdevaa, R. Jaina and P. Nagratha, "Credit Card Fraud Detection Using Different ML and DL Techniques".
- [18] G. B. Rajendran, U. M. Kumarasamy, C. Zarro, P. B. Divakarachari and S. L. Ullo, "Land-use and land-cover classification using a human group-based particle swarm optimization algorithm with an LSTM Classifier on hybrid pre-processing remote-sensing images," in *Remote Sensing* 12, 2020.
- [19] H. A. Shukur and S. Kurnaz, "Credit card fraud detection using machine learning methodology," in *Int. J. Comput. Sci. Mob. Comput.* 8, 2019.
- [20] Z. Guo, K. Yu, Y. Li, G. Srivastava and J. C. W. Lin, "Deep learning-embedded social internet of things for ambiguity-aware social recommendations," in *IEEE Trans. Netw. Sci. Eng.*, 2021.
- [21] N. Rtayli and N. Enneya, "Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization," in *J. Inf. Secur. Appl.* 55, 2020.
- [22] D. L. Vu, T. K. Nguyen, T. V. Nguyen, T. N. Nguyen, F. Massacci and P. H. Phung, "A convolutional transformation network for malware classification," in 2019 6th NAFOSTED conference on information and computer science (NICS); IEEE, 2019.
- [23] A. Pumsirirat and L. Yan, "Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine," in (IJACSA); *Int. J. Adv. Comput. Sci. Appl.* 9, 2018.

ACCEPTED MANUSCRIPT

# Physics Based Numerical Model of a Nanoscale Dielectric Modulated Step Graded Germanium Source Biotube FET Sensor: Modelling and Simulation

To cite this article before publication: Amit Das *et al* 2023 *Phys. Scr.* in press <https://doi.org/10.1088/1402-4896/acf4c9>

## Manuscript version: Accepted Manuscript

Accepted Manuscript is “the version of the article accepted for publication including all changes made as a result of the peer review process, and which may also include the addition to the article by IOP Publishing of a header, an article ID, a cover sheet and/or an ‘Accepted Manuscript’ watermark, but excluding any other editing, typesetting or other changes made by IOP Publishing and/or its licensors”

This Accepted Manuscript is © 2023 IOP Publishing Ltd.



During the embargo period (the 12 month period from the publication of the Version of Record of this article), the Accepted Manuscript is fully protected by copyright and cannot be reused or reposted elsewhere.

As the Version of Record of this article is going to be / has been published on a subscription basis, this Accepted Manuscript will be available for reuse under a CC BY-NC-ND 3.0 licence after the 12 month embargo period.

After the embargo period, everyone is permitted to use copy and redistribute this article for non-commercial purposes only, provided that they adhere to all the terms of the licence <https://creativecommons.org/licenses/by-nc-nd/3.0>

Although reasonable endeavours have been taken to obtain all necessary permissions from third parties to include their copyrighted content within this article, their full citation and copyright line may not be present in this Accepted Manuscript version. Before using any content from this article, please refer to the Version of Record on IOPscience once published for full citation and copyright details, as permissions may be required. All third party content is fully copyright protected, unless specifically stated otherwise in the figure caption in the Version of Record.

View the [article online](#) for updates and enhancements.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

# Physics Based Numerical Model of a Nanoscale Dielectric Modulated Step Graded Germanium Source Biotube FET Sensor: Modelling and Simulation

Amit Das<sup>1,\*</sup>, Sonam Rewari<sup>2,\*</sup>, Binod Kumar Kanaujia<sup>1</sup>, S.S. Deswal<sup>3</sup> and R.S. Gupta<sup>4</sup>

<sup>1</sup> School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi-110067, India  
<sup>2</sup> Department of Electronics and Communication Engineering, Delhi Technological University, New Delhi-110042, India  
<sup>3</sup> Department of Electrical and Electronics Engineering, Maharaja Agrasen Institute of Technology, New Delhi-110086, India  
<sup>4</sup> Department of Electronics and Communication Engineering, Maharaja Agrasen Institute of Technology, New Delhi-110086, India  
\* Authors to whom any correspondence should be addressed.

E-mail: [amitofficial7492@gmail.com](mailto:amitofficial7492@gmail.com) and [rewarisonam@gmail.com](mailto:rewarisonam@gmail.com)

Keywords: biotube, gate all around, MOSFET, MOSFET biosensors, SILVACO TCAD, step-graded doping, surrounding gate MOSFET

## Abstract

This paper proposes a novel dielectric modulated step-graded germanium source biotube FET for label-free biosensing applications. Its integrated structure and unique design combine the benefits of the gate stack, germanium source, triple-gate architecture, and a step-graded biotube channel, resulting in superior performance over existing biosensors. A compact two-dimensional analytical model for channel potential, drain current, threshold voltage, and subthreshold swing has been formulated and agrees well with the simulated results. The comprehensive investigation of different device parameters, including doping and bias, offers valuable insights into optimizing the biosensor's performance. The proposed biosensor exhibits remarkable sensitivity, achieving up to 263 mV and 1495.52 nA for certain biomolecules, which has been validated by a compact analytical model and simulations performed on the SILVACO TCAD simulator. Several parameters are employed to assess the biosensor's effectiveness: threshold voltage,  $I_{ON}/I_{OFF}$  ratio, subthreshold swing, off-current, peak trans-conductance, and on-current. Furthermore, the biotube channel design enables lightweight and cost-efficient biosensors, enhancing the biosensor's practicality. This work also includes an analysis of the effect of temperature on the biosensor's performance and characteristics, providing insights into practical applications. High sensitivity of the biosensor signifies a significant advancement in biosensing technology, suggesting a wide range of potential applications in biomedical field.

## List of abbreviations and symbols

Abbreviation	Description
CMOS	Complementary Metal Oxide Semiconductor
DM	Dielectric Modulated
DM-FET	Dielectric Modulated-Field Effect Transistor
DM-SGGS-BTFS	DM-Surrounding Gate Germanium Source-Biotube FET Sensor
FET	Field Effect Transistor
FTSGP	Forward Triple Step Graded Profile
IIFET	Impact Ionization Field Effect Transistor
$g_d$	Output Conductance
$g_m$	Transconductance
$g_{mp}$	Peak Transconductance
$I_{ON}$	On-Current
$I_{OFF}$	Off-Current
$I_{ON}/I_{OFF}$	Current Ratio
$K_{bio}$	Dielectric Constant of Biomolecules
MOSFET	Metal Oxide Semiconductor Field Effect Transistor
$N_f$	Charge Density of Biomolecules
RTSGP	Reverse Triple Step Graded Profile
SCE	Short Channel Effects
SG-MOSFET	Surrounding Gate-MOSFET
$S_{FCR}$	Fractional Sensitivity in $I_{ON}/I_{OFF}$ Ratio
$S_{gm}$	Transconductance Sensitivity

$S_{gd}$	Output Conductance Sensitivity
$S_{gmp}$	Peak Transconductance Sensitivity
$S_{IDS}$	Drain Current Sensitivity
$S_M$	Sensitivity with respect to sensing metric 'M'
$S_{SS}$	Subthreshold Swing Sensitivity
$S_{Vt}$	Threshold Voltage Sensitivity
TCAD	Technology Computer Aided Design
TFET	Tunnel Field Effect Transistor
$V_{GS}$	Gate Voltage
$V_{DS}$	Drain Voltage

## 1. Introduction

The exponential and rapid development of technology has led to the progressive evolution of Field Effect Transistor (FET) devices [1–4]. Field effect transistors are effectively used in sensing applications due to scalability and high sensitivity. Sensing technologies have evolved over the years, leveraging advancements in electronics and miniaturization to enhance their capabilities. From simple temperature and pressure sensors to sophisticated devices like pH/biomolecules/gas sensors, the range of sensing applications has expanded rapidly. Most of the FET sensors utilize the basic principles of dielectric modulation to detect and measure various physical parameters. To enhance the selectivity of these dielectric modulated FET based biosensors, the FET device is integrated with a sensing element, which could be a specific material, a biomolecule, or an enzyme. The development of FET sensing and biosensors has opened up new possibilities for real-time, label-free, and portable sensing devices, with the potential to transform healthcare, agriculture, and environmental monitoring.

FET based biosensors generally utilize the principle of dielectric modulation for label-free detection of biochemical species. Dielectric modulated FET (DM-FET) biosensors have gained attention in recent years for their potential applications in biomedical field [5,6]. The main advantage of a DM-FET based biosensor is its ability to detect biomolecules without the need for labeling (label-free detection). Generally, when biomolecules are trapped within the embedded cavity (without any bioreceptor layer), the characteristics of FET undergo changes that enable the detection of biomolecules by analyzing the relative change in the sensing metric (threshold voltage/drain current/current ratio/subthreshold swing). However, a significant drawback of this type of detection is its poor selectivity [6]. To address this limitation, the selectivity of the DM-FET biosensor can be greatly enhanced by incorporating an additional bioreceptor layer, which is functionalized on the gate oxide layer and specifically binds to the target biomolecule (biotarget). The incorporation of an additional bioreceptor layer significantly improves the selectivity but could increase the complexity of the fabrication process in the nanoscale regime and may lead to a decrease in sensitivity (considering the relative change in the value of the sensing metric). Various biotarget - bioreceptor systems have been employed in FET biosensors, as reported in the existing literature [4,7–10]. Examples include the interaction between Streptavidin and Biotin, the utilization of Iris antibody (anti-Iris) for binding to the Iris antigen, and the Avian influenza antibody (anti-AI) targeting the Avian influenza surface antigen (AIA).

In real scenario, DM-FET biosensor utilizes the concept of dielectric modulation to detect the presence or absence of biomolecules. The gate oxide layer is functionalized with a bioreceptor capable of specifically binding the desired biomolecule, known as the biotarget. The interaction between the biotarget and bioreceptor leads to alterations in the electrical characteristics of the device (FET), including electric potential, conduction current, and electron concentration in the channel. The observed changes in electrical properties (characteristics) are similar to the changes induced by the application of an external gate voltage (gate/drain voltage). The degree and extent of changes in the sensing metrics following the binding of biomolecules to the bioreceptor can be utilized for biomolecule detection. Therefore the sensitivity of a DM-FET biosensor can be defined as the change in the biosensor's output, which corresponds to variations in the concentration/charge density of analyte, serves as an indicator for detecting the presence or absence of biomolecules. This detection relies on properties such as the dielectric constant (or permittivity) as well as, charge density of the biomolecules [5–7]. However, DM-FET biosensors also have some disadvantages such as sensitivity to operating conditions and environmental conditions (temperature, humidity, high/low frequency noise etc); manufacturing complexity achieving a high level of accuracy and precision in creating nanometer-scale cavities with desired shape and perfection; selectivity issues (limitation in multi-analyte detection: simultaneous detection of multiple analytes in a single complex sample is challenging and non-specific binding: leads to false-positive output); and stability issues along with the reliability issues affecting the long-term performance (various environmental factors and aging effects can lead to changes in the properties of the dielectric layer, which can impact the biosensor's electrical response; and furthermore, the presence of different biomolecules trapped within the cavity can gradually cause tear over extended periods).

FET has specifically three primary variants which are utilized in biosensing applications: Metal Oxide Semiconductor Field Effect Transistor (MOSFET), Tunnel Field Effect Transistor (TFET), and Impact Ionization Field Effect Transistor (IIFET). TFET and IIFET are the newer variants, whereas MOSFET is one of the oldest and utilitarian variants in the FET family. The rapid progress in nano-bioelectronics has opened up potential applications for MOSFET

in label-free electrical detection of biomolecules. MOSFET has gained attention due to their high sensitivity, compatibility with CMOS technology, and the ease of fabricating biosensing systems onto integrated circuits. The development of TFET based biosensors has emerged as a promising avenue for sensitive and low-power biosensing applications [11–13]. TFET, which operate on the principle of quantum tunneling, offer distinct advantages over traditional MOSFET, such as lower power consumption and reduced leakage currents. These unique characteristics make TFET well-suited for biosensing, where high current sensitivity (subthreshold region) and minimal energy consumption are critical. Dwivedi et al. [14] has addressed the location-dependent sensitivity degradation in conventional TFET by introducing a pocket TFET biosensor, which combines lateral and vertical tunneling in a single architecture. Dwivedi et al. [7] has compared tunneling and accumulation mode p-type transistor architectures as dielectric modulated biosensors, showing that while TFET achieve higher sensitivity near the source-channel junction, accumulation mode transistors exhibit higher sensitivity for biomolecules located away from the junction. While TFET and IIFET offers unique advantages such as lower power consumption or impact ionization-induced amplification, MOSFET currently provide better overall performance, reliability, and compatibility with existing technology. Conventionally, MOSFET is often considered superior to TFET and IIFET for several reasons:-

- i. **Compatibility with existing and latest technology:** MOSFET has been widely used in the semiconductor industry for many years and are compatible with standard CMOS fabrication processes. The manufacturing process for MOSFET is well-established/known, resulting in high yield and reliable production. In contrast, TFET and IIFET are relatively new variants (technologies) and their fabrication processes are still being developed, making them potentially less accessible and reliable (industrial concern/drawback). Furthermore, the issue of random dopant fluctuations is more prevalent in TFET compared to MOSFET. This can be attributed to the conventional pin-structure of TFET, which necessitates high abrupt doping and the use of different materials in the source and drain regions (increasing fabrication complexity) [15,16].
- ii. **No ambipolar conduction:** The issue of ambipolarity in TFET makes them more vulnerable to temperature fluctuations, resulting in a degradation of their noise performance. Additionally, in the context of sensing applications, a noteworthy drawback emerges as various sensing metrics become highly temperature dependent. This means that even a slight change in temperature can cause a significant shift in the value of the sensing metric, ultimately leading to undesirable temperature-dependent sensitivity [15].
- iii. **Higher on-state current:** MOSFET generally exhibit higher on-current than TFET resulting in a slightly better transconductance and current ratio (comparatively, more suitable in amplifiers, and digital applications). Additionally, transconductance and current ratio can be valuable sensing metrics when utilizing MOSFET for different sensing applications.
- iv. **Performance and speed:** MOSFET typically offer higher performance and comparable speed compared to TFET and IIFET. MOSFET has been extensively optimized over the years, allowing for high-speed operation and efficient switching applications. TFET and IIFET, while offering potential advantages such as lower power consumption and lower subthreshold swing, often suffer from reduced performance [15,17–20] in terms of speed (tunneling phenomenon causing delay intrinsic delay in charge carrier transport), noise performance (ambipolar behavior) and lower on-state current driving capability.
- v. **Scalability and integration:** MOSFET has a well-established scaling roadmap, enabling the fabrication of smaller and more integrated devices. They can be readily integrated with other electronic components, allowing for the development of complex systems and sensor arrays. TFET and IIFET (being new emerging devices) may face challenges in terms of scalability and integration due to their unique device structures and associated fabrication complexities [5,21].
- vi. **Wide range of applications:** MOSFET has been extensively studied and applied in various fields, including electronics, telecommunications, and biosensing. The versatility, scalability and compatibility with existing technology make them suitable for a wide range of applications. TFET and IIFET, being relatively new technologies (in terms of usability), are still exploring their potential applications (sensing and biomedical field) and may have limited practical implementations.

The exposition of FET based biosensors for sensing nucleic acids [22], pH levels [23,24], proteins [25], and DNA [26] has been evaluated and shown in the past. Buitrago et al. [27] has experimentally demonstrated a silicon nanowire-based FET biosensor. Many non-planar variants of surrounding gate MOSFET biosensors have been reported in the recent decade. Pratap et al. [28] has proposed a novel MOSFET-based biosensor that utilizes the concept of a junctionless transistor [29]. Chakraborty et al. [30] reported a variant of surrounding gate MOSFET-based biosensor and has shown the effect of gate oxide stacking on the sensitivity. High-k dielectric has numerous advantages [31–33], but direct deposition of  $\text{HfO}_2$  over a silicon substrate is complex due to stability issues and can also degrade the performance of the biosensor. Therefore,  $\text{HfO}_2$  is deposited over a thin layer of  $\text{SiO}_2$  [34,35]. The literature survey reveals that channel doping [36,37] and source doping [38,39] must be kept low and high, respectively, to obtain superior characteristics. Therefore, a detailed study of doping-dependent sensitivity analysis becomes inevitable when investigating the performance of any FET-based biosensor.



MOSFET based biosensors [6] have gained attention due to their scalability and high sensitivity [30]. The key principle behind the biosensing ability of MOSFET-based biosensors is the relative change in the value of device metrics in the presence of biomolecules. Dielectric modulation enables the gate to control the flow of charge carriers. When biomolecules are immobilized inside the cavity, they change the net gate oxide capacitance, which in turn alters the potential and field distribution. This modulation affects device characteristics, such as subthreshold slope, off-current, and threshold voltage. The dielectric modulation of these parameters is the primary reason behind the biosensing capability of MOSFET. MOSFET-based biosensors are capable of label-free detection, which utilizes different molecular properties to sense biomolecules [40,41].

Optimization of device parameters can result in a highly sensitive biosensor. Goel et al. [42] revealed that the sensitivity of biosensors increases with the increase in the number of gates used. The proper choice of source material can substantially improve the sensitivity of the biosensor. Wu et al. [43] and Brunco et al. [44] have discussed the concept of germanium MOSFET in the past decade. Saha et al. [45] reported a germanium source-based TFET for label-free detection of biomolecules. Germanium offers more intrinsic charge carriers than silicon, which participates in the conduction process [46] and are involved in the biosensing action. The relative change in the various sensing metrics will be higher when the numbers of charge carriers are high. Hence, the sensitivity of the biosensor is higher in germanium-based biosensors. Using a graded channel instead of a non-graded channel improves the device's performance [47,48]. Recently, Singh et al. [49] reported a label-free FET biosensor that uses a graded channel. Although germanium-based biosensors are highly sensitive, they have some flaws and drawbacks. Germanium source-based biosensors are costlier than their silicon-based counterparts from an economic perspective, and their use is limited due to the limited availability of fabrication technology. Additionally, germanium has a lower operating temperature range than silicon, which limits the range of temperatures at which it can effectively operate.

This paper comprehensively investigates a dielectric-modulated step-graded germanium-source biotube FET sensor (DM-SGGS-BTFS). It shows enhanced sensitivity due to source and channel engineering, which involves using a germanium source and a graded-doped biotube filled with air [50]. Both physics-based modeling and simulation-based investigation are crucial to optimize the different parameters for designing an ultra-sensitive biosensor. The key novelty of this research work lies in two aspects: the structural innovation of the biotube sensor and the comprehensive investigation through analytical modeling, considering different sensing metrics and the impact of temperature. The structural novelty involves an integrated design of the surrounding gate MOSFET, which incorporates gate engineering (triple gate architecture), source engineering (germanium source), channel engineering (tubular channel with step-graded doping profile), and gate oxide engineering (oxide stacking:  $\text{HfO}_2 + \text{SiO}_2$ ) techniques. The biosensing performance of DM-SGGS-BTFS has been effectively analyzed using multiple key sensing metrics such as threshold voltage, subthreshold swing,  $I_{\text{ON}}/I_{\text{OFF}}$  ratio, drain current, peak transconductance, and peak output conductance. Additionally, the developed analytical model takes into account the influence of temperature on various analog performance parameters and sensing parameters of the proposed biosensor. One of the primary objectives of this paper is to present an analytical model of DM-SGGS-BTFS to investigate its biosensing performance and study the critical insights into the sensitivity dependency on different design parameters. To make this analysis more practical and realistic, the effect of partially filled cavities on sensitivity has also been considered. The analytical model of the biosensor is based on the non-linear parabolic solution of the 2D Poisson's equation [30]. The potential across the channel is expressed in terms of channel-surface potential [51,52]. The channel potential, threshold voltage, subthreshold swing, and drain current have been analytically derived, and the results show a remarkable agreement with the simulated results.

## 2. Device structure and software specifications

Figures 1(a) and 1(b) show the 3D cylindrical structure and a cross-sectional view, respectively, of the DM-SGGS-BTFS. A layer of  $\text{HfO}_2$  is formed over the thin layer of  $\text{SiO}_2$  to form the gate stack. The biotube has an inner radius (a) of 2 nm and an outer radius (b) of 10 nm. The biosensor considers a dual-sided cavity, which has the advantage of better fill-in factor probability than the single-sided cavity. Additionally, the asymmetric cavity has been preferred since the source-end open cavity is comparatively more sensitive to biomolecules than the drain-end open cavity [53]. The work function of the gate in the multi-gated structures can also affect the sensitivity of the biosensor [54]. Kumari et al. [54] in her work mentioned that the sensitivity obtained is highest if the work functions of the gate near to drain is smaller than that of the source ( $\phi_{\text{M1}} > \phi_{\text{M3}}$ ). Therefore, we have extended the same results for the proposed biosensor in our investigation. All the structural parameters of the biosensor are listed in table 1. Table 2 lists the lattice parameters of silicon and germanium in a tabular form. This paper primarily focuses on the biosensing performance of DM-SGGS-BTFS, and biosensing investigation requires changes in the value of sensing metrics. The net electric field across the channel can be resolved into two components: the lateral electric field (across the radial direction) and the horizontal electric field (across the z-direction). The vertical electric field is the primary factor that controls the sensitivity and affects the biosensor's performance. The vertical electrical field is changing due to the immobilization of biomolecules inside the cavity which ultimately changes the gate oxide capacitance [3]. The sensitivity of the biosensor is calculated based on the relative change in the value of the sensing metric. The horizontal electric field (doping dependent) is almost constant since the

doping of the source, channel, and drain are constant ( $N_S=N_D=10^{19}/\text{cm}^3$  and  $N_{\text{Cha},1}=10^{10}/\text{cm}^3$ ,  $N_{\text{Cha},2}=10^{14}/\text{cm}^3$ ,  $N_{\text{Cha},3}=10^{18}/\text{cm}^3$ ) [6,55], so it has a negligible effect on the sensitivity.

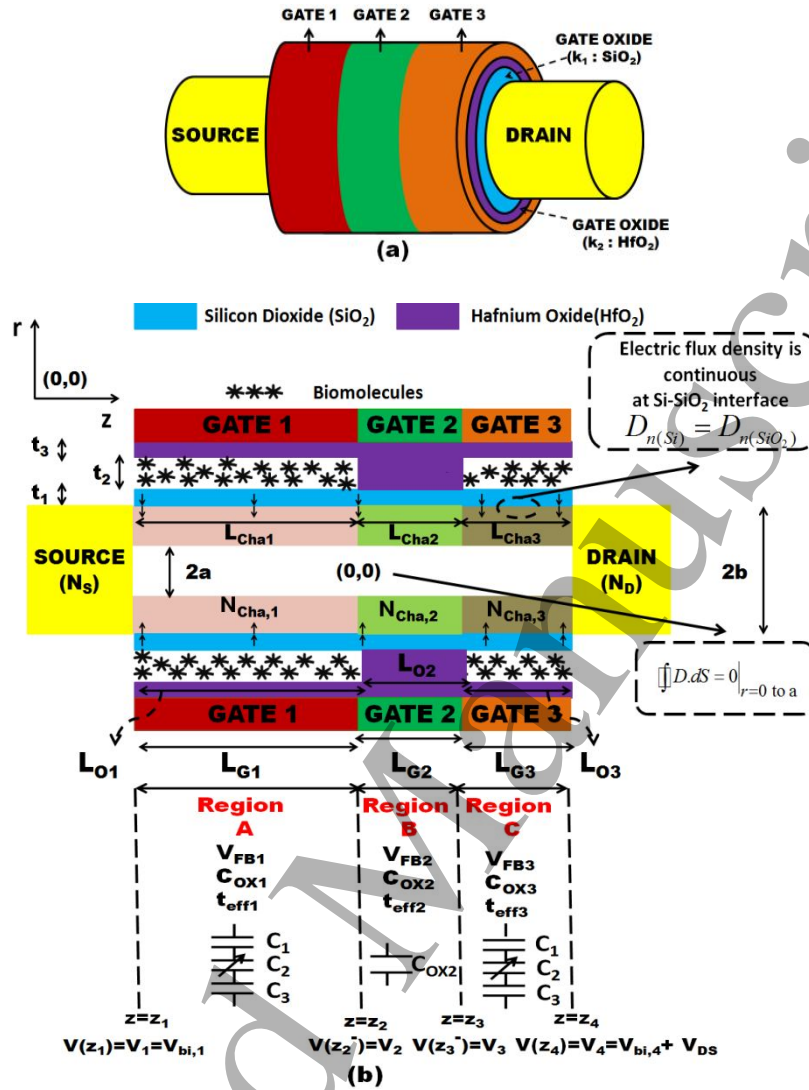


Table 1. Structural parameters used in the simulation of DM-SGGS-BTFS.

Parameter		Value		Parameter			Value	
Doping	Drain (N <sub>D</sub> )	10 <sup>19</sup> /cm <sup>3</sup>	Silicon	Gate length &	L <sub>G1</sub>	ϕ <sub>M1</sub>	30 nm	5.10 eV [56]
	Channel (N <sub>Cha</sub> )	10 <sup>10</sup> -10 <sup>18</sup> /cm <sup>3</sup>	Silicon		L <sub>G2</sub>	ϕ <sub>M2</sub>	15 nm	4.53 eV [56]
	Source (N <sub>S</sub> )	10 <sup>19</sup> /cm <sup>3</sup>	Germanium	Work function	L <sub>G3</sub>	ϕ <sub>M3</sub>	15 nm	4.10 eV [57]
Oxide thickness (t <sub>OX</sub> )		t <sub>O1</sub> = t <sub>O3</sub> =1nm	t <sub>O2</sub> =6nm	Source/Drain length			20 nm	
Oxide length (L <sub>TOT</sub> )		L <sub>O1</sub> =30nm	L <sub>O2</sub> =L <sub>O3</sub> =15nm	Channel radius (a & b)			a=2 nm	b=10nm
Cavity length (L <sub>CAV</sub> )		45 nm		Cavity thickness (t <sub>CAV</sub> )			6 nm	
n <sub>i</sub> (T=300K)		Si: 1.5*10 <sup>10</sup> /cm <sup>3</sup>	Ge: 2.5*10 <sup>13</sup> /cm <sup>3</sup>	Boltzmann Constant (K <sub>B</sub> )			1.38*10 <sup>-23</sup> J/K	
Biomolecules (K <sub>bio</sub> )	B <sub>1</sub> : Streptavidin (K <sub>bio</sub> =2.1) [42]		B <sub>2</sub> : Protein (K <sub>bio</sub> =2.5) [28]			B <sub>3</sub> : Biotin (K <sub>bio</sub> =2.63) [58]		
	B <sub>4</sub> : ChOx (K <sub>bio</sub> =3.3) [28]		B <sub>5</sub> : APTES (K <sub>bio</sub> =3.57) [30]			B <sub>6</sub> :Hydroprotein (K <sub>bio</sub> =5) [59]		
	B <sub>7</sub> : Keratin (K <sub>bio</sub> =8) [49]				B <sub>8</sub> : DNA (K <sub>bio</sub> =5 & N <sub>i</sub> =-1e10/cm <sup>2</sup> — -1e12/cm <sup>2</sup> ) [60,61]			
Physical models	BGN – Bandgap narrowing due to high doping							
	Boltzmann – Model carrier statistics				Lombardi CVT – Basic model for non planar MOSFET			
	Auger – To incorporate effects of recombination at high charge densities				FLDMOB – Model velocity saturation at high doping and temperature			
	CONMOB – Model concentration-dependent mobility of charge carriers				SRH – Recombination model used in surrounding gate MOSFET			

**Table 2.** Lattice parameters.

Lattice Parameters	Silicon	Germanium
Atomic number	14	32
Atomic weight	28.086 u	72.59 u
Density	2.33 g/cm <sup>3</sup>	5.32 g/cm <sup>3</sup>
Crystal structure	Diamond cubic (FCC)	Diamond cubic (FCC)
Atoms	5*10 <sup>22</sup> cm <sup>-3</sup>	4.42*10 <sup>22</sup> cm <sup>-3</sup>
Lattice constant	5.429 Å	5.657 Å
Nearest neighbour distance	0.235 nm	0.211 nm
Atomic radius	0.132 nm	0.137 nm
Nature of band gap	Indirect	Indirect
Dielectric constant	11.9	16
Energy gap	1.12 eV	0.67 eV
Electron affinity	4.05 eV	4 eV
Intrinsic carrier concentration	1.5*10 <sup>10</sup> cm <sup>-3</sup>	2.5*10 <sup>13</sup> cm <sup>-3</sup>
Intrinsic Debye length	24 µm	0.68 µm
Intrinsic resistivity	2.3*10 <sup>5</sup> Ω-cm	45 Ω-cm
Effective DOS in CB (N <sub>c</sub> )	2.8*10 <sup>19</sup> cm <sup>-3</sup>	1.04*10 <sup>19</sup> cm <sup>-3</sup>
Effective DOS in VB (N <sub>v</sub> )	1.04*10 <sup>19</sup> cm <sup>-3</sup>	6.0*10 <sup>18</sup> cm <sup>-3</sup>
Melting point	1415° C	936° C
Electronic configuration	1s <sup>2</sup> 2s <sup>2</sup> 2p <sup>6</sup> 3s <sup>2</sup> 3p <sup>2</sup>	1s <sup>2</sup> 2s <sup>2</sup> 2p <sup>6</sup> 3s <sup>2</sup> 3p <sup>6</sup> 4s <sup>2</sup> 3d <sup>10</sup> 4p <sup>2</sup>

Figure 2(a) shows the overall fabrication flowchart [62–64] of DM-SGGS-BTFS [65], while figure 2(b) shows the comparison of its sensitivity with an existing FET-based biosensor [28]. The calibration of the proposed device with the experimental work reported by Choi et al. [66] is shown in figure 2(c). To better understand the improved biosensing action of the proposed biosensor, table 3 lists the structural parameters of a conventional surrounding gate MOSFET, and table 4 shows a sensitivity comparison of DM-SGGS-BTFS with a conventional surrounding gate MOSFET-based biosensor. The basic parameters and structure of conventional MOSFET and DM-SGGS-BTFS were kept the same to facilitate the comparison, and sensitivity was obtained by calculating the relative change ( $S_M$ ) or the fractional change ( $S_{FM}$ ) [6] where, M denotes the metric of the biosensor. Different sensing metrics have been considered in the analysis for a more realistic and reasonable comparison. Also, the fractional sensitivity improvement in the current ratio can be seen in the proposed structure, which can also be used as a potential sensing metric. It can be seen that the sensitivity improves significantly due to the structural engineering in DM-SGGS-BTFS when compared to a conventional MOSFET. Silicon and germanium have different mobility characteristics for electrons and holes, which is one of the primary reasons behind the different sensitivity characteristics. To be more precise, the mobility of the conducting electrons is approximately twice as high, while the mobility of holes is approximately four times higher in germanium compared to silicon [7]. The high mobility in germanium can be attributed to the lower effective mass of its electrons and holes. This mobility enhancement leads to an improvement in the subthreshold characteristics. Therefore, for the same range of gate or drain voltage, a biosensor utilizing a germanium source (Ge-source) will exhibit a comparatively greater variation (swing) in current compared to one utilizing a silicon source. This means DM-SGGS-BTFS utilizing a germanium source will exhibit a larger variation in sensing metrics like subthreshold swing and threshold voltage, increasing its sensitivity and making it more sensitive to biomolecules when compared to a biosensor utilizing a silicon source.

DM-SGGS-BTFS is a label-free biosensor that uses the molecular and physical properties of biomolecules to detect and sense them [30,67–70]. Kim et al. [71] has practically demonstrated the entrapment of streptavidin (size ~5 nm) inside a cavity. Different biomolecules are variable in size ranging from few micrometers to few attometers [28]. A dual sided asymmetric cavity has been considered which facilitates a high fill-in factor probability along with enhanced sensitivity [72]. This work considered the different biomolecules ( $K_{bio}=2-8$  and size<6 nm) which can be easily localized and entrapped inside the cavity (size: 30\*6 nm<sup>2</sup>) of DM-SGGS-BTFS. The orientation of biomolecules also affects the localization, i.e. even few biomolecules whose size is greater than 6 nm (less than 30 nm) can be localized inside the cavity in horizontal orientation.

The numerical simulator used is SILVACO TCAD [73] which uses the NEWTON GUMMEL method to solve the complex coupled differential equations. Different models that have been used are mentioned in table 1. For the simulations performed on the TCAD simulator, the cavity is presumed to be filled with a material having some finite dielectric constant ( $K_{bio} \neq 1$ ) when the entire cavity is pervaded with different biomolecules. When the cavity is empty (no biomolecules are present), it is presumed to be filled with air ( $K_{air}=1$ ). The asymmetric cavity provides comparatively more sensitivity when compared to symmetric cavity. Various biomolecules have been considered in this work, which solidifies the biosensing investigation of DM-SGGS-BTFS. All the test biomolecules are mentioned in table 1 and have been used to analyze the sensitivity of biosensors in the past [28,30,42].

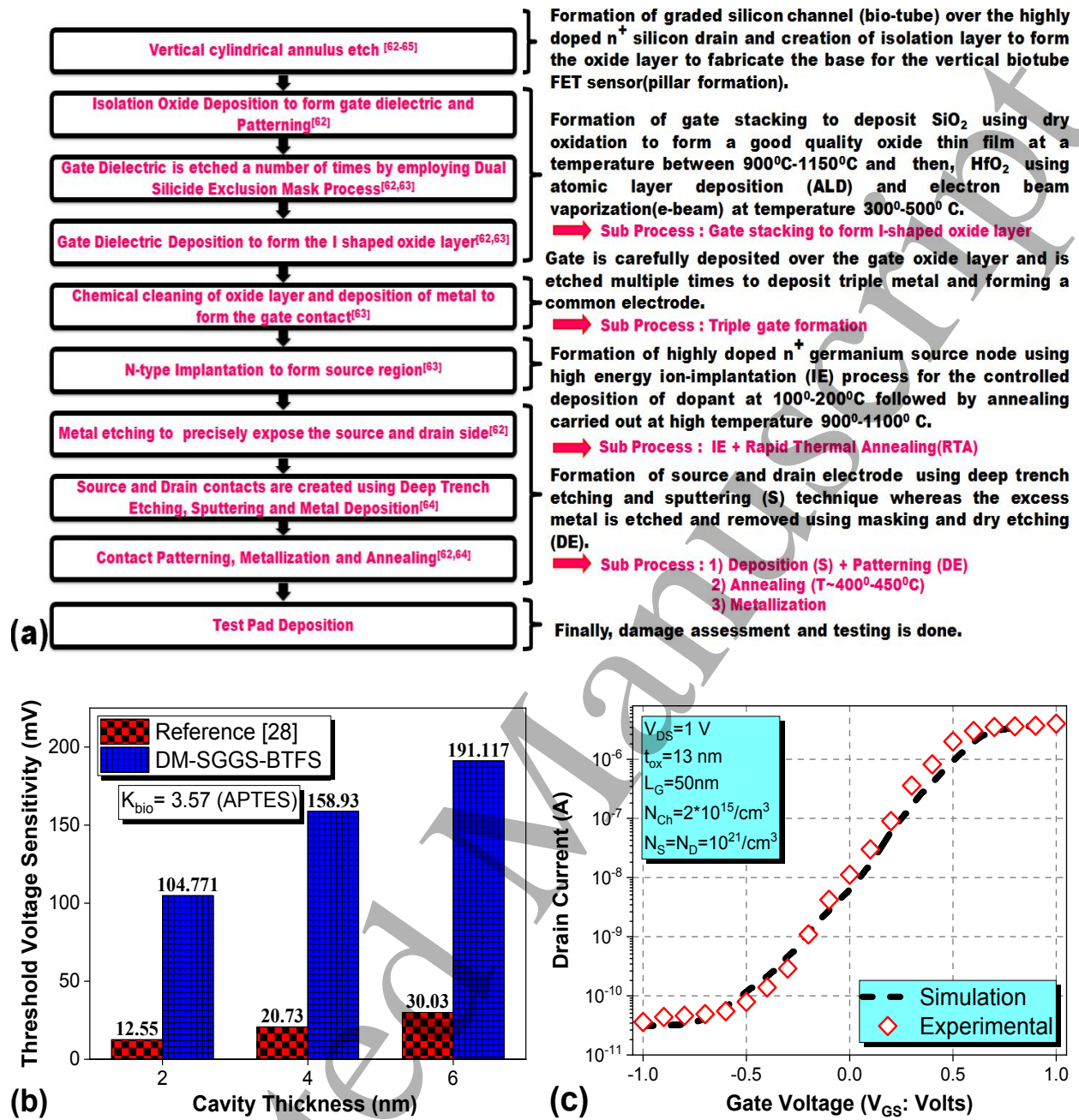


Figure 2. (a) Flow chart illustrating the basic fabrication steps [74], (b) sensitivity comparison and (c) calibration with experimental work [66] in DM-SGGS-BTFS.

Table 3. Parameters description of DM-SGGS-BTFS and conventional MOSFET.

Parameters	Conventional MOSFET	Our Work
Source/Drain doping	$10^{19} \text{ cm}^{-3}$	$10^{19} \text{ cm}^{-3}$
Channel doping	$10^{10} \text{ cm}^{-3}$	$N_{\text{Cha},1} = 10^{10} \text{ cm}^{-3} / N_{\text{Cha},2} = 10^{14} \text{ cm}^{-3} / N_{\text{Cha},3} = 10^{18} \text{ cm}^{-3}$
Work function	4.8 eV	$\phi_{\text{M}1} = 5.1 \text{ eV} / \phi_{\text{M}2} = 4.53 \text{ eV} / \phi_{\text{M}3} = 4.1 \text{ eV}$
Gate oxide	$\text{SiO}_2$ (8nm)	$\text{SiO}_2$ (1nm) + $\text{HfO}_2$ (7nm)
Channel length	60 nm	60 nm
Cavity length	30 nm + 15 nm	30 nm + 15 nm
Cavity thickness	6 nm	6 nm
Channel radius	10 nm	10 nm
Source material	Silicon	Germanium
$L_{\text{O}1} = 30 \text{ nm}, L_{\text{O}2} = 15 \text{ nm}, L_{\text{O}3} = 15 \text{ nm}$		$L_{\text{Cha},1} = 30 \text{ nm}, L_{\text{Cha},2} = 15 \text{ nm}, L_{\text{Cha},3} = 15 \text{ nm}$
$L_{\text{G}1} = 30 \text{ nm}, L_{\text{G}2} = 15 \text{ nm}, L_{\text{G}3} = 15 \text{ nm}$		



**Table 4.** Sensitivity comparison of DM-SGGS-BTFS with a conventional SG-MOSFET.

Biosensor	Sensitivity	$S_M =  M(\text{without\_biomolecules}) - M(\text{with\_biomolecules}) $ [28]						
		M : Threshold Voltage ( $S_{Vt}$ : mV)		M : Subthreshold Swing ( $S_{SS}$ : mV/decade)		M : Off Current ( $S_{IOFF}$ : fA)		
		$S_{FCR} = \left  \frac{CR(\text{with\_biomolecules})}{CR(\text{without\_biomolecules})} \right $		$CR = \frac{I_{ON}}{I_{OFF}}$ [70]		$I_{ON}$ : On Current $I_{OFF}$ : Off Current CR : Current Ratio (10 <sup>7</sup> )		
Neutral biomolecules		B <sub>1</sub>	B <sub>2</sub>	B <sub>3</sub>	B <sub>4</sub>	B <sub>5</sub>	B <sub>6</sub>	B <sub>7</sub>
Conventional SG-MOSFET	$S_{Vt}$	45.839	53.24	55.017	59.936	60.166	66.502	75.868
	$S_{IOFF}$	1358.282	1430.553	1445.428	1490.462	1500.333	1524.958	1537.601
	$S_{SS}$	6.0116	7.2199	7.5517	8.923	9.3524	10.924	12.448
	$S_{FCR}$	11.5616	20.13218	23.52495	45.36716	55.93344	121.5954	268.3308
	$I_{ON}/I_{OFF}$	46.8831	81.6374	95.3953	183.967	226.814	493.078	1088.1
DM-SGGS-BTFS	$S_{Vt}$	137.484	158.385	163.811	184.908	191.117	212.456	231.596
	$S_{IOFF}$	3367.4676	3397.0077	3401.0546	3409.24658	3410.26932	3411.729197	3411.967703
	$S_{SS}$	6.0618	7.8327	8.3253	10.3935	11.096	13.5831	15.7545
	$S_{FCR}$	73.07833	213.3475	290.6677	1126.405	1775.47	10886.79	86145.18
	$I_{ON}/I_{OFF}$	7.3804	21.5466	29.3554	113.759	179.31	1099.49	8700.06
Charged biomolecules		B <sub>81</sub>	B <sub>82</sub>	B <sub>83</sub>	B <sub>84</sub>	B <sub>85</sub>	B <sub>86</sub>	
Conventional SG-MOSFET	$S_{Vt}$	68.034	79.857	92.847	114.126	182.702	241.128	
	$S_{IOFF}$	1526.455	1530.443	1534.752	1542.807	1549.301	1549.735	
	$S_{SS}$	10.9149	10.8865	10.8511	10.7442	10.4423	10.2505	
	$S_{FCR}$	129.061643	154.413364	196.4094331	406.5429203	4666.701191	28556.10384	
	$I_{ON}/I_{OFF}$	523.354	626.157	796.454	1648.56	18923.8	115797	
DM-SGGS-BTFS	$S_{Vt}$	213.277	215.646	218.641	226.528	244.601	256.891	
	$S_{IOFF}$	3410.74998	3411.803	3414.857	3415.945	3416.998	3420	
	$S_{SS}$	13.5916	13.6154	13.6438	13.7228	13.7661	13.9795	
	$S_{FCR}$	11579.8124	13921.65794	17762.22114	36474.01305	441801.9071	1414157.417	
	$I_{ON}/I_{OFF}$	1169.48	1405.99	1793.86	3683.62	4519.6	142820	
B <sub>81</sub> : $N_f = -1*10^{10}/\text{cm}^2$ B <sub>82</sub> : $N_f = -4*10^{10}/\text{cm}^2$ B <sub>83</sub> : $N_f = -8*10^{10}/\text{cm}^2$ B <sub>84</sub> : $N_f = -2*10^{11}/\text{cm}^2$ B <sub>85</sub> : $N_f = -6*10^{11}/\text{cm}^2$ B <sub>86</sub> : $N_f = -1*10^{12}/\text{cm}^2$		$\phi_{M1} = 5.10 \text{ eV}$ [56] , $\phi_{M2} = 4.53 \text{ eV}$ [56] , $\phi_{M3} = 4.10 \text{ eV}$ [57] $\phi_{M(SGM)} = 4.80 \text{ eV}$ [75]						

### 3. Device modelling

#### 3.1 Channel potential

The potential distribution  $\psi(r, z, T)$  across the channel satisfies 2D Poisson's equation and is given by,

$$\frac{1}{r} \frac{\partial}{\partial r} \left[ r \frac{\partial}{\partial r} \{ \psi(r, z, T) \} \right] + \frac{\partial^2}{\partial z^2} \psi(r, z, T) = -\frac{qN_{Cha}}{\epsilon_{Cha}} \quad (1)$$

where  $q$ ,  $N_{Cha}$ , and  $\epsilon_{Cha}$  denote the electronic charge, channel doping, and channel permittivity, respectively. The radius of the channel is  $b$ , so  $r=b$  denotes the surface interface between the channel and oxide layer with potential expressed as  $\psi_s(z, T)$  or  $\psi(b, z, T)$ .  $\psi(r, z, T)$  is an implicit function of multiple variables, so its solution follows the parabolic potential profile and can be expressed as

$$\psi(r, z, T) = M_0(z, T) + M_1(z, T)r + M_2(z, T)r^2 \quad (2)$$

where coefficients  $M_0$ ,  $M_1$ , and  $M_2$  (mentioned in Appendix A) be determined using the different boundary conditions.

Various boundary conditions used are :-

$$(i) \psi(r, z_1, T) = V_{bi,1} = \frac{k_B T}{q} \left\{ \ln \left( \frac{N_S N_{Cha_1}}{n_{i(Si)} n_{i(Ge)}} \right) \right\} \quad (3)$$

where,  $V_{bi,1}$  denotes the potential barrier at the source-channel junction,  $n_{i(Ge)}/n_{i(Si)}$  are the intrinsic concentration of charge carriers in germanium and silicon,  $k_B$  denotes the Boltzmann constant ( $1.38066*10^{-23} \text{ J/K}$ ), and  $T$  denotes the temperature.

$$(ii) \psi(r, z_4, T) = V_{bi,4} + V_{DS} = \frac{k_B T}{q} \ln \left( \frac{N_{Cha_3} N_D}{n_{i(Si)}^2} \right) + V_{DS} \quad (4)$$

where,  $V_{bi,4}$  denotes the potential barrier at the drain-channel junction, and  $N_D$  ( $N_S$ ) denotes the doping of drain (source).

(iii) Electric field is continuous at  $z=z_2$  and  $z=z_3$ ,

$$\left. \frac{\partial \{\psi(r, z, T)\}}{\partial z} \right|_{z=z_2^-} = \left. \frac{\partial \{\psi(r, z, T)\}}{\partial z} \right|_{z=z_2^+}, \quad \left. \frac{\partial \{\psi(r, z, T)\}}{\partial z} \right|_{z=z_3^-} = \left. \frac{\partial \{\psi(r, z, T)\}}{\partial z} \right|_{z=z_3^+} \quad (5)$$

(iv) Electric field inside the bio-tube is zero.

$$\left. \frac{\partial \psi(r, z, T)}{\partial r} \right|_{r=a} = 0 \quad (6)$$

$$(v) \psi(r, z_2^-, T) = V_2, \quad \psi(r, z_3^-, T) = V_3 \quad (7)$$

(vi) Electric field is continuous at the Si-SiO<sub>2</sub> interface since electric flux density is the same just above and below the Si-SiO<sub>2</sub> interface in the absence of any interface trap charges.

$$\epsilon_{Cha} \left. \frac{\partial \{\psi(r, z, T)\}}{\partial r} \right|_{r=b} = C_{ox} [V_{GS} - V_{FB} - \psi(r, z, T)]_{r=b}, \quad \left. \frac{\partial \{\psi(r, z, T)\}}{\partial r} \right|_{r=b} = \frac{C_{ox}}{\epsilon_{Cha}} [V_{GS}^* - \psi_s(z, T)] \quad (8)$$

where,  $V_{GS}^* = V_{GS} - V_{FB}$  is the effective gate voltage that governs the bending in the energy bands, and  $C_{OX} / V_{GS} / V_{FB}$  is the gate oxide capacitance per unit area, gate voltage, and flat-band voltage.  $\psi_s(z, T)$  denotes the channel potential at  $r=b$ .

$$(vii) \psi(r, z_2^+) = \psi(r, z_2^-) + V_{bi,2}, \quad \psi(r, z_3^+) = \psi(r, z_3^-) + V_{bi,3} \quad (9)$$

$$(viii) V_{bi,2} = \frac{k_B T}{q} \left\{ \ln \left( \frac{N_{Cha1} N_{Cha2}}{n_i^2(Si)} \right) \right\}, \quad V_{bi,3} = \frac{k_B T}{q} \left\{ \ln \left( \frac{N_{Cha2} N_{Cha3}}{n_i^2(Si)} \right) \right\} \quad (10)$$

where,  $V_{bi,1}$ ,  $V_{bi,2}$ ,  $V_{bi,3}$ , and  $V_{bi,4}$  denote the built-in potential barrier at  $z_1$ ,  $z_2$ ,  $z_3$ , and  $z_4$ , respectively.  $N_{Cha,1}$ ,  $N_{Cha,2}$ , and  $N_{Cha,3}$  denotes the channel doping in region A, B, and C, respectively. Using the above boundary conditions, it is possible to express the potential in terms of device parameters as given below,

$$\psi(r, z, T) = \psi_s(z, T) + \frac{C_{OX} (2ab - b^2) [V_{GS}^* - \psi_s(z, T)]}{2\epsilon_{Cha} (b - a)} + \frac{C_{OX} a [V_{GS}^* - \psi_s(z, T)]}{\epsilon_{Cha} (b - a)} + \frac{C_{OX} [V_{GS}^* - \psi_s(z, T)]}{2\epsilon_{Cha} (b - a)} \quad (11)$$

For incorporating the effect of temperature, different relations have been formulated as given below [76–80],

$$E_G(T) = E_{G_0} + E_{G_\alpha} \left( \frac{(300)^2}{300 + E_{G_\beta}} - \frac{T^2}{T + E_{G_\beta}} \right), \quad \phi_{Sil}(T) = \chi_{Sil} + \frac{E_G(T)}{2} - q\phi_F(N_{Cha}, T) \quad (12a)$$

$$n_i(T) = A_0 \left( \frac{T}{300} \right)^{\frac{3}{2}} \exp \left( \frac{-qE_G(T)}{2K_B T} \right), \quad \phi_F(N_{Cha}, T) = \frac{K_B T}{q} \ln \left( \frac{N_{Cha}}{n_i(T)} \right) \quad (12b)$$

$$V_{GS}'(T) = V_{GS} - V_{FB}(T), \quad V_{FB}(T) = \phi_M - \phi_{Sil}(T) - \left( \frac{qN_f}{C} \right) \quad (12c)$$

Values of constant  $A_0$ ,  $E_{G_0}$ ,  $E_{G_\alpha}$  and  $E_{G_\beta}$  for silicon and germanium are  $2.2717 \times 10^{19}$ , 1.1,  $4.73 \times 10^{-4}$ , 636 and  $0.759144 \times 10^{19}$ , 0.66,  $4.77 \times 10^{-4}$ , 235 respectively [76]. The rest of the symbols has the usual meaning.  $\phi_{Sil}/\phi_M$ ,  $\chi_{Sil}$ , and  $\phi_F$  denote the work function of channel/metal, electron affinity, and Fermi potential of channel, respectively. Putting (11) back in (1) yields a differential equation in terms of the surface channel potential  $\psi_s(z, T)$ ,

$$\mu^2 \frac{\partial^2 \psi_s(z, T)}{\partial z^2} + \theta = \psi_s(z, T) \quad (13)$$

The generalized form of  $\psi_s(z, T)$  in three different regions are:-

$$\text{Region A } (z_1 \leq z \leq z_2): \quad \psi_s(z, T) = g_1(T) e^{\frac{z}{\mu_1}} + h_1(T) e^{\frac{-z}{\mu_1}} + \sigma_1(T) \quad (14)$$

$$\mu_1 = \sqrt{\frac{\epsilon_{Cha1} b(b-a)}{C_{OX1} (2b-a)}}, \quad \sigma_1(T) = V_{GS1}'(T) + \frac{qN_{Cha1} b^2}{C_{OX1} (2b-a)} - \frac{aqN_{Cha1} b}{C_{OX1} (2b-a)}, \quad V_{FB1}(T) = \phi_{M1} - \phi_{Sil}(T) - \frac{qN_f}{C_2} \quad (15a)$$

$$C_1 = \frac{\epsilon_{OXK1}}{b \ln \left( 1 + \frac{t_1}{b} \right)}, \quad C_2 = \frac{\epsilon_{bio}}{b \ln \left( 1 + \frac{t_2}{b} \right)}, \quad C_3 = \frac{\epsilon_{OXK2}}{b \ln \left( 1 + \frac{t_3}{b} \right)}, \quad C_{OX1} = \frac{C_1 C_2 C_3}{C_1 C_2 + C_2 C_3 + C_1 C_3} \quad (15b)$$



$$\text{Region B } (z_2 \leq z \leq z_3): \psi_{s_2}(z, T) = g_2(T) e^{\frac{z}{\mu_2}} + h_2(T) e^{\frac{-z}{\mu_2}} + \sigma_2(T) \quad (16)$$

$$\mu_2 = \sqrt{\frac{\varepsilon_{Cha_2} b(b-a)}{C_{OX_2} (2b-a)}}, \sigma_2(T) = V'_{GS_2}(T) + \frac{qN_{Cha_2} b^2}{C_{OX_2} (2b-a)} - \frac{aqN_{Cha_2} b}{C_{OX_2} (2b-a)} \quad (17a)$$

$$C_{OX_2} = \frac{\varepsilon_{SiO_2}}{b \ln\left(1 + \frac{t_{net}}{b}\right)}, V_{FB_2}(T) = \phi_{M_2} - \phi_{SiL}, t_{net} = t_1 + \left[ \frac{\varepsilon_{K_1}}{\varepsilon_{K_2}} (t_2 + t_3) \right] \quad (17b)$$

$$\text{Region C } (z_3 \leq z \leq z_4): \psi_{s_3}(z, T) = g_3(T) e^{\frac{z}{\mu_3}} + h_3(T) e^{\frac{-z}{\mu_3}} + \sigma_3(T) \quad (18)$$

$$\mu_3 = \sqrt{\frac{\varepsilon_{Cha_3} b(b-a)}{C_{OX_3} (2b-a)}}, \sigma_3(T) = V'_{GS_3}(T) + \frac{qN_{Cha_3} b^2}{C_{OX_3} (2b-a)} - \frac{aqN_{Cha_3} b}{C_{OX_3} (2b-a)} \quad (19a)$$

$$C_{OX_3} = C_{OX_1}, V_{FB_3}(T) = \phi_{M_3} - \phi_{SiL} - \frac{qN_f}{C_2} \quad (19b)$$

$V_1$  ( $V_4$ ) is the potential at the source (drain) junction, respectively. The coefficients  $g_1$ ,  $h_1$ ,  $g_2$ ,  $h_2$ ,  $g_3$ , and  $h_3$  can be determined using the boundary conditions and can be expressed as,

$$g_1(T) = \frac{0.5}{\sinh\left(\frac{z_1 - z_2}{\mu_1}\right)} \left[ \sigma_1(T) \left( e^{\frac{-z_1}{\mu_1}} - e^{\frac{-z_2}{\mu_1}} \right) + V_1 e^{\frac{-z_2}{\mu_1}} - V_2 e^{\frac{-z_1}{\mu_1}} \right] \quad (20a)$$

$$h_1(T) = \frac{0.5}{\sinh\left(\frac{z_1 - z_2}{\mu_1}\right)} \left[ \sigma_1(T) \left( e^{\frac{z_1}{\mu_1}} - e^{\frac{z_2}{\mu_1}} \right) + V_2 e^{\frac{z_1}{\mu_1}} - V_1 e^{\frac{z_2}{\mu_1}} \right] \quad (20b)$$

$$g_2(T) = \frac{0.5}{\sinh\left(\frac{z_2 - z_3}{\mu_2}\right)} \left[ \sigma_2(T) \left( e^{\frac{-z_2}{\mu_2}} - e^{\frac{-z_3}{\mu_2}} \right) + (V_2 + V_{bi,2}) e^{\frac{-z_3}{\mu_2}} - V_3 e^{\frac{-z_2}{\mu_2}} \right] \quad (21a)$$

$$h_2(T) = \frac{0.5}{\sinh\left(\frac{z_2 - z_3}{\mu_2}\right)} \left[ \sigma_2(T) \left( e^{\frac{z_2}{\mu_2}} - e^{\frac{z_3}{\mu_2}} \right) + V_3 e^{\frac{z_2}{\mu_2}} - (V_2 + V_{bi,2}) e^{\frac{z_3}{\mu_2}} \right] \quad (21b)$$

$$g_3(T) = \frac{0.5}{\sinh\left(\frac{z_3 - z_4}{\mu_3}\right)} \left[ \sigma_3(T) \left( e^{\frac{-z_3}{\mu_3}} - e^{\frac{-z_4}{\mu_3}} \right) + (V_3 + V_{bi,3}) e^{\frac{-z_4}{\mu_3}} - V_4 e^{\frac{-z_3}{\mu_3}} \right] \quad (22a)$$

$$h_3(T) = \frac{0.5}{\sinh\left(\frac{z_3 - z_4}{\mu_3}\right)} \left[ \sigma_3(T) \left( e^{\frac{z_3}{\mu_3}} - e^{\frac{z_4}{\mu_3}} \right) + V_4 e^{\frac{z_3}{\mu_3}} - (V_3 + V_{bi,3}) e^{\frac{z_4}{\mu_3}} \right] \quad (22b)$$

Boundary conditions at  $z_2$  and  $z_3$  can be used to evaluate  $V_2$  and  $V_3$  due to the continuity of the electric field at  $z_2$  and  $z_3$ .

$$\left. \frac{\partial \psi_{s_1}(z, T)}{\partial z} \right|_{z=z_2^-} = \left. \frac{\partial \psi_{s_2}(z, T)}{\partial z} \right|_{z=z_2^+}, \left. \frac{\partial \psi_{s_2}(z, T)}{\partial z} \right|_{z=z_3^-} = \left. \frac{\partial \psi_{s_3}(z, T)}{\partial z} \right|_{z=z_3^+} \quad (23a)$$

$$V_2 = f_4(s f_3 - t f_2), V_3 = f_4(s f_2 - t f_1) \quad (23b)$$

Values of  $s$ ,  $t$ , and  $f_1$ - $f_4$  are mentioned in Appendix A.

### 3.2 Threshold voltage

A significant amount of conduction current will flow only after applying a gate voltage greater than the threshold voltage. When the  $V_{GS}$  equals  $V_t$ , the minimum channel potential is  $2\phi_F$ . Since drain-to-source voltage decreases while moving towards the source and hence, minimum channel potential occurs in region-A. The threshold voltage is solved by

substituting the value of minimum channel potential (obtained by differentiating the surface potential), and putting  $V_{GS}=V_t$  will yield a quadratic equation in terms of  $V_t$ .

$$V_t = \frac{-r_{12} + \sqrt{r_{12}^2 - 4r_{11}r_{13}}}{2r_{11}} \quad (24)$$

Values of  $r_{11}$ ,  $r_{12}$ , and  $r_{13}$  are mentioned in Appendix B.

### 3.3 Drain current and subthreshold swing

The drain current can be modelled for linear region in the three different regions as [28,81,82],

$$I_{Lin,A} = \frac{\pi C_{ox1} \mu_n (b-a)}{L_{Cha1}} \left[ (V_{GS} - V_t)(V_2 - V_1) - \frac{(V_2 - V_1)^2}{2} \right] \quad (25)$$

$$I_{Lin,B} = \frac{\pi C_{ox2} \mu_n (b-a)}{L_{Cha2}} \left[ (V_{GS} - V_t)(V_3 - V_P) - \frac{(V_3 - V_P)^2}{2} \right] \quad (26)$$

$$I_{Lin,C} = \frac{\pi C_{ox3} \mu_n (b-a)}{L_{Cha3}} \left[ (V_{GS} - V_t)(V_4 - V_Q) - \frac{(V_4 - V_Q)^2}{2} \right] \quad (27)$$

Appropriate boundary conditions can be used to calculate  $V_P$  and  $V_Q$  and can be expressed as,

$$V_P = V_{bi,2} + \psi_{S1}(r, z_2), \quad V_Q = V_{bi,3} + \psi_{S2}(r, z_3) \quad (28)$$

The drain current is calculated in the saturation region by replacing  $V_{DS}$  with  $V_{DS,Sat}$ .

$$V_{DS,Sat} = \frac{(V_{GS} - V_t)}{1 + \frac{\mu_{efld}(V_{GS} - V_t)}{(L_{Cha1} + L_{Cha2} + L_{Cha3})v_{Sat}}}, \quad \mu_{efld} = \frac{\mu_n}{\left\{1 - \zeta(V_{GS} - V_t)\right\} \left[1 + \Omega \frac{V_{DS}\mu_n}{(L_{Cha1} + L_{Cha2} + L_{Cha3})v_{Sat}}\right]} \quad (29)$$

$$\Omega = \left[ \frac{V_{DS}\mu_n}{(L_{Cha1} + L_{Cha2} + L_{Cha3})v_{Sat}} \right] \left[ 1.5 + \left\{ \frac{V_{DS}\mu_n}{(L_{Cha1} + L_{Cha2} + L_{Cha3})v_{Sat}} \right\} \right]^{-1} \quad (30)$$

where  $v_{Sat}$  and  $\mu_{efld}$  are the saturation velocity of electrons ( $v_{Sat}=1*10^7$  cm/s) and maximum low field mobility.  $\zeta$  is the fitting parameter (0.43), and  $\mu_n$  is the mobility of electrons.  $C_{OX1}/C_{OX2}/C_{OX3}$  is the oxide capacitance per unit area. Subthreshold swing describes the swing in gate voltage obtained when the drain current is changed by a decade and is calculated as given below [83]

$$SS = V_T \ln 10 \left\{ \left( \frac{\partial \psi(b, z, T)}{\partial V_{GS}} \right)^{-1} \right\} \bigg|_{z_{min}} \quad (31)$$

### 3.4 Sensitivity

Relative change in the device metric (M) has been considered for the investigation of the biosensor and is given below,

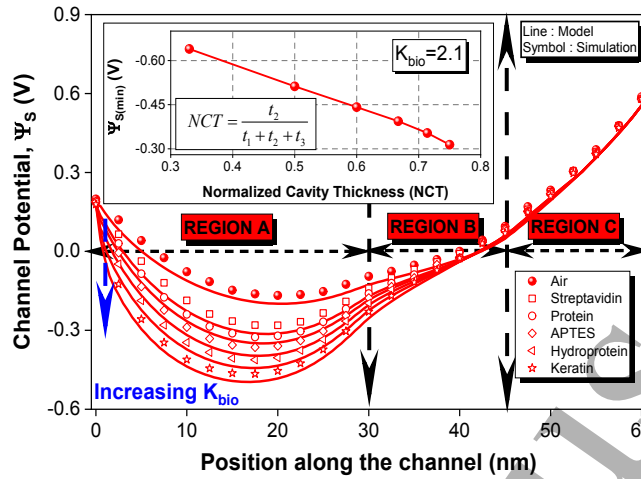
$$S_M = \left| \frac{M_{Air} - M_{Bio}}{M_{Bio}} \right| \quad (32)$$

where  $M_{Bio}/M_{Air}$  denotes the actual value of metric in the presence/absence of biomolecules and  $S_M$  is the sensitivity for the sensing metric 'M'.

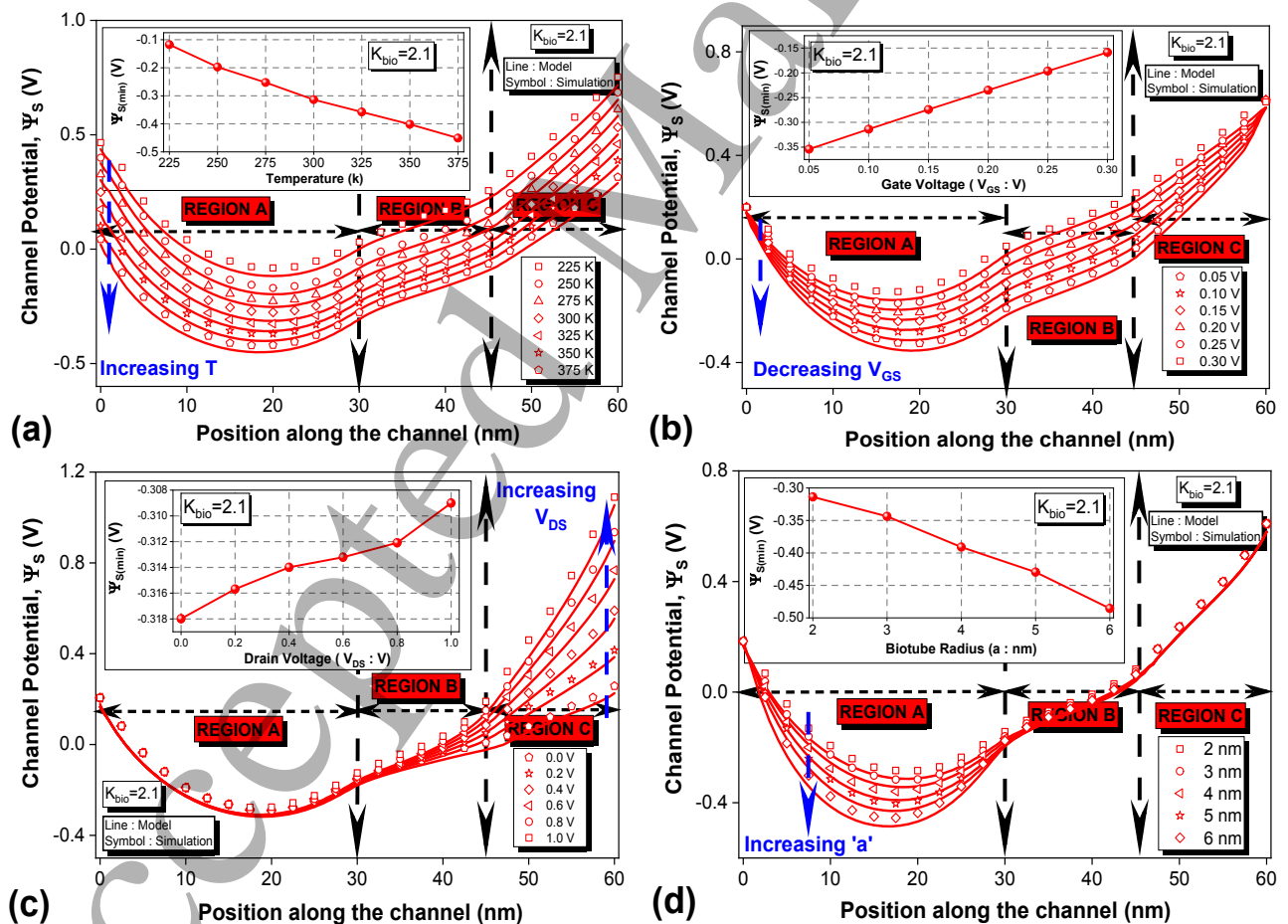
## 4. Results and discussion

When the biomolecules are completely entrapped inside the nanogap cavity, they change the effective gate oxide capacitance, which in turn changes the potential and electric field profile. Figure 3 shows the potential profile across the channel for different biomolecules, while its inset shows the minimum channel potential for different cavity thicknesses. The significant relative change in potential is an excellent qualitative indicator of the improved biosensing ability. A sharp change in the potential can be observed with the increasing dielectric constant of biomolecules. This sharp decrease in the minimum potential across the channel is due to the increased effective gate oxide capacitance. This improves

coupling and interaction between the charge carriers and the gate. Furthermore, the relative change in potential increases with the increasing  $K_{\text{bio}}$  [30]. The increase in the thickness of the cavity ( $t_{\text{CAV}}$ ) will increase the number of biomolecules that can be entrapped inside it. This increases the sensitivity of the biosensor. Hence, a significant change in the minimum channel potential can be seen with the changing  $t_{\text{CAV}}$ .



**Figure 3.** Channel potential profile for different  $K_{\text{bio}}$  ([INSET] shows the minimum channel potential for  $K_{\text{bio}} = 2.1$  at different normalized cavity thicknesses).



**Figure 4.** Channel potential profile for  $K_{\text{bio}} = 2.1$  at (a) different temperatures ([INSET] shows the minimum channel potential for  $K_{\text{bio}} = 2.1$  at different temperatures), (b) different  $V_{\text{GS}}$  ([INSET] shows the minimum channel potential for  $K_{\text{bio}} = 2.1$  at different temperature), (c) different  $V_{\text{DS}}$  ([INSET] shows the minimum channel potential for  $K_{\text{bio}} = 2.1$  at different temperature) and (d) different biotube inner radius 'a' ([INSET] shows the minimum channel potential for  $K_{\text{bio}} = 2.1$  at different 'a').

Figure 4(a) shows the potential profile across the channel for  $K_{\text{bio}}=2.1$  at different temperatures. Inset shows the minimum channel potential at different temperatures. The number of charge carriers increases with the increasing temperature. Hence, a narrow potential well can be seen at higher temperatures (minimum channel potential keeps on decreasing with the increasing temperature) [84]. Figure 4(b) shows the potential profile across the channel for  $K_{\text{bio}}=2.1$  at different gate voltages ( $V_{\text{GS}}$ ). Inset shows the minimum channel potential at different  $V_{\text{GS}}$  (step size of 50mV). At a fixed drain voltage ( $V_{\text{DS}}$ ), the subthreshold current increases with the increasing  $V_{\text{GS}}$ . This improvement in the subthreshold conduction is due to the decrease in the potential across the channel, and hence, minimum channel potential also decreases (magnitude) with the increasing gate voltage [84]. It is worth interesting to note that the curvature of the potential curve starts decreasing with increasing  $V_{\text{GS}}$  because high gate voltage increases the vertical electric field by a significant amount. Figure 4(c) shows the channel potential profile for  $K_{\text{bio}}=2.1$  at different drain voltages ( $V_{\text{DS}}$ ). Inset shows the minimum channel potential at different  $V_{\text{DS}}$ . Changing  $V_{\text{DS}}$  has a negligible effect on the channel potential variation (especially region A). Increasing  $V_{\text{DS}}$  will only change the potential barrier at the drain-channel junction; hence, the potential rises near the drain only [84]. It is interesting to note that there is no variation in the potential in region A (near the source) because the potential barrier at the source-channel junction [54] is almost unchanged. Hence, minimum channel potential also infinitesimally increases with the increasing drain voltage. Figure 4(d) shows the potential profile for  $K_{\text{bio}}=2.1$  at different inner radius 'a' of the biotube. Inset shows the minimum channel potential at different values of 'a'. With the increasing 'a', the cross-sectional area available for the conducting charge carriers decreases. This increases the threshold voltage of the biosensor due to increasing steepness in the channel potential profile; hence, minimum channel potential also increases (magnitude) with the increasing 'a'.

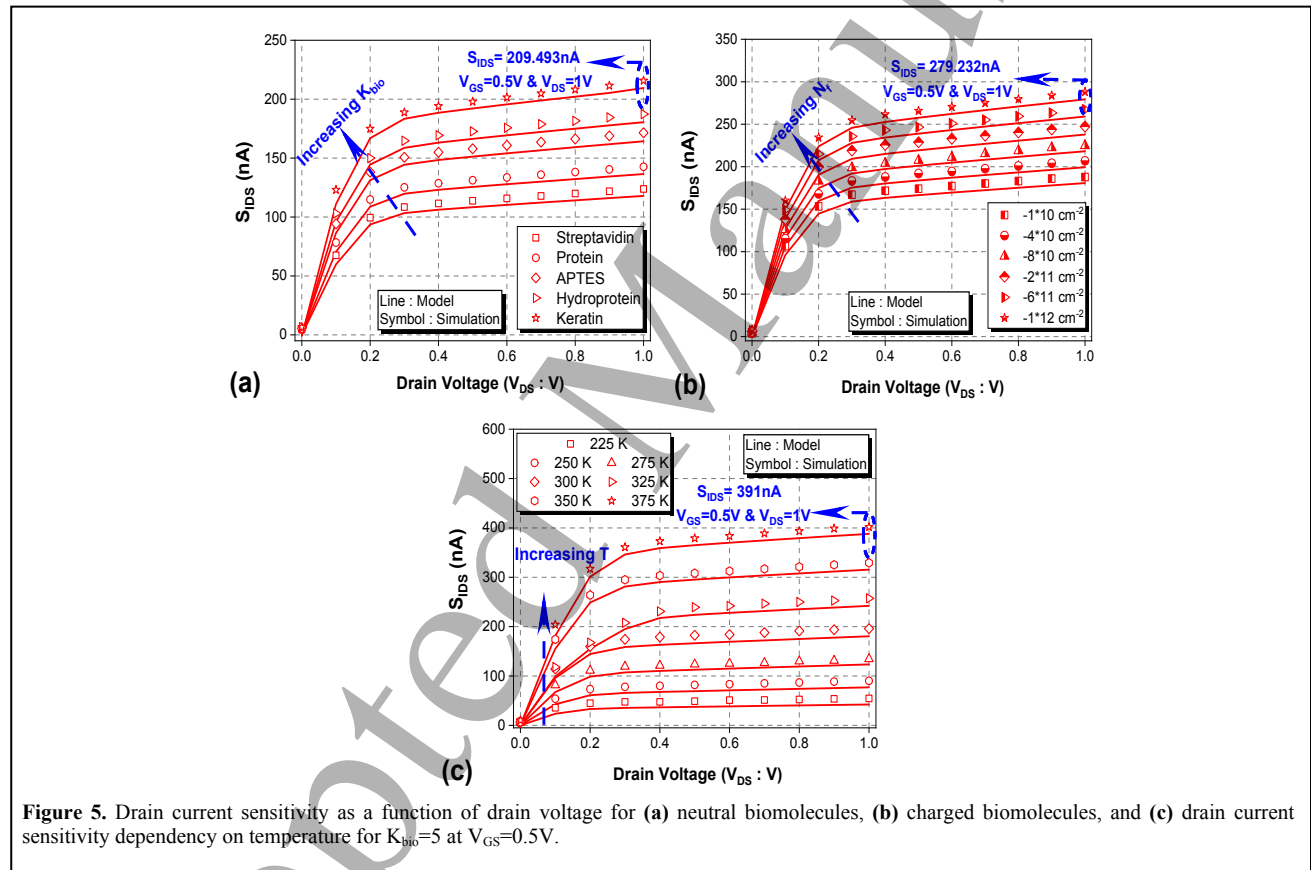
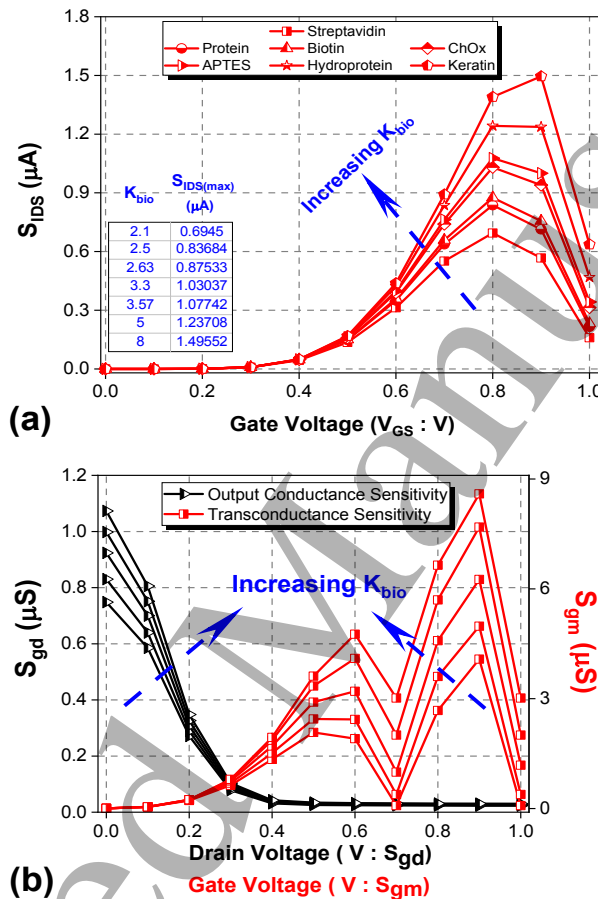


Figure 5 shows the drain current sensitivity ( $S_{\text{IDS}}$ ) as a function of  $V_{\text{DS}}$ . Drain current sensitivity increases with the increasing  $K_{\text{bio}}$ . The biomolecules, when entrapped inside the cavity, will increase the capacitance of the cavity. The increase in the oxide capacitance results in a stronger interaction between the gate and the flow of the charge carriers. This decreases the drain current but the change in drain current increases for the 'unfilled cavity case' [30]. Hence, the drain current sensitivity is high at higher values of  $K_{\text{bio}}$  which can be seen in figure 5(a). The presence of negatively charged biomolecules results in an increased interaction that accounts for the much stronger gate-to-channel coupling than the neutral biomolecules. Hence,  $S_{\text{IDS}}$  increases with the increasing charge density of biomolecules, as shown in figure 5(b). The negatively charged biomolecules decrease the channel potential by a factor of  $qN_f/C$  ( $N_f$  - charge density of biomolecules) in the nanogap cavity [30] that further decreases the drain current, but the change in drain current increases for the 'unfilled cavity case'. Figure 5(c) shows the impact of temperature on drain current sensitivity ( $S_{\text{IDS}}$ ) as

a function of  $V_{DS}$  for  $K_{bio}=5$ . As the temperature rises, a large number of charge carriers are generated, which increases the current flowing across the channel [85]. Hence,  $S_{IDS}$  increases with increasing temperature. For biosensing applications, operating the device in the ohmic region is advisable, which allows comparatively fast switching and sensing. The maximum  $S_{IDS}$  obtained here is roughly 391 nA for  $K_{bio}=5$  at  $T=375K$ .

Germanium has a comparatively larger intrinsic charge carrier density than silicon, due to which it offers a large number of intrinsic charge carriers that participates in the conduction process [9] and are also responsible for the biosensing action in DM-SGGS-BTFS. The intrinsic charge carrier density in germanium is approximately  $\sim 1650$  times larger than the silicon at room temperature. When the concentration of charge carriers is high, change in sensing metrics becomes more pronounced as the flow of many charge carriers is affected when different biomolecules localize inside the cavity [10]. Hence, the sensitivity of DM-SGGS-BTFS increases.



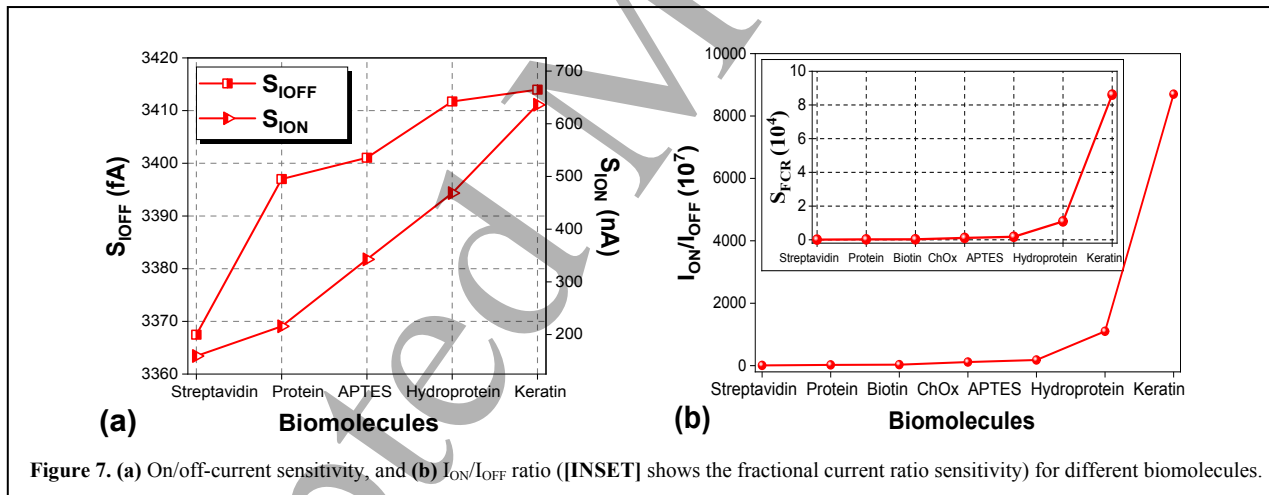
**Figure 6.** (a) Drain current sensitivity as a function of gate voltage, and (b) transconductance (output conductance) sensitivity as a function of gate voltage (drain voltage).

Figure 6(a) shows the drain current sensitivity as a non-linear implicit function of gate voltage. In contrast, figure 6(b) shows the transconductance/output conductance sensitivity ( $S_{gm}/S_{gd}$ ) for different biomolecules at a fixed value of  $V_{DS}/V_{GS}$ . At a lower value of gate voltage, the potential across the channel is nearly constant, but at higher  $V_{GS}$ , the potential distribution is not constant (unevenly distributed). It can be inferred that sensitivity is almost constant at lower values of gate voltage, whereas the peak of the drain current sensitivity ( $S_{IDS}=1.49552\mu A$ ) is obtained roughly at  $V_{GS}=0.9V$  and  $V_{DS}=0.5V$  for  $K_{bio}=8$ . Further, decreasing the value of  $V_{DS}$  will increase the threshold voltage sensitivity and drain current sensitivity. Interestingly, the biosensor is more sensitive to biomolecules at high  $V_{GS}$ , and  $S_{IDS}$  increases with increasing  $K_{bio}$  of the biomolecules. Transconductance is the derivative of drain current obtained from the transfer characteristics of the DM-SGGS-BTFS at a fixed drain voltage value [85]. The slope of the drain current in the  $I_{DS}-V_{GS}$  plot is higher at high  $V_{GS}$ , and it increases with the increasing  $K_{bio}$ , indicating high trans-conductance sensitivity ( $S_{gm}=8.5844\mu S$ ) is obtained for  $K_{bio}=8$  roughly at  $V_{GS}=0.9V$  and  $V_{DS}=0.5V$ . We have also used transconductance in determining the sensitivity of the proposed biosensor. Further, the peak transconductance ( $g_{mp}$ ) can be potentially used as a sensing metric ( $S_{gmp}$ ) that will increase the reliability of the biosensor. The relative change (compared to the 'unfilled cavity case') in the value of output conductance is high at lower values of drain voltage. This is because the output

conductance is calculated from the slope of the  $I_{DS}$ - $V_{DS}$  curve, and the slope is higher at low  $V_{DS}$  (in the ohmic region), which keeps on decreasing with the increasing  $V_{DS}$ . This slope tends to zero as the drain voltage increases at a fixed gate voltage value. The maximum output conductance ( $S_{gd}=1.073 \mu S$ ) is obtained for  $K_{bio}=8$  at  $V_{GS}=0.5$  V and  $V_{DS}=0$  V.

**Table 5.** Sensitivity variation in DM-SGGS-BTFS at different channel doping.

	STEP-GRADED CHANNEL DOPING (cm <sup>-3</sup> )					BIOMOLECULES						
						S <sub>V<sub>t</sub></sub> : Threshold Voltage Sensitivity (mV)						
						S <sub>V<sub>t</sub></sub> =  V <sub>t</sub> (without _bio) – V <sub>t</sub> (with _bio)						
N <sub>S</sub>	N <sub>Cha,1</sub>	N <sub>Cha,2</sub>	N <sub>Cha,3</sub>	N <sub>D</sub>	B <sub>1</sub>	B <sub>2</sub>	B <sub>3</sub>	B <sub>4</sub>	B <sub>5</sub>	B <sub>6</sub>	B <sub>7</sub>	
a.	10 <sup>19</sup>	10 <sup>10</sup>	10 <sup>11</sup>	10 <sup>12</sup>	10 <sup>19</sup>	134.629	155.303	160.676	181.603	187.771	209.007	228.094
b.		10 <sup>10</sup>	10 <sup>12</sup>	10 <sup>14</sup>		134.413	155.056	160.423	181.318	187.479	208.688	227.755
c.		10 <sup>10</sup>	10 <sup>14</sup>	10 <sup>18</sup>		137.484	158.385	163.811	184.908	191.117	212.456	231.596
d.		10 <sup>14</sup>	10 <sup>15</sup>	10 <sup>16</sup>		134.135	154.768	160.132	181.025	187.186	208.413	227.523
e.		10 <sup>16</sup>	10 <sup>17</sup>	10 <sup>18</sup>		134.583	155.121	160.458	181.237	187.361	208.441	227.404
f.		10 <sup>14</sup>	10 <sup>16</sup>	10 <sup>18</sup>		137.276	158.17	163.595	184.694	190.906	212.273	231.464
g.		10 <sup>12</sup>	10 <sup>11</sup>	10 <sup>10</sup>		134.742	155.445	160.827	181.796	187.966	209.241	228.371
h.		10 <sup>14</sup>	10 <sup>12</sup>	10 <sup>10</sup>		134.635	155.337	160.72	181.684	187.865	209.159	228.32
i.		10 <sup>18</sup>	10 <sup>14</sup>	10 <sup>10</sup>		46.51	54.367	56.46	64.857	67.423	76.69	85.865
j.		10 <sup>16</sup>	10 <sup>15</sup>	10 <sup>14</sup>		131.771	152.404	157.768	178.661	184.822	206.049	225.159
k.		10 <sup>18</sup>	10 <sup>17</sup>	10 <sup>16</sup>		45.773	53.527	55.593	63.89	66.427	75.597	84.69
l.		10 <sup>18</sup>	10 <sup>16</sup>	10 <sup>14</sup>		46.411	54.251	56.338	64.717	67.278	76.525	85.682
m.		10 <sup>10</sup>	10 <sup>10</sup>	10 <sup>10</sup>		134.851	155.558	160.941	181.901	188.077	209.34	228.448
n.		10 <sup>12</sup>	10 <sup>12</sup>	10 <sup>12</sup>		134.567	155.189	160.582	181.511	187.870	208.002	227.979
o.		10 <sup>14</sup>	10 <sup>14</sup>	10 <sup>14</sup>		134.234	154.876	160.242	181.147	187.312	207.555	227.678
p.		10 <sup>16</sup>	10 <sup>16</sup>	10 <sup>16</sup>		65.983	79.984	83.007	97.459	100.76	111.171	121.873
q.	10 <sup>18</sup>	10 <sup>18</sup>	10 <sup>18</sup>	39.681	46.582	48.432	55.908	58.213	66.611	75.065		
a-f: Forward tri-step graded profile (FTSGP)							g-l: Reverse tri-step graded profile (RTSGP)					
m-q: Non-step graded profile (Mono-doped profile)												



**Figure 7.** (a) On/off-current sensitivity, and (b)  $I_{ON}/I_{OFF}$  ratio ([INSET] shows the fractional current ratio sensitivity) for different biomolecules.

Using high doping in germanium increases the tunneling probability [8] and facilitates enhanced transportation of charge carriers which is due to the low energy band gap of germanium. As a result, this changes and alters the potential barrier at the source junction, which is bias and electric field dependent). The dependence of different device characteristics (threshold voltage, subthreshold swing, and drain current) on the source-channel potential barrier makes the germanium source biosensors more sensitive to biomolecules.

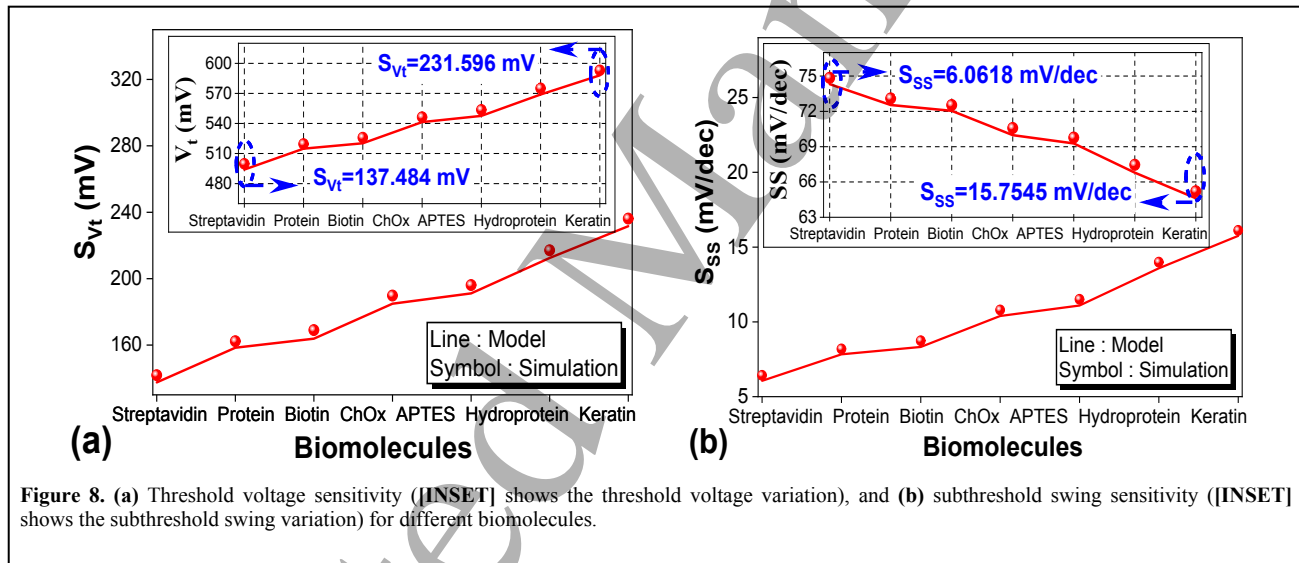
The graded channel lowers the short channel effects (SCEs) and improves the drive current of the device without scaling it [8],[86]. Using the tri-step graded doping causes a sudden change of potential across the channel, and this sharp change induces a larger electric field than the non-step graded doping [87]. This critical electric field imparts a more vital force on the charge carriers moving from the source to the drain, which improves the carrier transportation mechanism in a graded channel [88]. Using a step graded profile instead of non-step graded profile in channel makes the biosensor more sensitive towards biomolecules. Table 5 shows the sensitivity variation for different channel doping profiles which clearly indicates that the doping of channel must increase gradually while moving from source towards drain to obtain



high sensitivity. Forward tri step graded profile (FTSGP) yields comparatively high sensitivity as compared to reverse step graded profile (RTSGP), and this sensitivity increases further as the step doping gradient ( $|N_{\text{Cha},2} - N_{\text{Cha},1}|$  &  $|N_{\text{Cha},3} - N_{\text{Cha},2}|$ ) is increased. Here, the biosensor has shown maximum sensitivity at  $N_{\text{Cha},1}=10^{10}/\text{cm}^3$ ,  $N_{\text{Cha},2}=10^{14}/\text{cm}^3$  and  $N_{\text{Cha},3}=10^{18}/\text{cm}^3$  respectively.

Figure 7(a) shows the on-current sensitivity ( $S_{\text{ION}}$ ) and off-current sensitivity ( $S_{\text{IOFF}}$ ) for different biomolecules. The increasing  $K_{\text{bio}}$  of biomolecules decreases both the on-current and off-current [89], but the relative change increases due to increased gate-to-channel coupling. Hence,  $S_{\text{ION}}$  [90] and  $S_{\text{IOFF}}$  increase with the increasing  $K_{\text{bio}}$ . Figure 7(b) shows the  $I_{\text{ON}}/I_{\text{OFF}}$  ratio, while the inset shows the fractional sensitivity in  $I_{\text{ON}}/I_{\text{OFF}}$  ratio ( $S_{\text{FCR}}$ ) [91]. As both the on-current and off-current decrease, the overall  $I_{\text{ON}}/I_{\text{OFF}}$  ratio increases due to the dominant factor in the denominator (rate of decrease is more in off-current, thus increasing the overall  $I_{\text{ON}}/I_{\text{OFF}}$  ratio with the increasing  $K_{\text{bio}}$ ) [92]. Hence,  $I_{\text{ON}}/I_{\text{OFF}}$  ratio and its fractional sensitivity ( $S_{\text{FCR}}$ ) can also be used as a sensing metric along with threshold voltage and subthreshold swing.  $I_{\text{ON}}/I_{\text{OFF}}$  ratio changes approximately  $10^1$ - $10^4$  times when the cavity gets filled with biomolecules having  $K_{\text{bio}}=2.1$ -8. A larger  $I_{\text{ON}}/I_{\text{OFF}}$  ratio variation makes it a perfect sensing metric for biosensing applications. Using multiple sensing metrics to analyze the sensitivity of a biosensor increases its reliability.

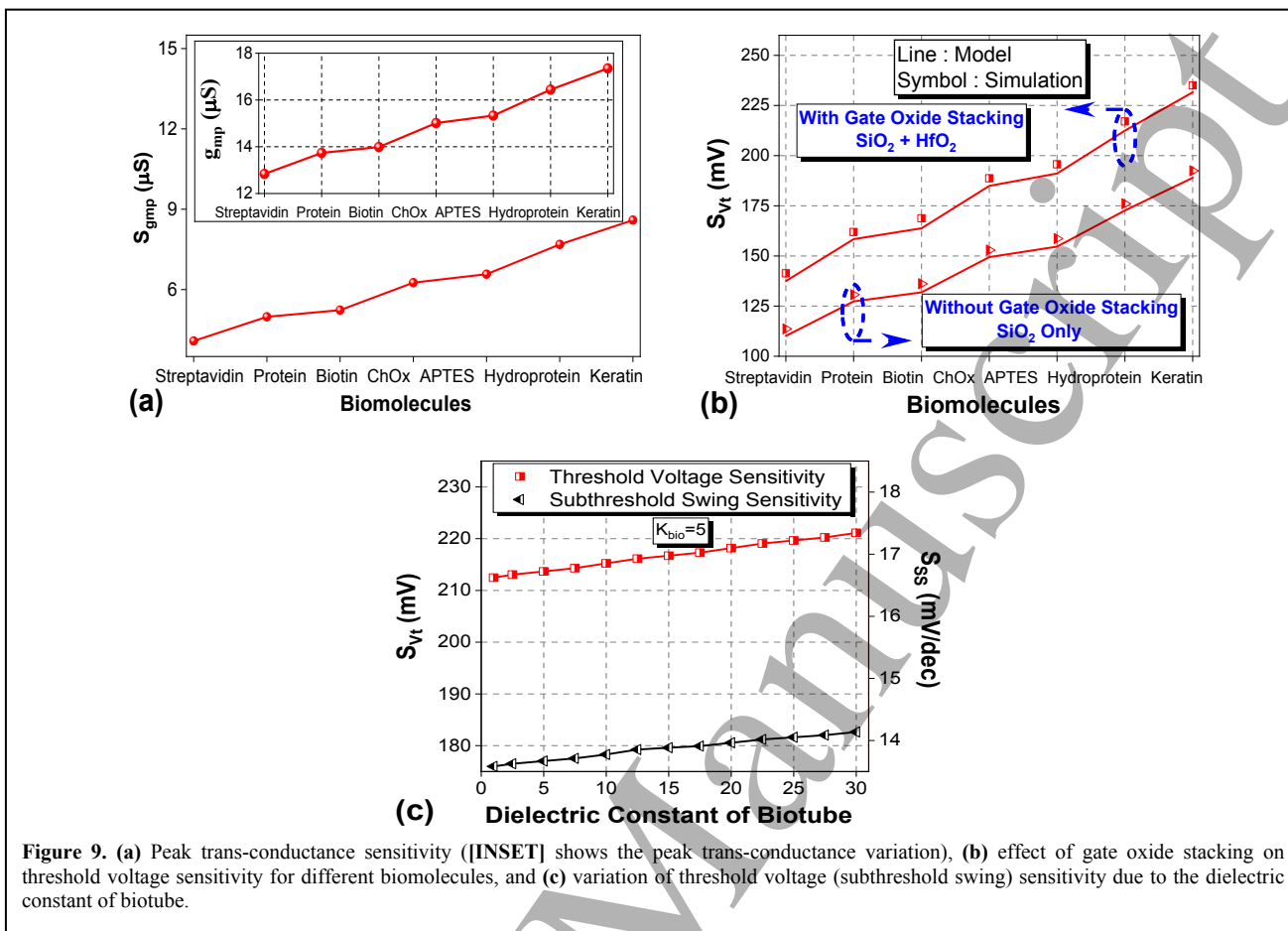
Figures 8(a) and 8(b) show the threshold voltage and subthreshold swing sensitivity for different biomolecules. The inset in Fig. 8 shows the variation of threshold voltage and subthreshold swing, respectively. With the increasing  $K_{\text{bio}}$ , more gate voltage will be required to start the conduction in the device because of the increased capacitance of the nanogap cavity (which leads to high capacitive coupling). This decreases the current flow but increases the threshold voltage. As a result, threshold voltage sensitivity increases due to the increase in the threshold voltage, which can be seen in figure 8(a). Subthreshold swing varies inversely with the gate oxide capacitance, and as a result, subthreshold swing decreases. However, the relative change (with respect to the 'unfilled cavity case') increases with the increasing  $K_{\text{bio}}$  [90], which is visible in figure 8(b). Threshold voltage has shown a more significant variation than the subthreshold swing for the same biomolecules and hence is a better sensing metric.



**Figure 8.** (a) Threshold voltage sensitivity ([INSET] shows the threshold voltage variation), and (b) subthreshold swing sensitivity ([INSET] shows the subthreshold swing variation) for different biomolecules.

Figure 9(a) shows the peak transconductance sensitivity ( $S_{\text{gmp}}$ ), while the inset shows the variation of peak transconductance for different biomolecules. The transconductance is an intrinsic parameter obtained from the  $I_{\text{DS}}-V_{\text{GS}}$  plot. As discussed above, transconductance is high at higher values of  $V_{\text{GS}}$ , and hence, the peak transconductance sensitivity is also obtained at high  $V_{\text{GS}}$ . The peak transconductance ( $g_{\text{mp}}=17.3507\mu\text{S}$ ) and peak transconductance sensitivity ( $S_{\text{gmp}}=8.5944\mu\text{S}$ ) is obtained for keratin at  $V_{\text{GS}}=0.9\text{ V}$  and  $V_{\text{DS}}=0.5\text{ V}$ . Gate oxide stacking increases the threshold voltage sensitivity by a substantial amount due to the increased gate oxide capacitance (capacitive coupling)[30]. Gate oxide stacking not only improves the sensitivity but also improves the electrostatics integrity of the device. Figure 9(b) shows the effect of the gate stack on threshold voltage sensitivity. A threshold voltage sensitivity improvement of approximately 42mV can be seen for  $K_{\text{bio}}=8$  due to gate oxide stacking ( $\text{SiO}_2+\text{HfO}_2$ ) which increases the coupling between the gate and the channel when compared to mono gate oxide ( $\text{SiO}_2$ ). We have considered the biotube to be filled with air only, but filling the biotube with other dielectric materials has a negligible effect on the threshold voltage and subthreshold swing. Threshold voltage sensitivity and subthreshold swing sensitivity changes merely 0.00866 V and 0.000537 V/decade when the dielectric constant of the material inside the biotube is changed by 30 times which can be seen in figure 9(c). Therefore, its effect can be neglected (the percentage change is merely below 1% for  $V_t$ ) in a practical scenario. This minute change is due to the leakage of charge carriers arising due to the non-ideality of

the dielectric material used in the biotube. More over, this change can further decrease (negligible) by using an ideal dielectric material inside the biotube.



**Figure 9.** (a) Peak trans-conductance sensitivity ([INSET] shows the peak trans-conductance variation), (b) effect of gate oxide stacking on threshold voltage sensitivity for different biomolecules, and (c) variation of threshold voltage (subthreshold swing) sensitivity due to the dielectric constant of biotube.

Table 6 shows the effect of the fill-in factor, along with the location of biomolecules, on the sensitivity. The fill-in factor implies the fraction of volume within the cavity occupied by the biomolecules. In practical cases, some portion of the cavity might remain unfilled for unavoidable reasons [6]. Hence, studying the impact of the fill-in factor becomes indispensable. It is worth noticing that the sensitivity of a biosensor mainly depends upon two prime factors: i) the location of biomolecules and ii) the fill-in factor of biomolecules. A high fill-in factor implies more biomolecules occupying the cavity which will change the gate oxide capacitance by a larger margin [93]. This changes the electric field and the potential profile across the channel. Hence, the lateral electric field will affect the flow of the charge carriers more strongly. Due to this, the device current changes relatively more [6]. Furthermore, the change in threshold voltage, subthreshold swing, and transconductance will increase if the fill-in factor is high.

Generally, a high fill-in factor of biomolecules results in high sensitivity. However, a high fill-in factor does not always guarantees high sensitivity, which can be seen in different cases. It can be seen that the sensitivity is high when the biomolecules are located near the source. The barrier potential generated across the source junction (between source and channel) mainly affects the device conduction. This potential hill across the source junction is mainly affected by the biomolecules near the source. When the biomolecules are entrapped in the cavity near the source, the potential barrier between the source and the channel is primarily affected. The change in the potential barrier across the source junction is predominantly responsible for a larger change in threshold voltage, transconductance, and subthreshold swing. The charge carriers generated from the source are involved in the biosensing action, and thus the sensitivity is affected by the fill-in factor and location of biomolecules. While the fill-in factor of biomolecules can impact sensitivity, the position of the biomolecules can have a more significant and dominating effect. The fill-in factor in case (3) is higher than in case (4), but the sensitivity relation is opposite to that of the fill-in factor. Similarly, even though case (1) and case (4) have the same fill-in factor but the sensitivity is multiple times larger in case (4) than in case (1), which strengthens the observation that the position of the biomolecules plays an important role in determining sensitivity.

Figure 10 shows the threshold voltage sensitivity plot of DM-SGGS-BTFS at different cavity thickness. Notably, the sensitivity pattern demonstrates a positive correlation with the increasing thickness of the cavity. It is worth noting that, as the cavity thickness increases, the ability of the gate to regulate the flow of charge carriers gradually decreases due to

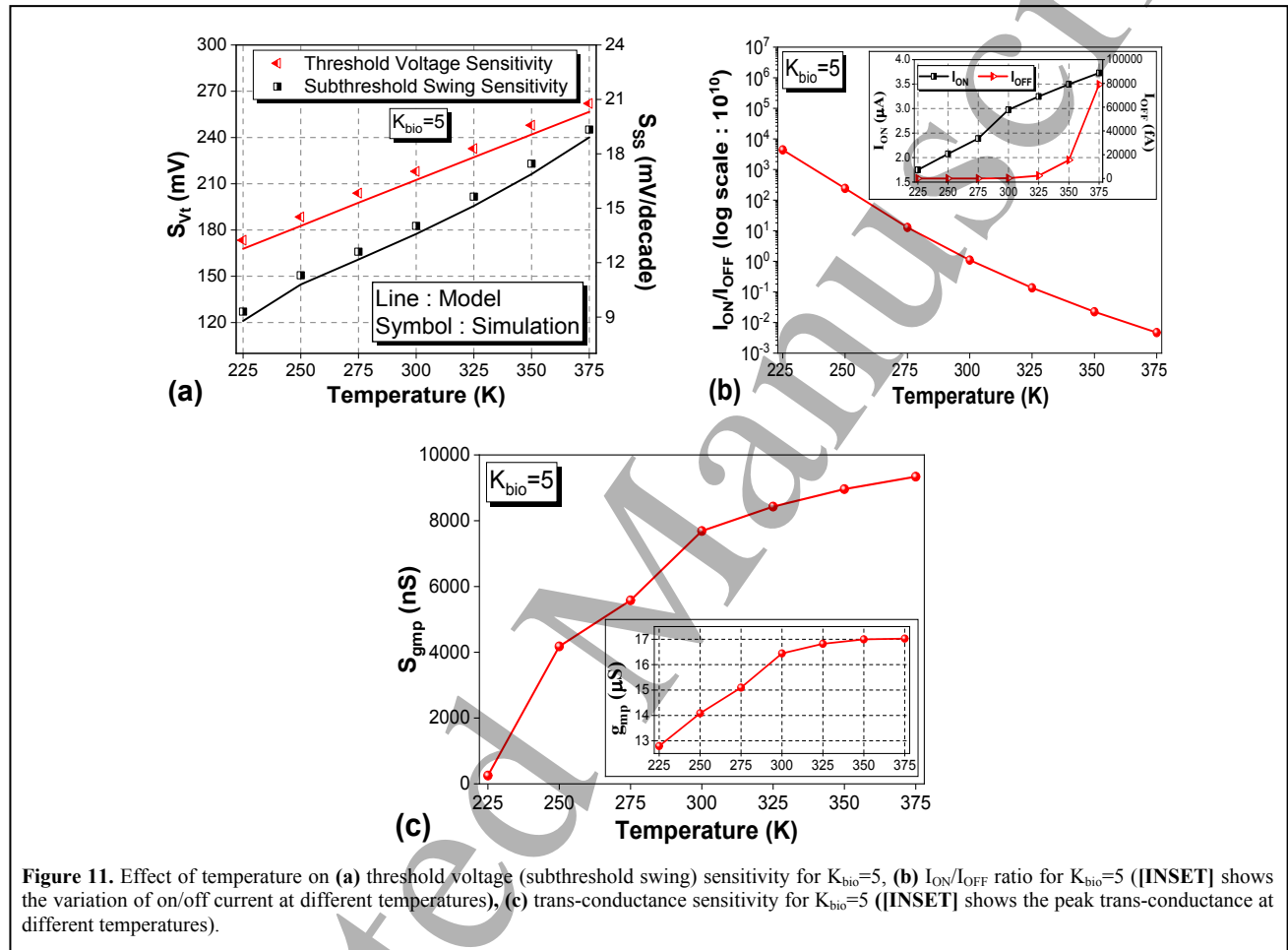
Figure 10 is a line graph showing the relationship between Cavity Thickness (nm) on the x-axis and Threshold Voltage ( $S_{vt} : V$ ) on the y-axis. The x-axis ranges from 6 to 20 nm, and the y-axis ranges from 0.21 to 0.33 V. A red line represents the model, and red symbols represent simulation data. The threshold voltage decreases as cavity thickness increases. A box in the graph specifies  $K_{bio} = 5$ . The legend indicates 'Line : Model' and 'Symbol : Simulation'.

Cavity Thickness (nm)	Threshold Voltage ( $S_{vt} : V$ )
6	0.215
8	0.245
10	0.265
12	0.280
14	0.290
16	0.295
18	0.298
20	0.300

**Table 6.** Effect of fill-in factor and location of different biomolecules on  $S_{Vt}$ ,  $S_{SS}$ , and  $S_{gmp}$ .

18

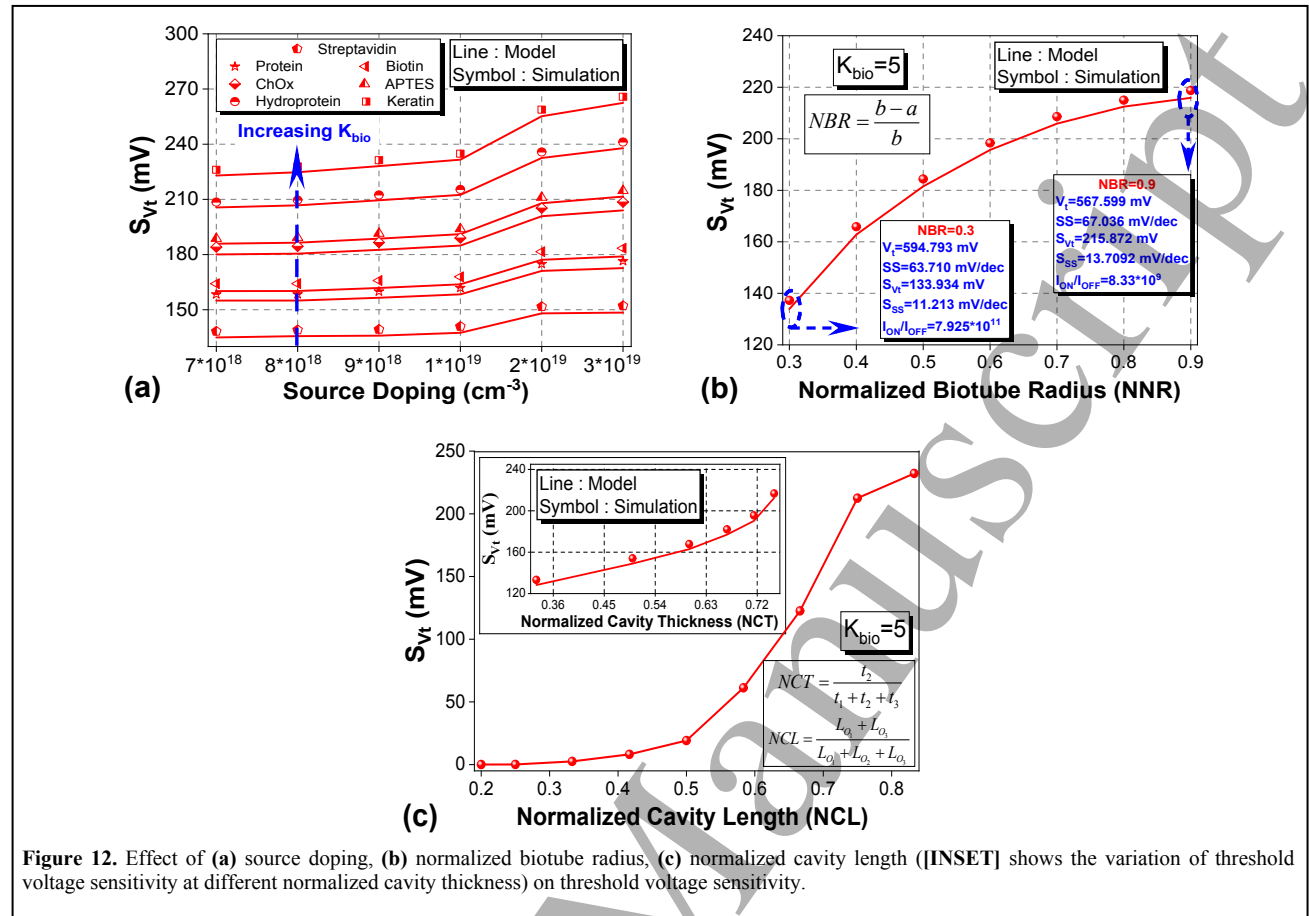
Figures 11(a) and 11(b) show the effect of temperature on threshold voltage/subthreshold swing sensitivity and  $I_{ON}/I_{OFF}$  ratio for  $K_{bio}=5$ . As the temperature increases, large numbers of charge carriers are generated due to which conduction starts at a lower value of gate voltage, and hence threshold voltage decreases [92]. But, the relative change in threshold voltage increases with increasing temperature, and hence the threshold voltage sensitivity increases. Due to a substantial increase in current with the rise in temperature, the swing in the drain current also increases, increasing the subthreshold swing sensitivity. This can be seen in figure 11(a). Both the on-current and off-current increase with an increase in temperature [92], but the rate of change of off-current dominates the  $I_{ON}/I_{OFF}$  ratio. Hence,  $I_{ON}/I_{OFF}$  ratio degrades badly at high temperatures, as seen in figure 11(b). The peak transconductance is directly proportional to the slope of the drain current in the transfer characteristic plot, and the slope increases with an increase in temperature. Therefore, as the temperature increases, the peak transconductance also increases, and the peak transconductance sensitivity also increases accordingly. This trend can be observed in figure 11(c).



**Figure 11.** Effect of temperature on (a) threshold voltage (subthreshold swing) sensitivity for  $K_{bio}=5$ , (b)  $I_{ON}/I_{OFF}$  ratio for  $K_{bio}=5$  ([INSET] shows the variation of on/off current at different temperatures), (c) trans-conductance sensitivity for  $K_{bio}=5$  ([INSET] shows the peak trans-conductance at different temperatures).

Figures 12(a), 12(b), and 12(c) show the effect of source doping, normalized biotube radius (NBR), and normalized cavity length/thickness on threshold voltage sensitivity. An increase in the source doping will increase the number of charge carriers that brings the nano-device into a conduction state at a low  $V_{GS}$  [55,94]. This increases the threshold voltage sensitivity, and the biosensor becomes more sensitive to biomolecules, as seen in figure 12(a). The increasing inner radius of the biotube ('a') will decrease the cross-sectional area of the channel available for the charge carriers, and hence, the conduction current decreases. This will increase the threshold voltage; hence, the relative change in threshold voltage increases with the increasing biotube radius, as seen in figure 12(b). It can be seen that at lower values of 'a' ( $a=1\text{nm}$ :  $NBR=0.9$ ), threshold voltage sensitivity is high, but the current ratio is comparatively low. Optimizing the value of 'a' become essential due to the tradeoff between sensitivity and  $I_{ON}/I_{OFF}$  ratio. As discussed above, the increasing thickness or length of the cavity will entrap more biomolecules inside it due to its increased volume. Hence, the potential and electric field profile changes relatively more, accounting for the more considerable threshold voltage change. So, threshold voltage sensitivity increases with the increasing cavity length or thickness [30], as seen in figure 12(c). Sensitivity in the label-free biosensors varies directly with the cavity dimensions, i.e., the more is the dimension, the

more its sensitivity will be [28,30]. Table 7 highlights the sensitivity comparison of DM-SGGS-BTFS with existing biosensors.



**Table 7.** Benchmarking the performance of DM-SGGS-BTFS with similar biosensors.

Parameter	Our Work	Reference [95]	Reference [96]	Reference [30]	Reference [97]	Reference [98]	Reference [99]
	DM-SGGS-BTFS	DM-DG-JL-MOSFET	GaN-GME-DE-SNW-FET	JLGSSRG	QG-MC-MOSFET	TGAA-NWFET	TGRC-MOSFET
Gate oxide	$\text{SiO}_2 + \text{HfO}_2$	$\text{SiO}_2 + \text{TiO}_2$	$\text{Al}_2\text{O}_3 + \text{HfO}_2$	$\text{SiO}_2 + \text{HfO}_2$	$\text{SiO}_2 + \text{Si}_3\text{N}_4$	$\text{SiO}_2 + \text{HfO}_2$	$\text{SiO}_2$
Cavity length (nm)	29	25	15	20	20	10	8
Cavity thickness (nm)	6	9	4	5	9	1	-
Channel length (nm)	60	100	50	50	40	20	40
Type of cavity	Two-sided	Two-sided	Two-sided	Two-sided	Two-sided	One-sided	Two-sided
$S_{vt}$ (V)	0.262	0.227	0.105	0.066	0.161	0.0172	0.0072
$K_{bio}$	8	10	8	12	12	2.1	8
Advantage	Very high sensitivity	High sensitivity	Low power	Fabrication simplicity	Low dependence of sensitivity on temperature	Low $I_{OFF}$	Low power
Disadvantage	Fabrication complexity	High threshold voltage	Fabrication complexity	Low sensitivity	Fabrication complexity	Very low sensitivity	Very low sensitivity

Selectivity is one of the most common problems in DM-FET based biosensors (designed without any bioreceptor layer). Hence, this issue has also been addressed here by discussing the merits/demerits of adding an additional bioreceptor layer. For this, the Anti-Iris antibody has been selected as the biomolecule, while the Thiol Linker and Iris Antigen act as the bioreceptor. The respective thicknesses of Thiol linker and Iris Antigen are 2 nm and 4 nm,



respectively. Although different papers typically consider a thickness of 2 nm for the Anti-Iris antibody, the thickness of the cavity (excluding the bioreceptor thickness) has been increased to 6 nm to accommodate the biomolecule effectively (considering a fill-in factor of 1) [7,9]. The Thiol Linker functions as a silica-binding protein, while the Iris Antigen serves as the bioreceptor for detecting the Anti-Iris (biotarget) layer. These layers possess individual dielectric constants of 2, 4, and 8, respectively. To provide a clear reference, figure 13 visually illustrates the arrangement of the three-layered stack (bioreceptor and biotarget). Within figure 13, the current sensitivity plot (transfer characteristics) for different cases (corresponding to the consecutive stacking of the Thiol Linker, Iris Antigen, and Anti-Iris Antibody) in DM-SGGS-BTFS is presented. In the absence of biomolecules, the drain current remains low. However, as the Thiol Linker, Iris Antigen, and Anti-Iris gradually accumulate within the cavity, the current increases significantly due to the strong electrostatic control of the gate over the channel because of strong capacitive coupling/effect. It can be also seen that threshold voltage sensitivity is higher for the case when no bioreceptor layer is used as compared to the case when bioreceptor layer is used (total oxide thickness, i.e.  $t_2$  is kept same in both the cases). This indicates the existence of a trade-off between sensitivity and selectivity in the proposed device. In non-critical primary sensing applications, the biosensor could be designed without an additional bioreceptor layer, reducing costs and simplifying the manufacturing process (the selectivity/specificity of the biochemical species could be further validated through secondary testing). The inclusion of a bioreceptor layer becomes pertinent when selectivity takes precedence in the sensing application.

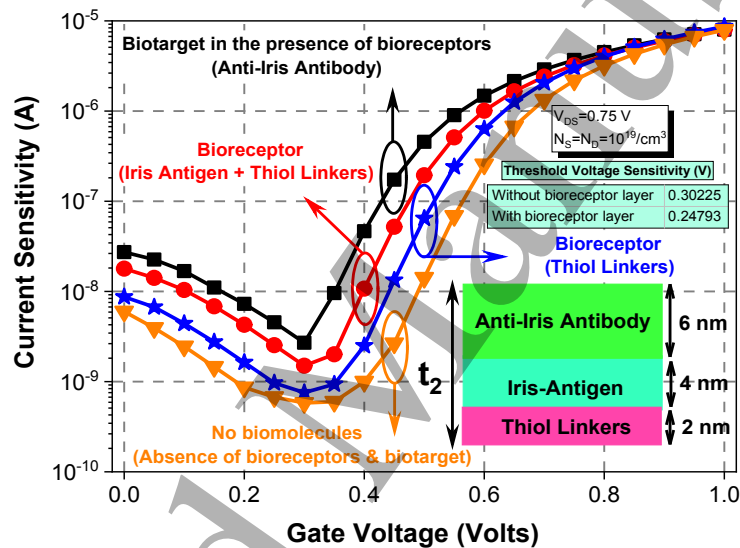


Figure 13. Current sensitivity plot (transfer characteristics) in DM-SGGS-BTFS with successive accumulation of the bioreceptor and biotarget.

#### 4. Conclusion

This paper thoroughly investigates and analyzes the biosensing performance of a novel DM-SGGS-BTFS. The improved biosensing performance is attributed to the use of a germanium source and step-graded biotube, which enhances the biosensing performance and electrostatic integrity of short-channel devices. The study extensively explores the effect of the fill-in factor and the location of biomolecules, demonstrating that the gate controls the flow of charge carriers effectively through capacitive coupling. Notably, the location of biomolecules has a more significant impact on sensitivity than the fill-in factor. Various metrics were used to evaluate the biosensing performance, including subthreshold swing, off-current, threshold voltage, peak trans-conductance,  $I_{ON}/I_{OFF}$  ratio, and on-current. The threshold voltage and current sensitivity reached remarkable values of up to 263 mV and 1495.52 nA, respectively, indicating excellent biosensing ability. Furthermore, the study discussed the effect of doping and bias on sensitivity and emphasized the need to optimize different parameters to achieve high sensitivity. For further enhancing the biosensing performance and sensitivity of DM-SGGS-BTFS, future investigations could explore implementing Gaussian/Double-Gaussian doping profile in the nanotube channel, drain engineering techniques using different drain materials, or introducing source/drain pockets. Overall, DM-SGGS-BTFS has the potential to serve as a viable alternative to existing silicon biosensors and can find widespread use in the biomedical industry.



## ACKNOWLEDGEMENT

The authors are grateful to the Director, Maharaja Agrasen Institute of Technology, Delhi for providing the facilities to carry out this research work. One of the authors (Amit Das) would like to thank UGC, Government of India for financial assistance to carry out this research work.

## APPENDIX-A

$$M_0(z,T) = \psi_S(z,T) + \frac{C_{OX}(2ab-b^2)[V_{GS}^* - \psi_S(z,T)]}{2\varepsilon_{Cha}(b-a)}, \quad M_1(z,T) = -\frac{C_{OX}a[V_{GS}^* - \psi_S(z,T)]}{\varepsilon_{Cha}(b-a)}, \quad M_2(z,T) = \frac{C_{OX}[V_{GS}^* - \psi_S(z,T)]}{2\varepsilon_{Cha}(b-a)}$$

$$f_3 = 2 \left[ \frac{\beta}{\mu_2} \cosh\left(\frac{z_2 - z_3}{\mu_2}\right) + \frac{\gamma}{\mu_3} \cosh\left(\frac{z_3 - z_4}{\mu_3}\right) \right], \quad s = \frac{\alpha}{\mu_1} s_a - \frac{\beta}{\mu_2} s_b, \quad f_1 = 2 \left[ \frac{\alpha}{\mu_1} \cosh\left(\frac{z_1 - z_2}{\mu_1}\right) + \frac{\beta}{\mu_2} \cosh\left(\frac{z_2 - z_3}{\mu_2}\right) \right], \quad f_2 = \frac{2\beta}{\mu_2}, \quad t = \frac{\gamma}{\mu_3} t_a - \frac{\beta}{\mu_2} t_b$$

$$s_a = \left[ \sigma_1 e^{\mu_1} \left( e^{\frac{z_2}{\mu_1}} - e^{\frac{z_1}{\mu_1}} \right) - \sigma_1 e^{\mu_1} \left( e^{\frac{z_2}{\mu_1}} - e^{\frac{z_1}{\mu_1}} \right) + 2V_{bi,1} \right], \quad s_b = \left[ \sigma_2 e^{\mu_2} \left( e^{\frac{z_2}{\mu_2}} - e^{\frac{z_3}{\mu_2}} \right) - \sigma_2 e^{\mu_2} \left( e^{\frac{z_2}{\mu_2}} - e^{\frac{z_3}{\mu_2}} \right) + 2 \cosh\left(\frac{z_2 - z_3}{\mu_2}\right) V_{bi,2} \right]$$

$$t_a = \left[ \sigma_3 e^{\mu_3} \left( e^{\frac{z_3}{\mu_3}} - e^{\frac{z_4}{\mu_3}} \right) - 2V_{DS} - 2V_{bi,4} - \sigma_3 e^{\mu_3} \left( e^{\frac{z_3}{\mu_3}} - e^{\frac{z_4}{\mu_3}} \right) + 2 \cosh\left(\frac{z_3 - z_4}{\mu_3}\right) V_{bi,3} \right], \quad t_b = \left[ \sigma_2 e^{\mu_2} \left( e^{\frac{z_2}{\mu_2}} - e^{\frac{z_3}{\mu_2}} \right) - \sigma_2 e^{\mu_2} \left( e^{\frac{z_2}{\mu_2}} - e^{\frac{z_3}{\mu_2}} \right) + 2V_{bi,2} \right], \quad f_4 = (f_2^2 - f_1 f_3)^{-1}$$

## APPENDIX-B

$$\eta_{11} = 4p_{11}^{-1}, \quad \eta_{12} = 4p_{22} + 2\xi_1 + 4\phi_F, \quad \xi_1 = \frac{qN_{Cha1}ab}{C_{ox1}(2b-a)} - \frac{qN_{Cha1}b^2}{C_{ox1}(2b-a)} + V_{FB1}, \quad \eta_{13} = 4p_{33} - 4\phi_F^2 - \xi_1^2 - 4\phi_F\xi_1, \quad p_{11} = p_{33} = p_1 p_3, \quad p_{22} = p_1 + p_3$$

$$\phi_F = V_T \ln\left(\frac{N_{Cha1}}{n_i(Si)}\right), \quad \alpha = \frac{1}{2 \sinh\left(\frac{z_1 - z_2}{\mu_1}\right)}, \quad \xi_3 = \frac{qN_{Cha3}ab}{C_{ox3}(2b-a)} - \frac{qN_{Cha3}b^2}{C_{ox3}(2b-a)} + V_{FB3}, \quad \xi_2 = \frac{qN_{Cha2}ab}{C_{ox2}(2b-a)} - \frac{qN_{Cha2}b^2}{C_{ox2}(2b-a)} + V_{FB2}$$

$$m_1 = \frac{\alpha}{\mu_1} \left[ \frac{z_2}{\mu_1} \left( e^{\frac{z_1}{\mu_1}} - e^{\frac{z_2}{\mu_1}} \right) - \frac{z_2}{\mu_1} \left( e^{\frac{z_2}{\mu_1}} - e^{\frac{z_1}{\mu_1}} \right) \right], \quad m_2 = \frac{\beta}{\mu_2} \left[ \frac{z_2}{\mu_2} \left( e^{\frac{z_2}{\mu_2}} - e^{\frac{z_3}{\mu_2}} \right) - \frac{z_2}{\mu_2} \left( e^{\frac{z_3}{\mu_2}} - e^{\frac{z_2}{\mu_2}} \right) \right], \quad m_3 = \frac{2\beta}{\mu_2} \cosh\left(\frac{z_2 - z_3}{\mu_2}\right) V_{bi2} - \frac{2\alpha}{\mu_1} V_{bi1}, \quad p_2 = \alpha n_1, \quad p_4 = \alpha n_2, \quad V_T = \frac{k_B T}{q}$$

$$p_1 = \alpha \left[ \frac{z_1}{\mu_1} \left( e^{\frac{z_1}{\mu_1}} - e^{\frac{z_2}{\mu_1}} \right) - \frac{z_1}{\mu_1} f_3 f_4 (m_1 - m_2) - e^{\mu_1} f_2 f_4 (m_4 - m_5) \right], \quad p_3 = \alpha \left[ \frac{z_1}{\mu_1} \left( e^{\frac{z_1}{\mu_1}} - e^{\frac{z_2}{\mu_1}} \right) + e^{\mu_1} f_3 f_4 (m_1 - m_2) - e^{\mu_1} f_2 f_4 (m_4 - m_5) \right], \quad m_5 = \frac{\beta}{\mu_2} \left[ \frac{z_3}{\mu_2} \left( e^{\frac{z_2}{\mu_2}} - e^{\frac{z_3}{\mu_2}} \right) - \frac{z_3}{\mu_2} \left( e^{\frac{z_3}{\mu_2}} - e^{\frac{z_2}{\mu_2}} \right) \right]$$

$$m_1 = \left[ \xi_1 e^{\mu_1} + V_{bi,1} e^{\mu_1} - e^{\mu_1} e_3 e_4 (-m_1 \xi_1 + m_2 \xi_2 - m_3) + e^{\mu_1} e_2 e_4 (-m_4 \xi_3 + m_5 \xi_2 - m_6) - \xi_1 e^{\mu_1} \right], \quad m_6 = \frac{2\beta}{\mu_2} V_{bi2} - \frac{2\gamma}{\mu_3} \cosh\left(\frac{z_3 - z_4}{\mu_3}\right) V_{bi3} - \frac{2(V_{DS} + V_{bi4})}{\mu_3} V_{bi2}$$

$$n_2 = \left[ \frac{z_1}{\mu_1} f_3 f_4 (-m_1 \xi_1 + m_2 \xi_2 - m_3) - \xi_1 \left( e^{\mu_1} - e^{\mu_1} \right) - V_{bi,1} e^{\mu_1} - e^{\mu_1} f_2 f_4 (-m_4 \xi_3 + m_5 \xi_2 - m_6) \right], \quad m_4 = \frac{\gamma}{\mu_3} \left[ \frac{z_3}{\mu_3} \left( e^{\frac{z_3}{\mu_3}} - e^{\frac{z_4}{\mu_3}} \right) - \frac{z_3}{\mu_3} \left( e^{\frac{z_4}{\mu_3}} - e^{\frac{z_3}{\mu_3}} \right) \right]$$

$$\xi_2 = \frac{qN_{Cha2}ab}{C_{ox2}(2b-a)} - \frac{qN_{Cha2}b^2}{C_{ox2}(2b-a)} + V_{FB2}, \quad \xi_3 = \frac{qN_{Cha3}ab}{C_{ox3}(2b-a)} - \frac{qN_{Cha3}b^2}{C_{ox3}(2b-a)} + V_{FB3}$$

## STATEMENTS & DECLARATIONS

### CONFLICT OF INTEREST

There are no conflicts of interest amongst the authors.

### AUTHOR CONTRIBUTIONS

Amit Das, Sonam Rewari, Binod Kumar Kanaujia, S.S. Deswal and R.S. Gupta have contributed mutually regarding this paper.

### AVAILABILITY OF DATA AND MATERIALS

Not Applicable.

### FUNDING

Not applicable.

### CONSENT TO PARTICIPATE

All the authors have complete consent to participate.

### CONSENT FOR PUBLICATION

All the authors have complete consent for publication.

### COMPLIANCE WITH ETHICAL STANDARDS

\* Disclosure of potential conflicts of interest

There are no conflicts of interest amongst the authors.

\* Research involving Human Participants and/or Animals

No animals or human beings were harmed during this research.

\* INFORMED CONSENT

All the authors have informed consent.

## References

- [1] Das A, Kanauija B K, Nath V, Rewari S and Gupta R S 2020 Impact of Reverse Gate Oxide Stacking on Gate All around Tunnel FET for High Frequency Analog and RF Applications *2020 IEEE 17th India Counc. Int. Conf. INDICON 2020* 1–6
- [2] Sharma S, Goel A, Rewari S, Vandana N and Gupta R S 2022 Enhanced Analog Performance and High-Frequency Applications of Dielectric Engineered High-K Schottky Nanowire FET *Silicon*
- [3] D. Singh, B. S. Sengar, P. Dwivedi, and V. Garg, "Comparative analysis of gate structure dependent FET-based biosensor," *Mater. Today Commun.*, p. 106301, 2023, doi: <https://doi.org/10.1016/j.mtcomm.2023.106301>
- [4] Im H, Huang X J, Gu B and Choi Y K 2007 A dielectric-modulated field-effect transistor for biosensing *Nat. Nanotechnol.* **2** 430–4
- [5] Wadhwa T, Kakkar D, Wadhwa G and Raj B 2019 Recent Advances and Progress in Development of the Field Effect Transistor Biosensor: A Review *J. Electron. Mater.* **48** 7635–7646
- [6] Das A, Rewari S, Kanauija B K and Gupta R S 2022 Recent Technological Advancement in Surrounding Gate MOSFET for Biosensing Applications - a Synoptic Study *Silicon* **14** 5133–5143
- [7] Dwivedi P and Kranti A 2018 Dielectric Modulated Biosensor Architecture: Tunneling or Accumulation Based Transistor? *IEEE Sens. J.* **18** 3228–35
- [8] Lee K W, Choi S J, Ahn J H, Moon D Il, Park T J, Lee S Y and Choi Y K 2010 An underlap field-effect transistor for electrical detection of influenza *Appl. Phys. Lett.* **96** 033703
- [9] Tang X, Jonas A M, Nysten B, Demoustier-Champagne S, Blondeau F, Prévot P P, Pampin R, Godfroid E, Iñiguez B, Colinge J P, Raskin J P, Flandre D and Bayot V 2009 Direct protein detection with a nano-interdigitated array gate MOSFET *Biosens. Bioelectron.* **24** 3531–7
- [10] Chung I Y, Jang H, Lee J, Moon H, Seo S M and Kim D H 2012 Simulation study on discrete charge effects of SiNW biosensors according to bound target position using a 3D TCAD simulator *Nanotechnology* **23**
- [11] Dwivedi P and Singh R 2020 Investigation the impact of the gate work-function and biases on the sensing metrics of TFET based biosensors *Eng. Res. Express* **2** 025043
- [12] Sarkar D and Banerjee K 2012 Proposal for tunnel-field-effect-transistor as ultra-sensitive and label-free biosensors *Appl. Phys. Lett.* **100** 143108
- [13] Dwivedi P, Singh R, Sengar B S, Kumar A and Garg V 2021 A New Simulation Approach of Transient Response to Enhance the Selectivity and Sensitivity in Tunneling Field Effect Transistor-Based Biosensor *IEEE Sens. J.* **21** 3201–9
- [14] Dwivedi P and Kranti A 2018 Overcoming Biomolecule Location-Dependent Sensitivity Degradation Through Point and Line Tunneling in Dielectric Modulated Biosensors *IEEE Sens. J.* **18** 9604–11
- [15] Goyal P, Srivastava G, Rewari S and Gupta R S 2020 Controlling Ambipolarity and Rising Ion in TFETs for Enhanced Reliability: A Review *2020 5th IEEE Int. Conf. Recent Adv. Innov. Eng. ICRAIE 2020 - Proceeding* 1–6
- [16] Goyal P, Srivastava G, Madan J, Pandey R and Gupta R S 2022 Source material valuation of charge plasma based DG-TFET for RFIC applications *Semicond. Sci. Technol.* **37** 095023
- [17] Chen Y N, Fan M L, Hu V P H, Su P and Chuang C T 2014 Evaluation of stability, performance of ultra-low voltage MOSFET, TFET, and Mixed TFET-MOSFET SRAM cell with write-assist circuits *IEEE J. Emerg. Sel. Top. Circuits Syst.* **4** 389–99
- [18] Soni D, Sharma D, Aslam M and Yadav S 2018 Approach for the improvement of sensitivity and sensing speed of TFET-based biosensor by using plasma formation concept *Micro Nano Lett.* **13** 1728–33
- [19] Avci U E, Rios R, Kuhn K J and Young I A 2011 Comparison of power and performance for the TFET and MOSFET and considerations for P-TFET *Proc. IEEE Conf. Nanotechnol.* 869–72
- [20] Strangio S, Settino F, Palestini P, Lanuzza M, Crupi F, Esseni D and Selmi L 2018 Digital and analog TFET circuits: Design and benchmark *Solid. State. Electron.* **146** 50–65
- [21] Reddy N N and Panda D K 2021 A Comprehensive Review on Tunnel Field-Effect Transistor (TFET) Based Biosensors: Recent Advances and Future Prospects on Device Structure and Sensitivity *Silicon* **13** 3085–100
- [22] Hwang M T, Heiranian M, Kim Y, You S, Leem J, Taqieddin A, Faramarzi V, Jing Y, Park I, van der Zande A M, Nam S, Aluru N R and Bashir R 2020 Ultrasensitive detection of nucleic acids using deformed graphene channel field effect biosensors *Nat. Commun.* **11** 1543
- [23] Sarkar D, Liu W, Xie X, Anselmo A C, Mitragotri S and Banerjee K 2014 MoS<sub>2</sub> field-effect transistor for next-generation label-free biosensors *ACS Nano* **8** 3992–4003
- [24] Lee J, Jang J, Choi B, Yoon J, Kim J Y, Choi Y K, Myong Kim D, Hwan Kim D and Choi S J 2015 A Highly Responsive Silicon Nanowire/Amplifier MOSFET Hybrid Biosensor *Sci. Rep.* **5** 12286
- [25] Jakob M H, Dong B, Gutsch S, Chatelle C, Krishnaraja A, Weber W and Zacharias M 2017 Label-free SnO<sub>2</sub> nanowire FET biosensor for protein detection *Nanotechnology* **28** 1–27
- [26] Kaisti M, Kerko A, Aarikka E, Saviranta P, Boeva Z, Soukka T and Lehmusvuori A 2017 Real-Time wash-free detection of unlabeled PNA-DNA hybridization using discrete FET sensor *Sci. Rep.* **7** 1–9
- [27] Buitrago E, Fagas G, Badia M F B, Georgiev Y M, Berthomé M and Ionescu A M 2013 Junctionless silicon nanowire transistors for the tunable operation of a highly sensitive, low power sensor *Sensors Actuators, B Chem.* **183** 1–10
- [28] Pratap Y, Kumar M, Kabra S, Haldar S, Gupta R S and Gupta M 2018 Analytical modeling of gate-all-around junctionless transistor based biosensors for detection of neutral biomolecule species *J. Comput. Electron.* **17** 288–96
- [29] Nowbahari A, Roy A and Marchetti L 2020 Junctionless transistors: State-of-the-art *Electron.* **9** 1–22
- [30] Chakraborty A and Sarkar A 2017 Analytical modeling and sensitivity analysis of dielectric-modulated junctionless gate stack surrounding gate MOSFET (JLGSSRG) for application as biosensor *J. Comput. Electron.* **16** 556–67
- [31] Wang B, Huang W, Chi L, Al-Hashimi M, Marks T J and Facchetti A 2018 High- k Gate Dielectrics for Emerging Flexible and Stretchable Electronics *Chem. Rev.* **118** 5690–754
- [32] Lee J C, Chò H J, Katng C S, Rhee S, Kim Y H, Choi R, Kang C Y, Choi C and Abkar M 2003 High-K Dielectrics and MOSFET Characteristics *IEEE Int. Electron Devices Meet.* 4.4.1–4.4.4
- [33] Mech B C and Kumar J 2017 Effect of high-k dielectric on the performance of Si, InAs and CNT FET *Micro Nano Lett.* **12** 624–9
- [34] Wilk G D, Wallace R M and Anthony J M 2001 High-κ gate dielectrics: Current status and materials properties considerations *J. Appl. Phys.* **89** 5243–75
- [35] Kasturi P, Saxena M and Gupta R S 2005 Modeling and simulation of STacked Gate Oxide (STGO) architecture in Silicon-On-

- Nothing (SON) MOSFET *Solid. State. Electron.* **49** 1639–48
- [36] Gupta M and Hu V P-H 2021 Influence of Channel Doping on Junctionless and Negative Capacitance Junctionless Transistors *ECS J. Solid State Sci. Technol.* **10** 065009
- [37] Fallahnejad M, Vadzadeh M, Salehi A, Kashaniniya A and Razaghian F 2020 Impact of channel doping engineering on the high-frequency noise performance of junctionless In<sub>0.3</sub>Ga<sub>0.7</sub>As/GaAs FET: A numerical simulation study *Phys. E Low-Dimensional Syst. Nanostructures* **115** 113715
- [38] Kim R, Avci U E and Young I A 2015 Source/drain doping effects and performance analysis of ballistic III-V n-MOSFETs *IEEE J. Electron Devices Soc.* **3** 37–43
- [39] Mosfets S D G, Lin H, Member S and Taur Y 2017 Effect of Source – Drain Doping on Subthreshold Characteristics of Short-Channel DG MOSFETs *IEEE Trans. Electron Devices* **64** 4856–60
- [40] Curreli M, Zhang R, Ishikawa F N, Chang H K, Cote R J, Zhou C and Thompson M E 2008 Real-time, label-free detection of biological entities using nanowire-based FETs *IEEE Trans. Nanotechnol.* **7** 651–67
- [41] Syahir A, Usui K, Tomizaki K, Kajikawa K and Mihara H 2015 Label and Label-Free Detection Techniques for Protein Microarrays *Microarrays* **4** 228–44
- [42] Goel A, Rewari S, Verma S and Gupta R S 2018 Dielectric Modulated Triple Metal Gate All around MOSFET (TMGAA) for DNA Bio-Molecule Detection *Proc. Int. Conf. 2018 IEEE Electron Device Kolkata Conf. EDKCON 2018* 337–40
- [43] Wu H, Si M, Dong L, Gu J, Zhang J and Ye P D 2015 Germanium nMOSFETs with recessed channel and S/D: Contact, scalability, interface, and drain current exceeding 1 A/mm *IEEE Trans. Electron Devices* **62** 1419–26
- [44] Brunco D P, De Jaeger B, Eneman G, Mitard J, Hellings G, Satta A, Terzieva V, Souriau L, Leys F E, Pourtois G, Houssa M, Winderickx G, Vrancken E, Sioncke S, Opsomer K, Nicholas G, Caymax M, Stesmans A, Van Steenberghe J, Mertens P W, Meuris M and Heyns M M 2008 Germanium MOSFET Devices: Advances in Materials Understanding, Process Development, and Electrical Performance *J. Electrochem. Soc.* **155** H552
- [45] Saha R, Hirpara Y and Hoque S 2021 Sensitivity Analysis on Dielectric Modulated Ge-Source DMDG TFET Based Label-Free Biosensor *IEEE Trans. Nanotechnol.* **20** 552–60
- [46] Zhou J, Zhang C, Liu Q, You J, Zheng X, Cheng X and Jiang T 2020 Controllable all-optical modulation speed in hybrid silicon-germanium devices utilizing the electromagnetically induced transparency effect *Nanophotonics* **9** 2797–807
- [47] Lim B S, Arshad M K M, Othman N, Fathil M F M, Fatin M F and Hashim U 2014 The impact of channel doping in junctionless field effect transistor *IEEE Int. Conf. Semicond. Electron. Proceedings, ICSE* 112–4
- [48] Zhao H, Zhu F, Chen Y T, Yum J H, Wang Y and Lee J C 2009 Effect of channel doping concentration and thickness on device performance for In<sub>0.53</sub>Ga<sub>0.47</sub>As metal-oxide-semiconductor transistors with atomic-layer-deposited Al<sub>2</sub>O<sub>3</sub> dielectrics *Appl. Phys. Lett.* **94** 1–3
- [49] Singh K N and Dutta P K 2021 Analytical modeling of underlap graded channel field effect transistor as a label-free biosensor *Superlattices Microstruct.* **155** 106897
- [50] Kumar A, Bhusan S and Tiwari P K 2017 A Threshold Voltage Model of Quantum Confinement Effects *IEEE Trans. Nanobioscience* **16** 868–75
- [51] Pal A and Sarkar A 2014 Analytical study of Dual Material Surrounding Gate MOSFET to suppress short-channel effects (SCEs) *Eng. Sci. Technol. an Int. J.* **17** 205–12
- [52] Singh S, Yadav S and Bhalla S K 2022 An Improved Analytical Modeling and Simulation of Gate Stacked Linearly Graded Work Function Vertical TFET *Silicon* **14** 4647–60
- [53] Narang R, Saxena M and Gupta M 2017 Modeling of gate underlap junctionless double gate MOSFET as bio-sensor *Mater. Sci. Semicond. Process.* **71** 240–51
- [54] Kumari M, Singh N K, Sahoo M and Rahaman H 2021 Work function optimization for enhancement of sensitivity of dual-material (DM), double-gate (DG), junctionless MOSFET-based biosensor *Appl. Phys. A Mater. Sci. Process.* **127** 1–8
- [55] Das A, Kanaujia B K, Deswal S S, Rewari S and Gupta R S 2022 Doping induced threshold voltage and ION/IOFF ratio modulation in surrounding gate MOSFET for analog applications *2022 IEEE International Conference of Electron Devices Society Kolkata Chapter (EDKCON), Kolkata, India* pp 1–6
- [56] Das R, Chanda M and Sarkar C K 2018 Analytical Modeling of Charge Plasma-Based Optimized Nanogap Embedded Surrounding Gate MOSFET for Label-Free Biosensing *IEEE Trans. Electron Devices* **65** 5487–93
- [57] Kumar M, Halder S, Gupta M and Gupta R S 2017 Ambipolarity reduction in DMG asymmetric vacuum dielectric Schottky Barrier GAA MOSFET to improve hot carrier reliability *Superlattices Microstruct.* **111** 10–22
- [58] Busse S, Scheumann V, Menges B and Mittler S 2002 Sensitivity studies for specific binding reactions using the biotin/streptavidin system by evanescent optical methods *Biosens. Bioelectron.* **17** 704–10
- [59] Goel A, Rewari S, Verma S and Gupta R S 2020 Physics-based analytic modeling and simulation of gate-induced drain leakage and linearity assessment in dual-metal junctionless accumulation nano-tube FET (DM-JAM-TFET) *Appl. Phys. A Mater. Sci. Process.* **126** 1–14
- [60] Lodhi A, Rajan C, Kumar A, Dip B, Samajdar P and Soni D 2020 Sensitivity and sensing speed analysis of extended nano - cavity and source over electrode in Si / SiGe based TFET biosensor *Appl. Phys. A* **126** 1–8
- [61] Kim C H, Jung C, Park H G and Choi Y K 2009 Novel dielectric-modulated field-effect transistor for label-free DNA detection *Biochip J.* **2** 127–34
- [62] Ganesh A, Goel K, Mayall J S and Rewari S 2021 Subthreshold Analytical Model of Asymmetric Gate Stack Triple Metal Gate all Around MOSFET (AGSTMGA AFET) for Improved Analog Applications *Silicon* **14** 4063–4073
- [63] Li C, Liu F, Han R and Zhuang Y 2021 A Vertically Stacked Nanosheet Gate-All-Around FET for Biosensing Application *IEEE Access* **9** 63602–10
- [64] Zhang Y, Han K and Li J 2020 A simulation study of a gate-all-around nanowire transistor with a core-insulator *Micromachines* **11** 1–12
- [65] Ye S, Yamabe K and Endoh T 2021 Ultimate vertical gate-all-around metal-oxide-semiconductor field-effect transistor and its three-dimensional integrated circuits *Mater. Sci. Semicond. Process.* **134** 106046
- [66] Choi S J, Moon D II, Kim S, Duarte J P and Choi Y K 2011 Sensitivity of threshold voltage to nanowire width variation in junctionless transistors *IEEE Electron Device Lett.* **32** 125–7
- [67] Kumari M, Singh N K and Sahoo M 2022 A detailed investigation of dielectric-modulated dual-gate TMD FET based label-free biosensor via analytical modelling *Sci. Rep.* **12** 1–17

- [68] Narang R, Reddy K V S, Saxena M, Gupta R S and Gupta M 2012 A dielectric-modulated tunnel-FET-based biosensor for label-free detection: Analytical modeling study and sensitivity analysis *IEEE Trans. Electron Devices* **59** 2809–17
- [69] Mohanty S S, Mishra S, Mohapatra M and Mishra G P 2022 Hetero Channel Double Gate MOSFET for Label-free Biosensing Application *Silicon* **14** 8109–18
- [70] Goel A, Rewari S, Verma S, Deswal S S and Gupta R S 2021 Dielectric Modulated Junctionless Biotube FET (DM-JL-BT-FET) Bio-Sensor *IEEE Sens. J.* **21** 16731–43
- [71] Kim S, Ahn J H, Park T J, Lee S Y and Choi Y K 2009 A biomolecular detection method based on charge pumping in a nanogap embedded field-effect-transistor biosensor *Appl. Phys. Lett.* **94** 1–4
- [72] Das A, Rewari S, Kanaujia B K, Deswal S S and Gupta R S 2023 Ge/Si interfaced label free nanowire BIOFET for biomolecules detection-analytical analysis *Microelectronics J.* **138** 105832
- [73] Software D S 2018 ATLAS User's Manual Device Simulation Software Volume I **1** 318
- [74] Sze S M 2017 *VLSI Technology* (McGraw Hill Education)
- [75] Hafiz S A, Iltesha, Ehteshamuddin M and Loan S A 2019 Dielectrically Modulated Source-Engineered Charge-Plasma-Based Schottky-FET as a Label-Free Biosensor *IEEE Trans. Electron Devices* **66** 1905–10
- [76] Goel A, Rewari S, Verma S and Gupta R S 2019 Temperature-Dependent Gate-Induced Drain Leakages Assessment of Dual-Metal Nanowire Field-Effect Transistor - Analytical Model *IEEE Trans. Electron Devices* **66** 2437–45
- [77] Ravindra N M and Srivastava V K 1979 Temperature dependence of the energy gap in semiconductors *J. Phys. Chem. Solids* **40** 791–3
- [78] Misiakos K and Tsamakis D 1993 Accurate measurements of the silicon intrinsic carrier density from 78 to 340 K *J. Appl. Phys.* **74** 3293–7
- [79] Singh J 2005 *Smart Electronic Materials: Fundamentals and Applications* (Cambridge University Press)
- [80] Rewari S, Nath V, Halder S, Deswal S S and Gupta R S 2018 Gate-Induced Drain Leakage Reduction in Cylindrical Dual-Metal Hetero-Dielectric Gate All Around MOSFET *IEEE Trans. Electron Devices* **65** 3–10
- [81] Ajay, Narang R, Saxena M and Gupta M 2017 Modeling and simulation investigation of sensitivity of symmetric split gate junctionless fet for biosensing application *IEEE Sens. J.* **17** 4853–61
- [82] Das A, Rewari S, Kanaujia B K, Deswal S S and Gupta R S 2023 Numerical modeling of a dielectric modulated surrounding-triple-gate germanium-source MOSFET (DM-STGGS-MOSFET)-based biosensor *J. Comput. Electron.*
- [83] Jung H 2019 Analysis of subthreshold swing in symmetric junctionless double gate MOSFET using high-k gate oxides *Int. J. Electr. Electron. Eng. Telecommun.* **8** 334–9
- [84] Agarwal A, Tiwari R, Ranjan S, Pradhan P C and Swain B P 2018 2D Analytical Modeling of Surface Potential for GaAs based Nanowire Gate All Around MOSFET 2D Analytical Modeling of Surface Potential for GaAs based Nanowire Gate All Around MOSFET *IOP Conf. Ser. Sci. Eng.* 012103
- [85] Gaspari V, Fobelets K, Velazquez-Perez J E, Ferguson R, Michelakis K, Despotopoulos S and Papavassiliou C 2004 Effect of temperature on the transfer characteristic of a 0.5  $\mu\text{m}$ -gate Si/SiGe depletion-mode n-MOSFET *Appl. Surf. Sci.* **224** 390–3
- [86] Narendar V and Girdhardas K A 2018 Surface Potential Modeling of Graded-Channel Gate-Stack (GCGS) High-K Dielectric Dual-Material Double-Gate (DMDG) MOSFET and Analog/RF Performance Study *Silicon* **10** 2865–75
- [87] Das A, Rewari S, Kanaujia B K, Deswal S S and Gupta R S 2023 Analytical investigation of a triple surrounding gate germanium source metal – oxide – semiconductor field-effect transistor with step graded channel for biosensing applications *Int. J. Numer. Model. Electron. Networks, Devices Fields* 1–25
- [88] Pathak V and Saini G 2018 A Graded Channel Dual-Material Gate Junctionless MOSFET for Analog Applications *Procedia Comput. Sci.* 825–31
- [89] Shamim Sarker M, Mainul Islam M, Nur Kutubul Alam M and Rafiqul Islam M 2016 Gate dielectric strength dependent performance of CNT MOSFET and CNT TFET: A tight binding study *Results Phys.* **6** 879–83
- [90] Jana G, Sen D, Debnath P and Chanda M 2022 Power and delay analysis of dielectric modulated dual cavity Junctionless double gate field effect transistor based label-free biosensor *Comput. Electr. Eng.* **99** 107828
- [91] Bind M K and Nigam K 2022 Sensitivity Analysis of Junction Free Electrostatically Doped Tunnel-FET Based Biosensor *Silicon* **14** 7755–7767
- [92] Agha F N A K, Hashim Y and Shakib M N 2020 Temperature Impact on the ION/IOFF Ratio of Gate All around Nanowire TFET 2020 *IEEE Int. Conf. Semicond. Electron.* 1–4
- [93] Rahman E, Shadman A and Khosru Q D M 2017 Effect of biomolecule position and fill in factor on sensitivity of a Dielectric Modulated Double Gate Junctionless MOSFET biosensor *Sens. Bio-Sensing Res.* **13** 49–54
- [94] Das A, Rewari S, Kanaujia B K, Deswal S S and Gupta R S 2023 Analytical modeling and doping optimization for enhanced analog performance in a Ge / Si interfaced nanowire MOSFET *Phys. Scr.* **98** 74005
- [95] Kumari M, Singh N K, Sahoo M and Rahaman H 2022 2-D Analytical Modeling and Simulation of Dual Material, Double Gate, Gate Stack Engineered, Junctionless MOSFET based Biosensor with Enhanced Sensitivity *Silicon* **14** 4473–84
- [96] Sharma S, Nath V, Deswal S S and Gupta R S 2022 Analytical modelling and sensitivity analysis of Gallium Nitride-Gate Material and, dielectric engineered- Schottky nano-wire fet(GaN-GME-DE-SNW-fet) based label-free biosensor *Microelectronics J.* **129** 105599
- [97] Maiti S, De A and Sarkar S K 2022 Analytical Modelling of Symmetric Gate Underlap Quadruple Gate Multichannel Junctionless MOSFET Biosensor *Silicon* **14** 6921–32
- [98] Getnet M and Chaujar R 2022 Sensitivity Analysis of Biomolecule Nanocavity Immobilization in a Dielectric Modulated Triple-Hybrid Metal Gate-All-Around Junctionless NWFET Biosensor for Detecting Various Diseases *J. Electron. Mater.* **51** 2236–47
- [99] Kumar A, Tripathi M M and Chaujar R 2018 Ultralow-power dielectric-modulated nanogap-embedded sub-20-nm TGRC-MOSFET for biosensing applications *J. Comput. Electron.* **17** 1807–15



# Polymer nanocomposite film based piezoelectric nanogenerator for biomechanical energy harvesting and motion monitoring

Shilpa Rana<sup>1</sup>, and Bharti Singh<sup>1,\*</sup>

<sup>1</sup> Department of Applied Physics, Delhi Technological University, Main Bawana Road, Delhi 110042, India

**Received:** 3 April 2023

**Accepted:** 24 August 2023

**Published online:**  
4 September 2023

© The Author(s), under  
exclusive licence to Springer  
Science+Business Media, LLC,  
part of Springer Nature, 2023

## ABSTRACT

With the advancement in the wearable technologies such as, smart watches, electronic skin, and wearable portable device, scavenging the biomechanical energy from human movements have gained considerable attention for designing self-sustainable power system. Here, we have reported a flexible piezoelectric device that can be conformably adhered to the human body in order to harness the energy from diversification of touch and motion. For this, we have fabricated a polyvinyl difluoride (PVDF) polymer based flexible piezoelectric nanogenerator (PNG) device that can harness energy from various human motions and convert it to useful electrical energy. To further improve the performance of PVDF based nanogenerator, hydrothermally synthesized nanosheets of reduce graphene oxide (rGO) and boron doped rGO are embedded in PVDF matrix as a conductive nanofiller materials to enhance the device output performance. Among all fabricated devices based on pristine PVDF (P), rGO doped PVDF (PR) and, boron doped rGO (PBR), the latter generates a maximum voltage and power density of 13.8 V and  $\sim 42.3 \mu\text{W}/\text{cm}^2$  respectively, which is then used to light up series of commercially available LEDs. Finally, PBR film based PNG is demonstrated to harvest energy from different types of human motion which includes finger tapping, elbow bending, foot tapping, leg folding, and wrist movements. This device demonstrates the potential use of polymer nanocomposite films in self-powered wearable devices.

## 1 Introduction

With the technological advancement, flexible, wearable and, miniaturized microelectronic devices with diverse functionality are rapidly becoming the part of our mundane life by affecting various aspects of human life from health monitoring, and entertainment to information and communication technology [1, 2]. One of the

important factors for continuous and stable operation of these flexible and wearable electronic devices is the sustainable power supply. Till now, batteries have been used as one of the most reliable and economical sources to supply power to these electronic devices but their frequent recharging and replacement process, bulky nature and disposal difficulty are becoming challenge for their long term application and environmental safety [3, 4].

Address correspondence to E-mail: bhartisingh@dtu.ac.in

This problem can be overcome by scavenging the energy from our living environment as there are ample green sources of energy present in our environment including motion or movements, vibrations, waves, solar energy, wind, and thermal energy sources which are capable of generating electricity [5–8]. Among which scavenging mechanical energy particularly from the human body, where sufficient energy is generated by human motion can become potential alternative to supply power to the portable, and wearable electronic device. In this regard, piezoelectric nanogenerators (PNG) have gathered significant attention because of their outstanding ability to convert tiny and irregular mechanical energy from running, walking, typing, muscle movements, and respiration etc. into electricity with the help of piezoelectric materials [9–12]. Several PNG have been developed over the past few years using PZT, BaTiO<sub>3</sub>, GaN, LiNbO<sub>3</sub>, PVDF, and its copolymers to increase the energy conversion efficiency since the invention of first PNG by Prof. Zhong Lin Wang group in 2006 [13–17]. However, inorganic piezoelectric materials suffer from poor toughness/ brittleness, low durability, toxicity and heavy weight which restrict their application in flexible piezoelectric energy harvesting device. In this regard, piezoelectric polymer for example PVDF and its copolymer is a good choice for fabricating flexible nanogenerator due to its flexible, lightweight, biocompatibility, and conformable nature makes them favorable to design flexible device. The electroactive polar (i.e.  $\beta$  and  $\gamma$ ) phases are responsible for facilitating the piezoelectric properties in PVDF film. The piezoelectric response of the PVDF can be further enhanced by several mechanisms, such as, electric poling, mechanical stretching, and the addition of nanofillers i.e. carbon based materials, metal nanoparticles (NPS), metal oxide, and ceramics in polymer matrix which helps in nucleation of electroactive polar phase [16, 18]. Anand et al. synthesized the PVDF composite films via the solution casting method in which addition of rGO in the PVDF matrix and applying the UV Visible light exposure on the sample enhances the piezoresponse of the polymer composite films upto 8 times in comparison to pristine PVDF film and hence producing a maximum voltage of  $\sim 1.9$  V [19]. Hu et al. fabricated a wearable piezoelectric nanogenerator using rGO/PVDF-TrFE film and uses in situ polarization method to increase the energy harvesting performance of the device [20]. The fabricated nanogenerator can produce a maximum voltage, current, and power density of 8.3 V, 0.6  $\mu$ A, and 28.7 W/m<sup>3</sup> respectively. Also, it has been observed from the literature that doping of

rGO with other atoms, such as Fe reported by Karan et al. increases the output performance of the PENG by enhancing the electrical properties [21]. In the reported work, the addition of Fe doped rGO not only helps in the nucleation of electroactive  $\gamma$  phase to  $\sim 99\%$  ( $\pm 0.18$  of relative proportion) in PVDF but also increases the surface conductivity up to  $\sim 3.30 \times 10^{-3}$  Scm<sup>-1</sup> as a result enhance piezoelectric properties of nanocomposite films. Also, the doping of nitrogen, iron, and co-doping of two different heteroatoms in rGO are also reported to enhance the piezoelectric performance of PVDF based nanocomposite film [22, 23]. However, studies on the enhancement of piezoelectric properties in boron doped rGO (B-rGO) have been found to be relatively limited. Furthermore, boron doping in rGO framework leads to redistribution of the energy density, which alters the band structure, charge transfer characteristics, and electronic properties, facilitating its applications in numerous fields, such as, supercapacitors [24], batteries [25], solar cell [26], biomedical applications [27, 28] and can also affects the piezoelectric response of rGO.

In the current study, we have reported a flexible PVDF composite film based piezoelectric nanogenerator for harnessing the mechanical energy from the human body. Firstly, a PVDF based nanogenerator is fabricated, and then to enhance its output performance rGO and boron doped rGO are added in polymer matrix. Then output performance of all the fabricated PNG are compared which shows that incorporation of conducting nanofillers in PVDF matrix improves the output performance. A maximum open-circuit voltage, short-circuit current and power density of 13.8 V, 5.5  $\mu$ A and  $\sim 42.3$   $\mu$ W/cm<sup>2</sup> is obtained for PVDF/boron doped rGO (PBR) nanocomposite film based piezoelectric nanogenerator. To demonstrate the PNG's practical applicability, energy produced by PNG is stored in 1  $\mu$ F capacitor and utilized to power LEDs connected in series. Finally, the fabricated device is demonstrated to harness energy from daily life activities of human body such as wrist bending, leg folding, foot tapping and elbow bending.

## 2 Synthesis and experimental details

The rGO is synthesized by our previous reported paper using hydrothermal approach [22]. For the synthesis of boron doped rGO, 1 g of boric acid is mixed into an aqueous solution of GO and sonicated for 30 min for homogeneous dispersion. The prepared



solution is then transferred to Teflon lined autoclave and kept in oven at 180 °C for 12 h. After being washed multiple times with DI water and ethanol, the resulting product is kept in an oven overnight at 60 °C for drying to get nanosheets of boron doped rGO (B-rGO). Subsequently, a simple drop casting method is used for synthesis of the PVDF and PVDF nanocomposite films. First, 1 g of PVDF is mixed in 10 ml *N,N*-dimethylformamide (DMF) with the help of magnetic stirrer until PVDF powder is completely dissolved in the solvent to obtain transparent solution. Then 10 mg of B-rGO is added to the above solution and stir for 45 min. After that the resultant solution is drop casted on clean glass substrate with the help of micropipette and placed in oven for 2 h at 90 °C to remove any solvent. Then after natural cool down, the films are dipped in DI water to obtain freestanding films. The rGO doped film is also prepared by the same procedure by adding 10 mg of rGO in the PVDF/DMF solution.

## 2.1 Fabrication of the piezoelectric nanogenerator

To fabricate PNG, first the PVDF and PVDF composite films are cut into the requisite shape and aluminum electrode of thickness 100 nm is deposited on both sides of film of area  $2 \times 2 \text{ cm}^2$  with thermal evaporation technique. After that two copper wires are drawn from the top and bottom aluminum electrode surface for the external connection and finally the entire device is encapsulated with Kapton tape. The detailed schematic diagram showing the fabrication of piezoelectric nanogenerator has already been published by our group [22].

## 3 Characterization techniques

Firstly, synthesized powder samples of rGO and B-rGO are characterized by X-ray diffractometer and Raman spectroscopy technique using Rigaku, Ultima diffractometer and WiTec alpha 300 RA model with a laser source of 532 nm wavelength. X-ray photoelectron spectroscopy (XPS) measurements are recorded by Nexsa base spectrometer with Aluminium  $K_{\alpha}$  source for chemical composition analysis. The surface morphologies of the samples together with the energy-dispersive X-ray spectroscopy (EDS) are observed under a field emission scanning electron microscopy

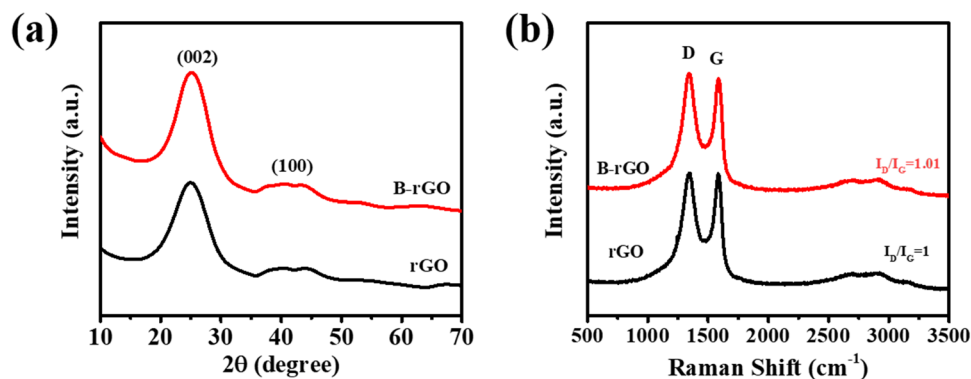
via Nova Nano SEM. Fourier transform infrared spectroscopy (FTIR) spectra of the polymer composite films are obtained by Perkin Elmer Spectrum II spectrometer. The Marine India PE loop tracer is used to measure polarization versus electric field (P–E) loops and YE2730A  $d_{33}$  meter is used to measure the piezoelectric coefficient ( $d_{33}$ ) of PVDF and PVDF composite films. The crystallinity of the films are measured using Differential Scanning calorimeter (DSC 8000, Perkin Elmer). Finally, in order to measure output performance of PNG, the  $V_{OC}$  and  $I_{SC}$  measurements are carried out by Tektronix MDO500 oscilloscope and Keithley DMM7510 digital multimeter.

## 4 Results and discussion

Figure 1a shows the XRD pattern of synthesized rGO and B-rGO samples. Both the samples show the diffraction peak corresponding to (002) plane which is characteristic peak of graphitic materials. For rGO, the peak corresponding to (002) plane is appeared at  $2\theta = 24.84^\circ$ , whereas in B-rGO case the peak is shifted right and appear at  $25.14^\circ$ . The corresponding value of interlayer spacing for rGO and B-rGO is 3.58 Å and 3.53 Å respectively. The decrease in the interlayer spacing can be attributed due to the fact that when we dope boron atom in rGO matrix, it gives rise to strain because of the different atomic size of boron and carbon atom resulting in peak shift [29, 30]. Also, the reduction of  $\pi$  electron density between graphene planes may be the another possible reason for low interlayer spacing in B-rGO due to the replacement of carbon by boron atom [31].

As we know, the crystallinity and homogeneity of the rGO and heteroatom doped rGO is highly sensitive to the synthesis route. Therefore, to evaluate the degree of disorder within the carbon samples, Raman analysis for the rGO and B-rGO are collected at an excitation wavelength of 532 nm. The Raman spectra of all samples exhibit two major Raman peaks at 1344 and  $1586 \text{ cm}^{-1}$  corresponds to the D and G band and a wide 2D-band around  $2800 \text{ cm}^{-1}$  corresponding to the carbon structure in rGO as depicted in Fig. 1b. In the Raman spectra, D band stems from the structural defects and disorder structure in graphitic lattice while G bands ascribed to inplane stretching of C=C bonds and first order  $E_{2g}$  optical mode of graphite. To further deduce the degree of disorder and defects present in rGO sample after doping,  $I_D/I_G$  ratio of

**Fig. 1** **a** XRD and **b** Raman spectra of synthesized rGO and B-rGO samples

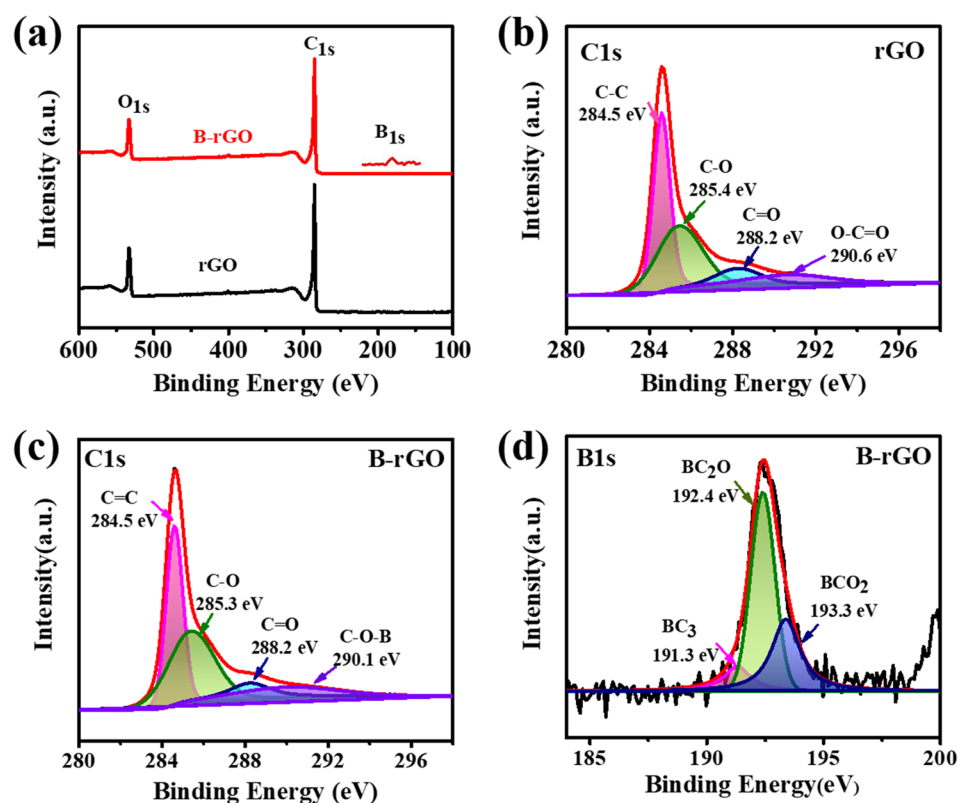


the rGO and B-rGO is calculated which shows that when we dope the boron atom in rGO, the  $I_D/I_G$  ratio increases slightly from 1 to 1.01. This slight increment in the value of  $I_D/I_G$  ratio shows the successful doping of boron atom and increased amount of disorder in  $sp^2$  domains and structural defects in B-rGO [32, 33].

To further support the successful incorporation of boron atom in rGO, the X-ray photoelectron spectroscopy (XPS) analysis is carried out for rGO and B-rGO samples where in the survey spectrum, rGO has only C1s (285.2 eV) and O1s (532.9 eV) peaks, whereas for B-rGO an additional B1s (192.7 eV) peak is observed

showing that boron atoms are bonded to rGO network (Fig. 2a). Figure 2b shows the high resolution deconvoluted C1s spectra of rGO which is divided into four peaks at different binding energy levels (284.5, 285.4, 288.2, and 290.6 eV). The peak at 284.5 eV corresponds to  $sp^2$  hybridized C–C bond, while the peaks at 285.4, 288.2, and 290.6 eV corresponds to C–O, C=O and O–C=O bonds respectively [22, 29, 34]. Figure 2c shows the high resolution C1s spectra of B-rGO, where B-rGO sample have all the functional groups as that in case of rGO with the exception that peak at 290.6 eV is shifted to 290.1 eV and this new peak corresponds to

**Fig. 2** **a** XPS survey spectra of rGO and B-rGO, High resolution deconvoluted C1s spectra of **b** rGO and **c** B-rGO, **d** B1s spectra of B-rGO



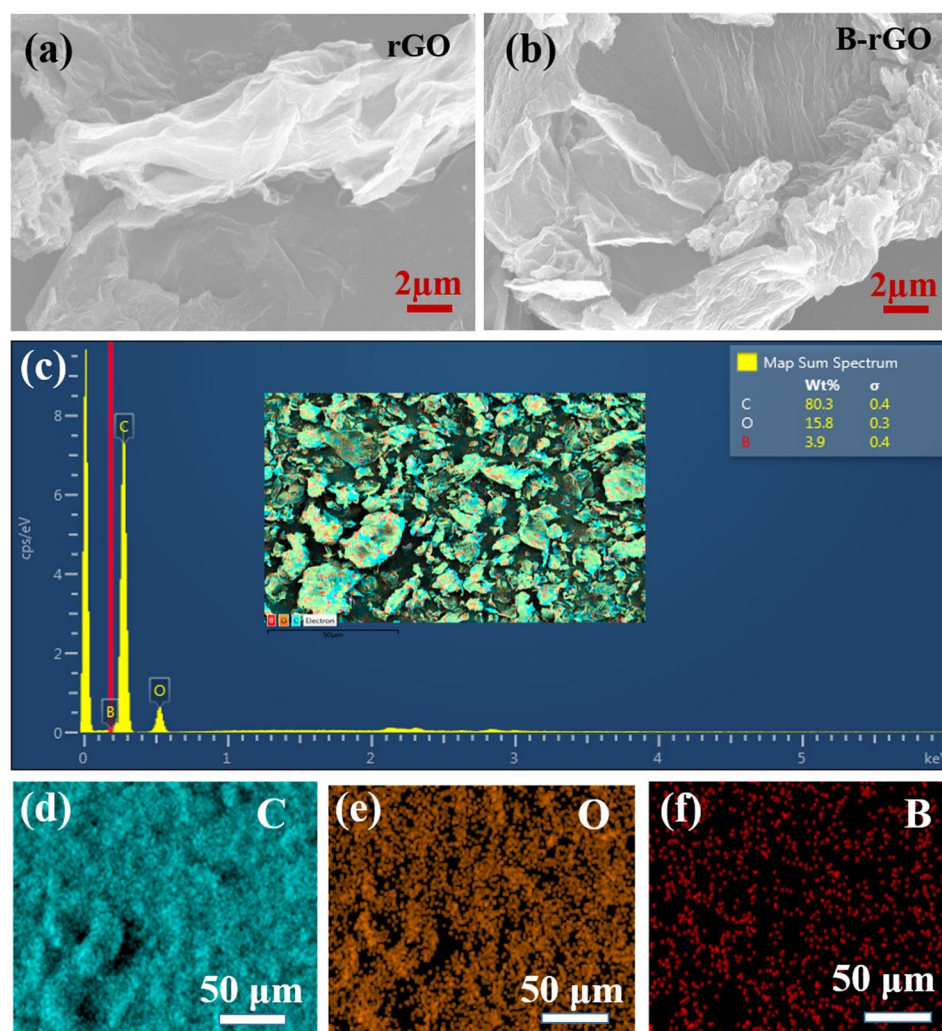
the C–O–B bond. Figure 2d shows the high resolution B1s spectra of B-rGO, which is deconvoluted in three peaks at binding energies 191.3, 192.4 and 193.3 eV corresponding to  $BC_3$ ,  $BC_2O$  and  $BCO_2$  bond respectively [26, 35]. The B1s spectra of B-rGO demonstrates that boron can exist in three different form, first is by substitution of carbon atom by boron atom in the rGO network to form bond  $BC_3$ , second is co-doping with the oxygen to form  $BC_2O$  bond which creates nano-voids in layer and third is  $BCO_2$  bond resulting from the formation of functional groups [36]. These results not only confirm the successful incorporation of boron atoms in rGO network but also show that the inter-layer spacing increases and more defects are introduced in rGO network.

For structural and morphological analysis, FESEM micrographs are obtained for rGO and boron doped rGO samples. As depicted in Fig. 3a–b, rGO exhibit

of randomly aggregated layered structure where few layer of sheets that are closely associated with one another and after doping of boron atom in rGO the aggregation increases. In addition, Energy Dispersive X-ray spectroscopy (EDX) analysis have also been carried out on B-rGO for the elemental analysis of the sample. The EDX spectra confirms the presence of C, O, and B elements for B-rGO sample is illustrated in Fig. 3c–f that further confirm the presence of boron atom in rGO network. After successful synthesis of rGO and B-rGO samples, they are mixed in PVDF and DMF solution to fabricate the polymer composite films. The samples of pure PVDF, rGO doped PVDF and B-rGO doped PVDF are described as P, PR and PBR respectively.

To confirm the existence of different phases in pristine PVDF and PVDF composite films, XRD and FTIR analysis is carried out. Figure 4a depicts the

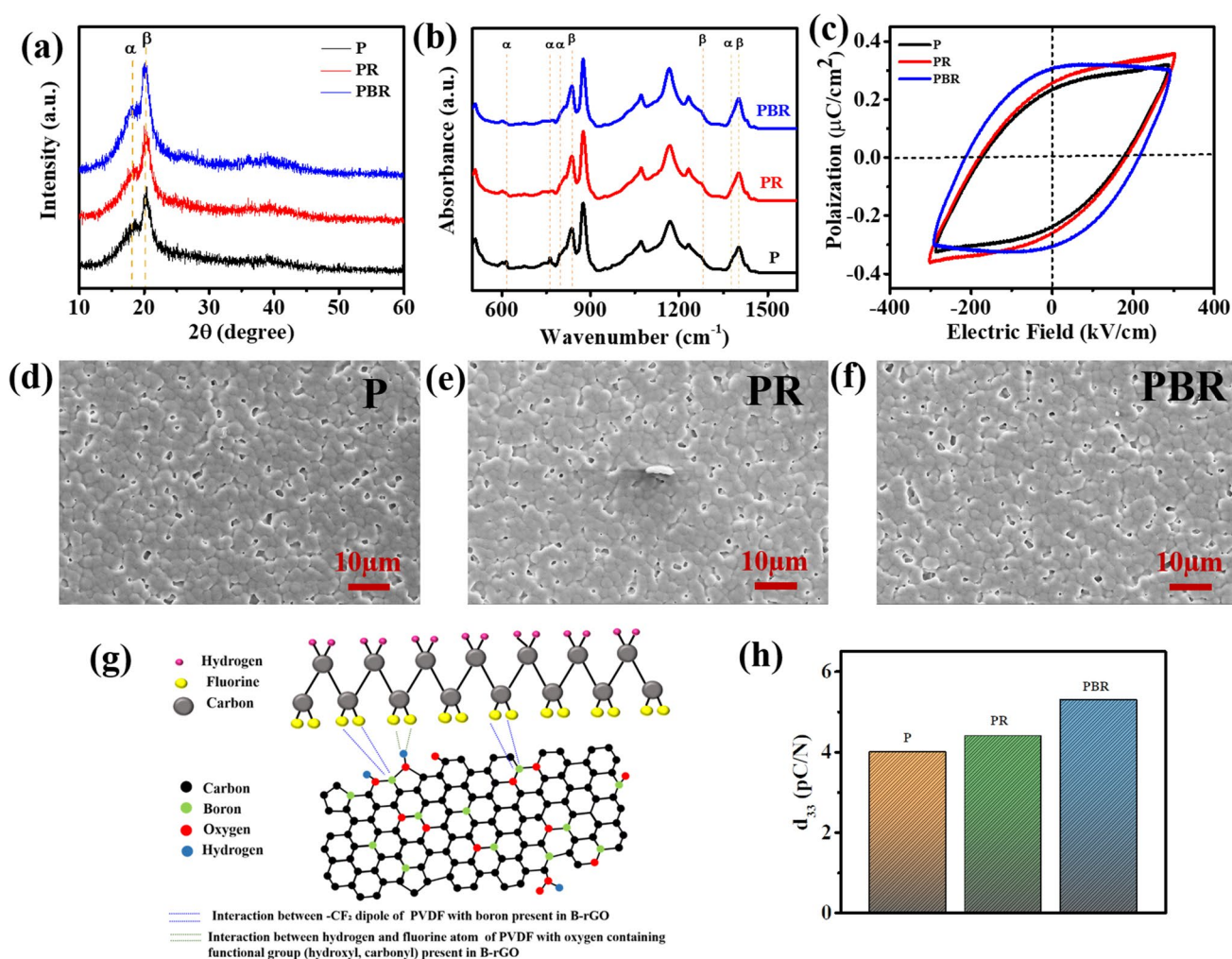
**Fig. 3** Morphological characterization of the synthesized **a** rGO and **b** B-rGO powder sample. **c** EDX spectra and **d** elemental mapping of C, O, and B atom present in the B-rGO sample





XRD pattern of P, PR and PBR composite films where all the samples exhibit two characteristic peak. The presence of a small characteristic peak at  $\sim 18.4^\circ$  correspond to (100) plane belongs to the non-polar  $\alpha$  phase, whereas the presence of sharp intense peak at  $\sim 20.4^\circ$  corresponds to (200) plane of polar  $\beta$  phase of the PVDF [37, 38]. It can be clearly seen from the graph, that the intensity of peak at  $\sim 20.4^\circ$  grows, which proves enhanced polar phase in PVDF with addition of rGO and B-rGO as a nanofillers materials. Moreover, to further investigate the crystal phases of PVDF, FTIR analysis is carried out on the synthesized polymer and polymer composite films. Figure 4b shows the FTIR spectra of pristine PVDF and

PVDF composite films, where peaks at 612, 764, 795, and  $1382\text{ cm}^{-1}$  are attributed to non-polar  $\alpha$  phase with trans-gauche-trans-gauche (TGTG) conformation, whereas the peaks at  $840\text{ cm}^{-1}$  and  $1275\text{ cm}^{-1}$  are related to electroactive polar  $\beta$  phase of PVDF and arises due to the asymmetrical stretching of  $-\text{CF}_2$  and oscillating vibration of  $-\text{CH}_2$  dipoles [39, 40]. In comparison to pristine PVDF film, the intensity of absorbance peaks corresponding to  $\alpha$  phase is decreasing in PR and PBR films indicating that when rGO and B-rGO are doped in PVDF matrix, they aid in the formation of crystalline phase. The relative content of the polar phases in PVDF is calculated with the help of Lambert-Beer law using absorption



**Fig. 4** a–c XRD, FTIR and P–E loops of the pure PVDF (P), PVDF/rGO (PR) and PVDF/B-rGO films, respectively, d–f FESEM images showing the morphology of P, PR and PBR films, g A schematic representation showing the interaction

between dipoles present in PVDF and the filled B-rGO in the PVDF matrix, h piezoelectric strain coefficient  $d_{33}$  of P, PR and PBR films

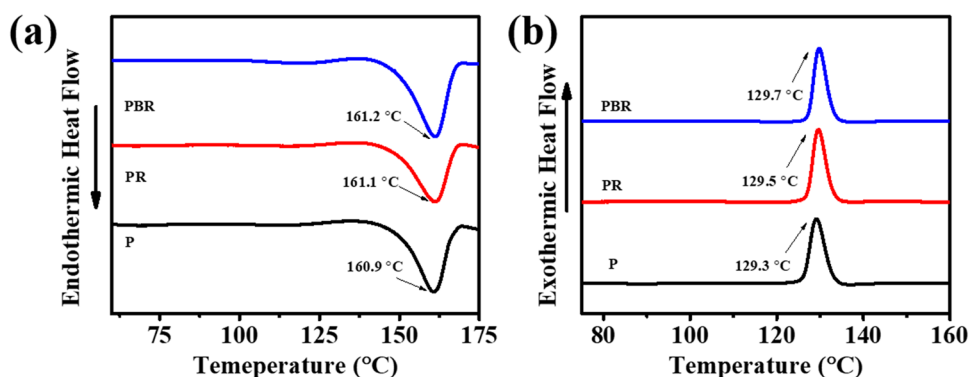
bands at  $761\text{ cm}^{-1}$  and  $840\text{ cm}^{-1}$  and is given by the equation.

$$F(\beta) = \frac{A_\beta}{\left(\frac{k_\beta}{k_\alpha}\right)A_\alpha + A_\beta} \times 100\%,$$

where,  $A_\alpha$  and  $A_\beta$  are the absorbance of  $\alpha$  and  $\beta$  phase at  $761\text{ cm}^{-1}$  and  $840\text{ cm}^{-1}$  and  $k_\alpha$  and  $k_\beta$  are the absorption coefficient at respective wavenumber ( $k_\alpha = 6.1 \times 10^4\text{ cm}^2\text{ mol}^{-1}$ ,  $k_\beta = 7.7 \times 10^4\text{ cm}^2\text{ mol}^{-1}$ ) and  $F(\beta)$  represents the relative fraction of  $\beta$  phase [41]. The value of  $F(\beta)$  obtained for pristine PVDF is  $\sim 69\%$  and after loading the rGO and B-rGO the value of  $F(\beta)$  becomes  $\sim 77\%$  and  $\sim 80\%$  respectively. The substantially increased value of  $F(\beta)$  can be understood by the fact that when we add rGO and B-rGO in the PVDF matrix, there is strong interaction between the surface charges arise due to delocalized  $\pi$  electron, boron atom and oxygen containing functional group in rGO and B-rGO with the  $-\text{CH}_2$  and  $-\text{CF}_2$  dipoles of the polymer chain [21, 42]. The fluorine group of PVDF interacts with boron and oxygen containing functional groups present on basal plane of B-rGO, allowing PVDF chain to crystallize on their surface in trans-trans-trans-trans(TTTT) conformation, facilitating the transformation of non-polar  $\alpha$  phase to the polar  $\beta$  phase. The proposed mechanism for the formation of electroactive polar phase after doping is shown in Fig. 4g. Thus, it can be concluded that the addition of rGO and B-rGO acts as a nucleating agent, which helps in inducing electroactive polar phase in PVDF segment by rotating the polymer chain in all trans configuration. Furthermore, to verify the enhanced piezoelectric properties of the composite films, a comparison between the polarization versus electric field (P–E) hysteresis loop analysis is done on the films to reveal the effect of rGO and B-rGO doping on the ferroelectricity of nanocomposite

films. The P–E loops show that after incorporating rGO and B-rGO in PVDF matrix, the ferroelectric properties of polymer film improve. The value of remnant polarization which indicates the internal dipole moment per unit volume when applied electric field is zero obtained from the P–E loops are 0.23, 0.26 and  $0.31\text{ }\mu\text{C}/\text{cm}^2$  for P, PR and PBR films respectively as shown in Fig. 4c. The improved value of remnant polarization after addition of rGO and B-rGO illustrate that addition nanofillers helps in alignment of dipole in the PVDF film and a maximum value of the remnant polarization corresponds to PBR composite film. In order to further confirm the enhancement in the piezoelectric properties of PVDF after doping of rGO and B-rGO, the piezoelectric strain coefficient  $d_{33}$  is measured with the help of Sino Cera YE2730  $d_{33}$  meter, which relates the electric field produced for an applied electric stress. The value of  $d_{33}$  for P, PR and PBR films are 4, 4.4 and  $5.3\text{ pC}/\text{N}$  respectively as shown in Fig. 4h. The increase in the value of  $d_{33}$  in the composite is attributed due to increase in interfacial polarization after addition of the rGO and B-rGO in the PVDF matrix which is in agreement with P–E and FTIR studies. The effect of addition of rGO and B-rGO on the crystallization and melting temperature of PVDF composite films were analyzed by DSC as the piezoelectric behavior of the nanocomposite films are dependent on the crystalline structure as well as the stability of the electroactive polar phase of the nanocomposite films [43]. Figure 5a–b shows the DSC thermographs of pristine PVDF and PVDF nanocomposite films, which shows that when we add nanofillers in PVDF matrix, the melting temperature ( $T_m$ ), melting enthalpy ( $\Delta H_m$ ), and crystalline temperature ( $T_c$ ) increases gradually as shown in Table 1. The shifting of peak towards high temperature is attributed to the transformation from  $\alpha$  phase to  $\beta$  phase which is

**Fig. 5** DSC **a** heating and **b** cooling curves of PVDF and PVDF nanocomposite films



**Table 1** The DSC parameters of pristine PVDF and PVDF nanocomposite films

Sample	T <sub>m</sub> (°C)	T <sub>c</sub> (°C)	ΔH <sub>m</sub>	χ <sub>c</sub> (%)
P	160.9	129.3	28.73	27.49
PR	161.1	129.5	31.25	30.21
PBR	161.2	129.7	33.27	32.16

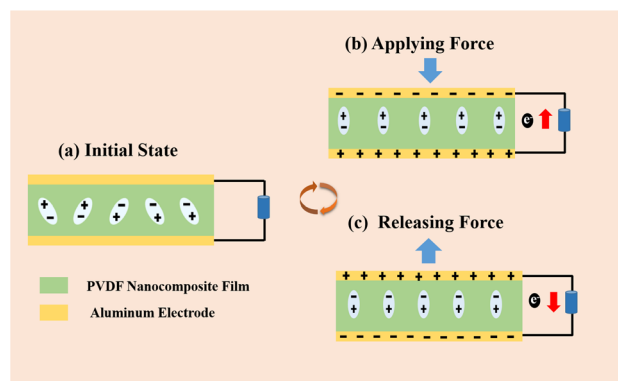
in accordance with XRD and FTIR studies [44]. The percentage crystallinity is determined by the following equation.

$$\chi_c(\%) = \frac{\Delta H_m}{(1 - \phi)\Delta H_m^0} \times 100\%,$$

where, ΔH<sub>m</sub> is melting enthalpy, φ is filler loading percentage, and ΔH<sub>m</sub><sup>0</sup> is the melting enthalpy of 100% crystalline PVDF (i.e. 104.5 J/g). The calculated value of χ<sub>c</sub> for pristine PVDF and PVDF nanocomposite films are shown in Table 1. The percentage crystallinity also increase from ~ 27% in pristine PVDF film to ~ 32% in PBR film. Furthermore, to check the surface morphology of the polymer and polymer composite films, FESEM images are taken for the PVDF and PVDF nanocomposite films (Fig. 4d–f) which shows that rGO and B-rGO nanosheets are evenly distributed in the PVDF polymer matrix without aggregation.

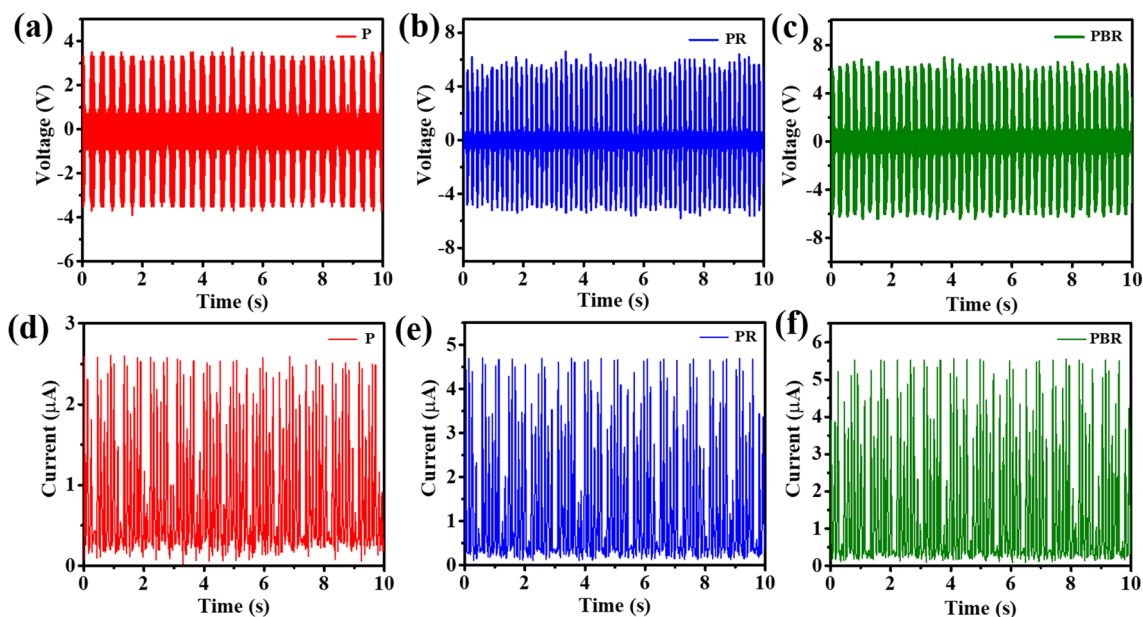
The PNG structure with the corresponding working principle is illustrated in Fig. 6. In initial state, when no force is imparted on device (Fig. 6a), PNG does not deliver any electrical output due to presence of the existing unpolarized electric dipoles in the film. Therefore, when we start to apply force on device (Fig. 6b), film starts deforming to polarize the electric dipoles inside the film and due to polarization effect, electric charges are induced on both electrodes, causing charges to flow from bottom to top electrode. When force is removed (Fig. 6c) the compressive stress fades away in the film leading the electron to flow in the opposite direction until PNG device returns to its initial state. This periodic process results in generation of positive and negative potential cycles in output of nanogenerator.

The piezoelectric energy harvesting performance of PNG device is analyzed by periodic tapping of nanogenerator with the help of the dynamic shaker and measuring the corresponding open-circuit

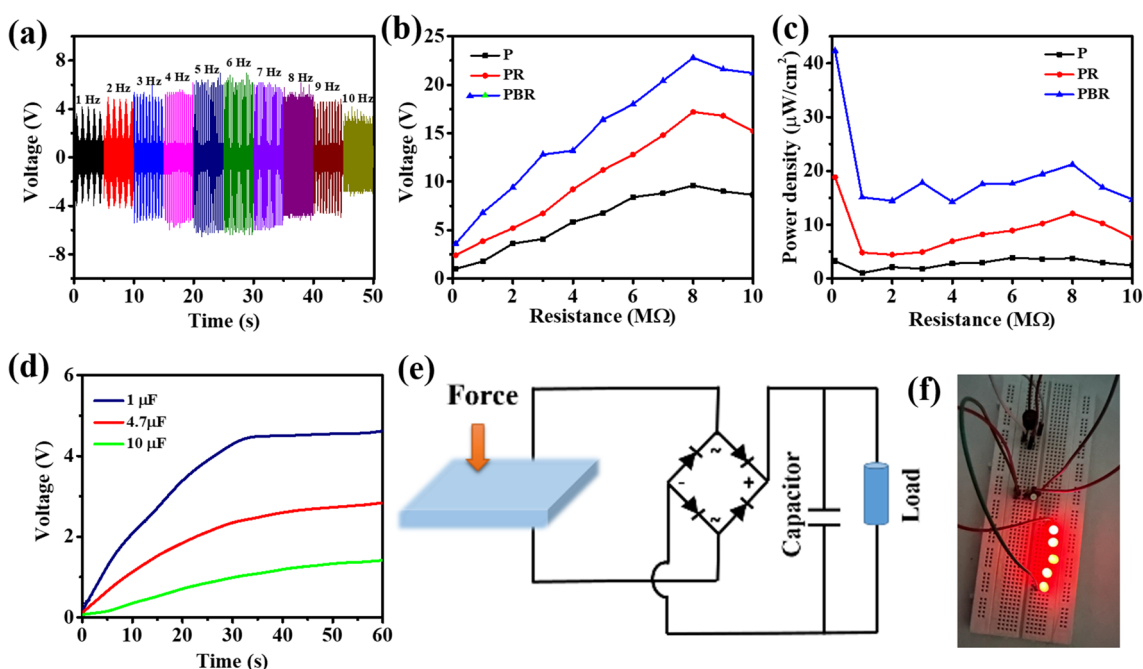
**Fig. 6** A schematic showing the working mechanism of piezo-electric nanogenerator

voltage (V<sub>OC</sub>) and short-circuit (I<sub>SC</sub>) current with the help of the oscilloscope and digital multimeter. In our previous study, we have demonstrated that when we tap the nanogenerator at different frequency, maximum voltage is obtained corresponding to 6 Hz tapping frequency. Therefore, in this paper we have measured the open-circuit voltage and short-circuit current of P, PR, and PBR films at 6 Hz frequency only. Figure 7a–c depicts the measured open-circuit voltage of the fabricated PNG, in which maximum voltage of 13.8 V is produced by PBR film based PNG, whereas pristine PVDF and PR film based PNG produce an output voltage of 7.6 and 12.2 V respectively. The value of rectified short-circuit current obtained for P, PR and PBR film based PNG are 2.6 μA, 4.72 μA and 5.57 μA respectively (Fig. 7d–f). Furthermore, the effect of tapping frequency on the output voltage of PBR film based PNG is also examined which also further verify that maximum output voltage corresponds to the 6 Hz frequency as shown in Fig. 8a. Among all the fabricated device, PNG made by PBR film show maximum output in comparison to PVDF and PR film based PNG. The enhanced performance of PBR film based PNG can be ascribed due to the following aspects. First, adding boron doped rGO in PVDF helps in nucleation of β phase which is supported by XRD, FTIR and P–E studies. Second, it has been reported in the literature that selective surface adsorption of atoms in graphene can induce the piezoelectricity in it by breaking the inversion symmetry, therefore it is possible that when we add boron atom in rGO it will break the inversion symmetry and induce piezoelectricity in it [45]. Third, rGO is conducting in nature





**Fig. 7** a–c Open-circuit voltage and d–f rectified short-circuit current measurements of P, PR, and PBR film based PNG device



**Fig. 8** a The measured output voltage of PBR film based PNG as a function of frequency. b Variation of output voltage and c power density as a function of load resistance, d Charging of 1, 4.7, and 10  $\mu\text{F}$  capacitor by the generated voltage of PBR based

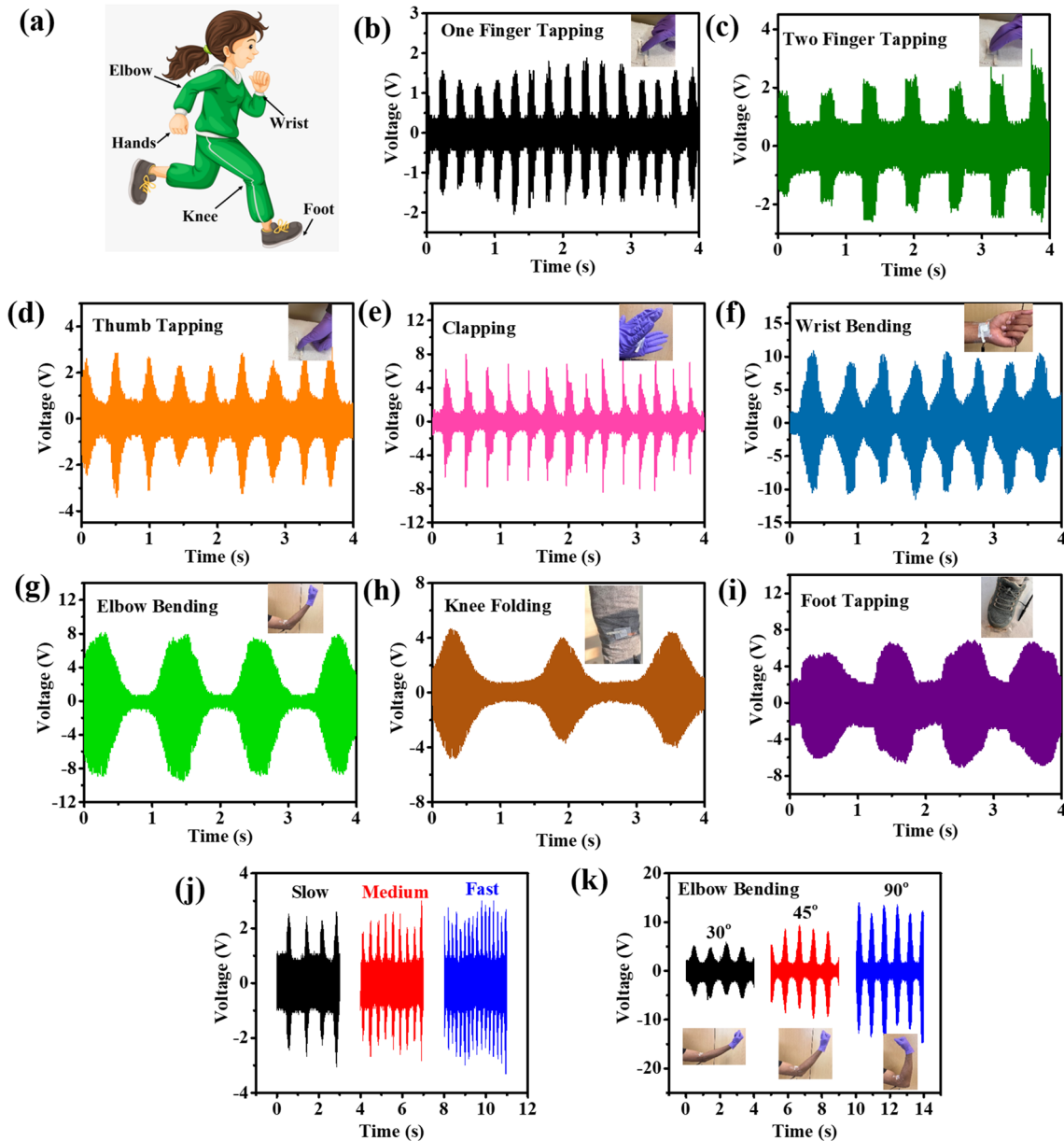
device, e A circuit diagram for storing the energy generated by PNG in the capacitor to power up LEDs f digital photograph showing LEDs powered by PNG

therefore when we add it in the polymer matrix it forms conducting channels which helps in the better charge transfer as a result increase the output performance of the PBR film based PNG device [46].

The variation of the output voltage of P, PR PBR film based PNG devices across the varied load resistance are shown in Fig. 8b, where maximum voltage is corresponding to 8  $\text{M}\Omega$  resistance after that voltage

first saturates and then start decreasing. Figure 8c depicts the power density curves of the corresponding films, with a maximum power density of  $\sim 42.3 \mu\text{W}/\text{cm}^2$  is attained for PBR based PNG device. As the output of the PNG is AC in nature, therefore to store the generated energy a bridge rectifier is utilized to convert AC voltage to DC voltage. Figure 8d shows the energy generated by PBR film based PNG device and is stored in different capacitors (1, 4.7 and  $10 \mu\text{F}$ ). It is clearly seen from the graph that 1

$\mu\text{F}$  capacitor can charge upto  $\sim 4.6 \text{ V}$  in 60 s while 4.7 and  $10 \mu\text{F}$  capacitor can charge upto 2.8 V and 1.4 V respectively. The stored energy is then further used to light up different LEDs and their corresponding circuit design is shown in Fig. 8e. The PBR film based PNG device can light up maximum 5 LEDs with the digital picture illustrated in Fig. 8f and corresponding video is placed in supplementary. Furthermore, the as fabricated nanogenerator is utilized to harness biomechanical energy from human body



**Fig. 9** **a** A schematic representation of different parts of human body from which energy can be harvested. The output voltage generated by PBR film based PNG by **b** one finger, **c** two fingers, **d** thumb tapping, **e** clapping, **f** wrist bending, **g** elbow bend-

ing, **h** knee folding, and **i** foot tapping, **j** variation in the output performance of PNG by slow, moderate and fast finger tapping, **k** The voltage produced by PNG by bending elbow at different angles

by attaching it to different parts of the body and a schematic representation of human body parts from which energy has been harvested is shown in Fig. 9a. Figure 9b–d illustrates the output voltage generated by PNG by tapping it with one finger, two finger and thumb. The variation of the tapping with finger in slow, moderate and high speed is also illustrated in Fig. 9j. The output voltage generated by nanogenerator by clapping and attaching it to wrist, elbow, knee, and tapping it with foot are illustrated in Fig. 9e–i. We also measured PNG's output voltage by bending elbow at various angles (30°, 45° and 90°) as illustrated in Fig. 9k. The above results demonstrate the flexible PBR film based PNG is an effective device that can harness biomechanical energy from the human motion. Thus, by harvesting the energy associated with these various biomechanical motions, it is possible to design a pervasive, sustainable and environmental friendly energy solution for the bioelectronics which will revolutionize the future of wearable electronics in upcoming era of IoT and artificial intelligence [47–49].

## 5 Conclusion

In this work, rGO and B-rGO are successfully synthesized by the hydrothermal technique and are incorporated in the PVDF matrix to synthesize polymer nanocomposite films. It is realized that incorporating rGO and B-rGO into the PVDF matrix improves device piezoresponse by increasing nucleation, which improves content of  $\beta$  phase in PVDF nanocomposite film. A maximum open-circuit voltage and short-circuit current of 13.8 V and 5.5  $\mu$ A is produced by the boron doped rGO/PVDF film based device whereas pristine PVDF based PNG can only generate an output voltage and current of 7.6 V and 2.6  $\mu$ A, respectively by tapping the nanogenerator with 6 Hz frequency. The practical applicability of the as fabricated nanogenerator is demonstrated by storing the harvested energy in the capacitor which is used to power up 5 LEDs connected in series. The fabricated PNG also demonstrated energy harvesting from the human movements, which includes finger tapping, foot tapping, elbow bending, wrist movements and leg folding. Hence, the current study proposes an effective strategy of using biocompatible PVDF flexible films for harvesting the various biomechanical energy for biomedical sector such

as drug delivery, cell stimulation, body patchable device, sensors, and so on, paving a new path for designing a self-powered system by providing a pervasive energy solution for smart and versatile wearable bioelectronics devices.

## Acknowledgements

The authors are grateful to Council of Scientific and Industrial Research (CSIR) with award no (08/133(0042)/2019-EMR-I) for providing the fellowship.

## Author contributions

SR: conceptualization, data curation, methodology, fabrication, writing original draft. BS: supervision, writing review & editing.

## Data availability

The data will be made available on reasonable request.

## Declarations

**Conflict of interest** The authors declare that there is no conflicts of interest.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1007/s10854-023-11207-x>.

## References

1. M. Zhou, M.S.H. Al-Furjan, J. Zou, W. Liu, A review on heat and mechanical energy harvesting from human—Principles, prototypes and perspectives. *Renew. Sustain. Energy Rev.* **82**, 3582–3609 (2018). <https://doi.org/10.1016/j.rser.2017.10.102>
2. F. Mokhtari, Z. Cheng, R. Raad, J. Xi, J. Foroughi, Piezofibers to smart textiles: a review on recent advances and future outlook for wearable technology. *J. Mater. Chem.*

- A **8**, 9496–9522 (2020). <https://doi.org/10.1039/D0TA00227E>
3. S. Li, J. Wang, W. Peng, L. Lin, Y. Zi, S. Wang, G. Zhang, Z.L. Wang, Sustainable energy source for wearable electronics based on multilayer elastomeric triboelectric nanogenerators. *Adv. Energy Mater.* **7**, 1602832 (2017). <https://doi.org/10.1002/aenm.201602832>
  4. X. Chen, Y. Song, Z. Su, H. Chen, X. Cheng, J. Zhang, M. Han, H. Zhang, Flexible fiber-based hybrid nanogenerator for biomechanical energy harvesting and physiological monitoring. *Nano Energy* **38**, 43–50 (2017). <https://doi.org/10.1016/j.nanoen.2017.05.047>
  5. Q. Zhang, Q. Liang, D.K. Nandakumar, H. Qu, Q. Shi, F.I. Alzakia, D.J.J. Tay, L. Yang, X. Zhang, L. Suresh, Shadow enhanced self-charging power system for wave and solar energy harvesting from the ocean. *Nat. Commun.* **12**, 616 (2021). <https://doi.org/10.1038/s41467-021-20919-9>
  6. S. Wu, T. Li, Z. Tong, J. Chao, T. Zhai, J. Xu, T. Yan, M. Wu, Z. Xu, H. Bao, High-performance thermally conductive phase change composites by large-size oriented graphite sheets for scalable thermal energy harvesting. *Adv. Mater.* **31**, 1905099 (2019). <https://doi.org/10.1002/adma.201905099>
  7. J. Wang, S. Zhou, Z. Zhang, D. Yurchenko, High-performance piezoelectric wind energy harvester with Y-shaped attachments. *Energy. Convers. Manag.* **181**, 645–652 (2019). <https://doi.org/10.1016/j.enconman.2018.12.034>
  8. M. Liu, F. Qian, J. Mi, L. Zuo, Biomechanical energy harvesting for wearable and mobile devices: state-of-the-art and future directions. *Appl. Energy* **321**, 119379 (2022). <https://doi.org/10.1016/j.apenergy.2022.119379>
  9. B. Dudem, D.H. Kim, L.K. Bharat, J.S. Yu, Highly-flexible piezoelectric nanogenerators with silver nanowires and barium titanate embedded composite films for mechanical energy harvesting. *Appl. Energy* **230**, 865–874 (2018). <https://doi.org/10.1016/j.apenergy.2018.09.009>
  10. S. Siddiqui, H.B. Lee, D.I. Kim, L.T. Duy, A. Hanif, N.E. Lee, An omnidirectionally stretchable piezoelectric nanogenerator based on hybrid nanofibers and carbon electrodes for multimodal straining and human kinematics energy harvesting. *Adv. Energy Mater.* **8**, 1701520 (2018). <https://doi.org/10.1002/aenm.201701520>
  11. P. Yingyong, P. Thainirarn, S. Jayasvasti, N. Thanach-Issarasak, D. Isarakorn, Evaluation of harvesting energy from pedestrians using piezoelectric floor tile energy harvester. *Sens. Actuators A* **331**, 113035 (2021). <https://doi.org/10.1016/j.sna.2021.113035>
  12. Z. Liu, S. Zhang, Y. Jin, H. Ouyang, Y. Zou, X. Wang, L. Xie, Z. Li, Flexible piezoelectric nanogenerator in wearable self-powered active sensor for respiration and healthcare monitoring. *Semicond. Sci. Technol.* **32**, 064004 (2017). <https://doi.org/10.1088/1361-6641/aa68d1>
  13. S.H. Wankhade, S. Tiwari, A. Gaur, P. Maiti, PVDF—PZT nanohybrid based nanogenerator for energy harvesting applications. *Energy Rep.* **6**, 358–364 (2020). <https://doi.org/10.1016/j.egy.2020.02.003>
  14. Z.L. Wang, J. Song, Piezoelectric nanogenerators based on zinc oxide nanowire arrays. *Science* **312**, 242–246 (2006). <https://doi.org/10.1126/science.1124005>
  15. H. Su, X. Wang, C. Li, Z. Wang, Y. Wu, J. Zhang, Y. Zhang, C. Zhao, J. Wu, H. Zheng, Enhanced energy harvesting ability of polydimethylsiloxane-BaTiO<sub>3</sub>-based flexible piezoelectric nanogenerator for tactile imitation application. *Nano Energy* **83**, 105809 (2021). <https://doi.org/10.1016/j.nanoen.2021.105809>
  16. L. Lu, W. Ding, J. Liu, B. Yang, Flexible PVDF based piezoelectric nanogenerators. *Nano Energy* **78**, 105251 (2020). <https://doi.org/10.1016/j.nanoen.2020.105251>
  17. M.A. Johar, J.-H. Kang, M.A. Hassan, S.-W. Ryu, A scalable, flexible and transparent GaN based heterojunction piezoelectric nanogenerator for bending, air-flow and vibration energy harvesting. *Appl. Energy* **222**, 781–789 (2018). <https://doi.org/10.1016/j.apenergy.2018.04.038>
  18. S. Badatya, D.K. Bharti, N. Sathish, A.K. Srivastava, M.K. Gupta, Humidity sustainable hydrophobic poly (vinylidene fluoride)-carbon nanotubes foam based piezoelectric nanogenerator. *ACS Appl. Mater. Interfaces* **13**, 27245–27254 (2021). <https://doi.org/10.1021/acsami.1c02237>
  19. A. Anand, D. Meena, K.K. Dey, M.C. Bhatnagar, Enhanced piezoelectricity properties of reduced graphene oxide (RGO) loaded polyvinylidene fluoride (PVDF) nanocomposite films for nanogenerator application. *J. Polym. Res.* **27**, 1–11 (2020). <https://doi.org/10.1007/s10965-020-02323-x>
  20. X. Hu, Z. Ding, L. Fei, Y. Xiang, Y. Lin, Wearable piezoelectric nanogenerators based on reduced graphene oxide and in situ polarization-enhanced PVDF-TrFE films. *J. Mater. Sci.* **54**, 6401–6409 (2019). <https://doi.org/10.1007/s10853-019-03339-5>
  21. S.K. Karan, D. Mandal, B.B. Khatua, Self-powered flexible Fe-doped RGO/PVDF nanocomposite: an excellent material for a piezoelectric energy harvester. *Nanoscale* **7**, 10655–10666 (2015). <https://doi.org/10.1039/C5NR02067K>
  22. S. Rana, V. Singh, B. Singh, Tailoring the output performance of PVDF-based piezo-tribo hybridized nanogenerators via B, N-codoped reduced graphene oxide. *ACS Appl. Electron. Mater.* **4**, 5893–5904 (2022). <https://doi.org/10.1021/acsaelm.2c01085>

23. J.-H. Ji, B.S. Kim, J. Kang, J.-H. Koh, Improved output performance of hybrid composite films with nitrogen-doped reduced graphene oxide. *Ceram. Int.* (2021). <https://doi.org/10.1016/j.ceramint.2021.12.271>
24. S. Li, Z. Wang, H. Jiang, L. Zhang, J. Ren, M. Zheng, L. Dong, L. Sun, Plasma-induced highly efficient synthesis of boron doped reduced graphene oxide for supercapacitors. *Chem. Commun.* **52**, 10988–10991 (2016). <https://doi.org/10.1039/C6CC04052G>
25. N. Venkatesan, K.S. Archana, S. Suresh, R. Aswathy, M. Ulaganthan, P. Periasamy, P. Ragupathy, Boron-doped graphene as efficient electrocatalyst for zinc–bromine redox flow batteries. *ChemElectroChem* **6**, 1107–1114 (2019). <https://doi.org/10.1002/celec.201801465>
26. H. Fang, C. Yu, T. Ma, J. Qiu, Boron-doped graphene as a high-efficiency counter electrode for dye-sensitized solar cells. *Chem. Commun.* **50**, 3328–3330 (2014). <https://doi.org/10.1039/C3CC48258H>
27. Z. Fan, Y. Li, X. Li, L. Fan, S. Zhou, D. Fang, S. Yang, Surrounding media sensitive photoluminescence of boron-doped graphene quantum dots for highly fluorescent dyed crystals, chemical sensing and bioimaging. *Carbon* **70**, 149–156 (2014). <https://doi.org/10.1016/j.carbon.2013.12.085>
28. R.S. Sahu, K. Bindumadhavan, R. Doong, Boron-doped reduced graphene oxide-based bimetallic Ni/Fe nanohybrids for the rapid dechlorination of trichloroethylene. *Environ. Sci.* **4**, 565–576 (2017). <https://doi.org/10.1039/C6EN00575F>
29. R.N. Muthu, S.S.V. Tatiparti, Electrode and symmetric supercapacitor device performance of boron-incorporated reduced graphene oxide synthesized by electrochemical exfoliation. *Energy Storage* **2**, e134 (2020). <https://doi.org/10.1002/est2.134>
30. L.K. Putri, B.-J. Ng, W.-J. Ong, H.W. Lee, W.S. Chang, S.-P. Chai, Heteroatom nitrogen-and boron-doping as a facile strategy to improve photocatalytic activity of standalone reduced graphene oxide in hydrogen evolution. *ACS Appl. Mater. Interfaces* **9**, 4558–4569 (2017). <https://doi.org/10.1021/acsami.6b12060>
31. M. Endo, C. Kim, T. Karaki, T. Tamaki, Y. Nishimura, M.J. Matthews, S.D.M. Brown, M.S. Dresselhaus, Structural analysis of the B-doped mesophase pitch-based graphite fibers by Raman spectroscopy. *Phys. Rev. B* **58**, 8991 (1998). <https://doi.org/10.1103/PhysRevB.58.8991>
32. Y. Hishiyama, H. Irumano, Y. Kaburagi, Y. Soneda, Structure, Raman scattering, and transport properties of boron-doped graphite. *Phys. Rev. B* **63**, 245406 (2001)
33. J. Li, X. Li, D. Xiong, L. Wang, D. Li, Enhanced capacitance of boron-doped graphene aerogels for aqueous symmetric supercapacitors. *Appl. Surf. Sci.* **475**, 285–293 (2019). <https://doi.org/10.1016/j.apsusc.2018.12.152>
34. T. Zhu, S. Li, B. Ren, L. Zhang, L. Dong, L. Tan, Plasma-induced synthesis of boron and nitrogen co-doped reduced graphene oxide for super-capacitors. *J. Mater. Sci.* **54**, 9632–9642 (2019)
35. W. Cheng, X. Liu, N. Li, J. Han, S. Li, S. Yu, Boron-doped graphene as a metal-free catalyst for gas-phase oxidation of benzyl alcohol to benzaldehyde. *RSC Adv.* **8**, 11222–11229 (2018). <https://doi.org/10.1039/C8RA00290H>
36. Y. Wang, C. Wang, Y. Wang, H. Liu, Z. Huang, Boric acid assisted reduction of graphene oxide: a promising material for sodium-ion batteries. *ACS Appl. Mater. Interfaces* **8**, 18860–18866 (2016). <https://doi.org/10.1021/acsami.6b04774>
37. V. Singh, D. Meena, H. Sharma, A. Trivedi, B. Singh, Investigating the role of chalcogen atom in the piezoelectric performance of PVDF/TMDCs based flexible nanogenerator. *Energy* **239**, 122125 (2022). <https://doi.org/10.1016/j.energy.2021.122125>
38. B. Jaleh, A. Jabbari, Evaluation of reduced graphene oxide/ZnO effect on properties of PVDF nanocomposite films. *Appl. Surf. Sci.* **320**, 339–347 (2014). <https://doi.org/10.1016/j.apsusc.2014.09.030>
39. X. Cai, T. Lei, D. Sun, L. Lin, A critical analysis of the  $\alpha$ ,  $\beta$  and  $\gamma$  phases in poly (vinylidene fluoride) using FTIR. *RSC Adv.* **7**, 15382–15389 (2017). <https://doi.org/10.1039/C7RA01267E>
40. F. Wang, H. Sun, H. Guo, H. Sui, Q. Wu, X. Liu, D. Huang, High performance piezoelectric nanogenerator with silver nanowires embedded in polymer matrix for mechanical energy harvesting. *Ceram. Int.* **47**, 35096–35104 (2021). <https://doi.org/10.1016/j.ceramint.2021.09.052>
41. P. Martins, A. Lopes, S. Lanceros-Mendez, Electroactive phases of poly (vinylidene fluoride): determination, processing and applications. *Prog. Polym. Sci.* **39**, 683–706 (2014). <https://doi.org/10.1016/j.progpolymsci.2013.07.006>
42. R. Bhunia, S. Gupta, B. Fatma, Prateek, R.K. Gupta, A. Garg, Milli-watt power harvesting from dual triboelectric and piezoelectric effects of multifunctional green and robust reduced graphene oxide/P (VDF-TrFE) composite flexible films. *ACS Appl. Mater. Interfaces* **11**, 38177–38189 (2019). <https://doi.org/10.1021/acsami.9b13360>
43. S. Ojha, S. Paria, S.K. Karan, S.K. Si, A. Maitra, A.K. Das, L. Halder, A. Bera, A. De, B.B. Khatua, Morphological interference of two different cobalt oxides derived from a hydrothermal protocol and a single two-dimensional metal organic framework precursor to stabilize the  $\beta$ -phase of PVDF for flexible piezoelectric nanogenerators. *Nanoscale*



- 11, 22989–22999 (2019). <https://doi.org/10.1039/C9NR08315D>
44. F. Mokhtari, G.M. Spinks, S. Sayyar, J. Foroughi, Dynamic mechanical and creep behaviour of meltspun pvdf nano-composite fibers. *Nanomaterials* **11**, 2153 (2021). <https://doi.org/10.3390/nano11082153>
45. M.T. Ong, E.J. Reed, Engineered piezoelectricity in graphene. *ACS nano* **6**, 1387–1394 (2012). <https://doi.org/10.1021/nn204198g>
46. K. Shi, B. Sun, X. Huang, P. Jiang, Synergistic effect of graphene nanosheet and BaTiO<sub>3</sub> nanoparticles on performance enhancement of electrospun PVDF nanofiber mat for flexible piezoelectric nanogenerators. *Nano Energy* **52**, 153–162 (2018). <https://doi.org/10.1016/j.nanoen.2018.07.053>
47. S. Niu, X. Wang, F. Yi, Y.S. Zhou, Z.L. Wang, A universal self-charging system driven by random biomechanical energy for sustainable operation of mobile electronics. *Nat. Commun.* **6**, 8975 (2015). <https://doi.org/10.1038/ncomms9975>
48. Y. Zou, V. Raveendran, J. Chen, Wearable triboelectric nanogenerators for biomechanical energy harvesting. *Nano Energy* **77**, 105303 (2020). <https://doi.org/10.1016/j.nanoen.2020.105303>
49. T. He, H. Wang, J. Wang, X. Tian, F. Wen, Q. Shi, J.S. Ho, C. Lee, Self-sustainable wearable textile nano-energy nano-system (NENS) for next-generation healthcare applications. *Adv. Sci.* **6**, 1901437 (2019). <https://doi.org/10.1002/advs.201901437>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.





# Potato Peel Waste as an Economic Feedstock for PHA Production by *Bacillus circulans*

Sonika Kag<sup>1</sup> · Pravir Kumar<sup>1</sup> · Rashmi Kataria<sup>1</sup>

Accepted: 15 September 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

## Abstract

Polymers of hydroxy alkanooates (PHA), also known as biodegradable, biocompatible plastic, are potential alternatives to petrochemical-based plastics. PHA is synthesized by microbes in their cytoplasm in the form of inclusion bodies in stress conditions such as nitrogen, oxygen, and phosphorus with excessive amounts of carbon. Sugar extracted from potato peel in the form of hydrolysate was employed as a carbon source for PHA production after acidic hydrolysis. The acid hydrolysis conditions are optimized for dilute acid concentrations and temperatures. The highest sugar-yielding condition (2% 15 min at 121 °C) was used for submerged fermentation for PHA production by *Bacillus circulans* MTCC 8167. Fourier transform infrared spectroscopy, nuclear magnetic resonance, and differential scanning calorimetry were used for polymer characterization. Gas chromatography coupled with mass spectrometry confirmed the monomers such as hexadecenoic acid 3-hydroxy, methyl esters, pentadecanoic acid 14 methyl esters, and tetradecanoic acid 12-methyl esters. Crotonic acid assay was used for quantification of PHA and it was found highest ( $0.232 \pm 0.04$  g/L) at 37 °C and 36 h of incubation. Hence, potato peel waste could be a potential feedstock for waste to valuable production.

**Keywords** Cell dry weight · Fermentation · Bioplastics · Potato peel hydrolysate · Reducing sugars

## Introduction

Plastics play an indispensable role in our day-to-day lives due to their quality and diverse applications. The increasing human population has led to the deposition of commercial plastics in the form of microplastics, which have adversely affected the marine environment. Consequently, many countries are now focusing on sustainable alternative materials with properties similar to those of petroleum-based synthetic plastics [20]. Polyhydroxybutyrate (PHB) is a homopolymer and one of the most extensively studied types of polyhydroxyalkanoates (PHAs). These biopolymers are synthesized as intracellular inclusion

---

✉ Rashmi Kataria  
rashmikataria@gmail.com

<sup>1</sup> Department of Biotechnology, Delhi Technological University (DTU), Shahbad Daulatpur Village, Bawana Road, Delhi 110042, India

bodies by various bacterial species under adverse conditions and exhibit similarities to commercial plastics [50].

Based on their monomeric subunits, PHAs can be classified into two categories: short-chain-length (SCL) PHAs, which contain 3 to 5 carbons, and medium-chain-length (MCL) PHAs, with structures containing 6 to 14 carbons [11]. Structurally, PHAs can be found in the form of homopolymers (P(HB)) or heteropolymers (P(HB-CO-HV)). These biopolymers have various industrial and biomedical applications, including drug delivery carriers, dental implants, and the synthesis of biodegradable bioplastics [39].

However, the application of these biodegradable plastics has been challenging at an industrial level due to the high production costs [39]. This challenge can be mitigated by using cost-effective substrates for microorganisms that accumulate PHAs [37]. Sugar-rich wastes from different sources, such as wheat straw, rice straw, corn straw, potato peels, whey protein, and glycerol, are gaining attention for renewable production [16]. Wheat, rice, maize, and potatoes are among the most consumed crops worldwide, leading to substantial peel waste generation from both industries and households. Potato peel waste, for instance, contains 30–40% glucose in the form of starch, which can contribute to environmental pollution if disposed of in open areas [24].

Starch hydrolysis yields monomeric hexose sugar in the form of glucose, which serves as the preferred carbon source for energy generation within bacterial cells.

Starch hydrolysis can be achieved through acid hydrolysis or by using amylase and glucoamylase enzymes [6]. While enzymatic hydrolysis is a preferable and rapid method for generating reducing sugars, the use of commercial enzymes may not be cost-effective for large-scale processes [27]. An alternative option for glucose generation is the use of dilute acid, which conserves both energy and time. This method can be particularly useful for various industrially important chemicals derived from potato peel waste, such as organic acids [29], biofuels [2, 27], carbohydrate polymers [1], and bioplastics [8]. Several studies are underway to enhance product yield using different approaches, including advanced pretreatment strategies and simultaneous saccharification and fermentation [41]. However, microbial production of PHA from potato peel waste remains relatively unexplored.

In the present work, potato peel waste was subjected to hydrolysis using dilute acid, and the extracted sugar was utilized for bacterial polymer production. The production medium for the bacterial strain *Bacillus circulans* MTCC 8167 was prepared by adding sugar hydrolysate, 0.5% NaCl, and 0.1% tryptone. PHA was detected and characterized through a crotonic acid assay, GC–MS, FTIR, NMR, and DSC.

## Material and Methods

### Potato Peel Collection and Primarily Processing

A potato peel sample was collected from the Delhi Technological University (DTU) canteen and hostels. Processing of the sample was done by washing followed by drying at 40 °C until constant weight. The dried sample was grinded to reduce particle size and sieved through a 1-mm mesh.

## Bacterial Culture

An isolated bacterial culture, *Bacillus circulans* MTCC 8167, was procured from the Microbial Type Culture Collection (MTCC) in Chandigarh, India. Culture was maintained on Luria Bertani (LB) media containing, yeast extract (5 g/L), peptone (10 g/L), NaCl (10 g/L), and agar (15 g/L), the pH was adjusted to 7. The culture media after inoculation were incubated at 37 °C with 180 rpm for the growth studies.

## Composition Analysis and Acid Hydrolysis of Potato Peel Waste

One gram of dried powder of potato peel was subjected to composition analysis in a triplicate. Starch analysis was done using the spectrophotometric method [28]; protein analysis was done by multiplying plant factor 6.25 with obtained nitrogen content after three-step procedure digestion, distillation, and titration as given in literature [45]; and lignin analysis was done using the protocol established by National Renewable Energy Laboratory [44]. Structural carbohydrate such as cellulose and hemicellulose analysis was done gravimetrically using chlorite method [51]. In which holocellulose was determined with NaClO<sub>2</sub> treatment, cellulose extraction from holocellulose was done using 17.5% NaOH treatment and hemicellulose was estimated by subtracting cellulose from holocellulose.

Potato peel biomass is composed of starch, cellulose, lignin, protein, etc. [7, 13]. To release fermentable sugar from polysaccharides, a proper treatment strategy is required before subjecting the biomass to microbial fermentation [41]. Dilute acid treatment at high temperatures is done to convert the potato peel powder into reducing sugar. For chemical hydrolysis, 1 L of 3.2 M HCl stock solution was prepared by mixing 275.86 mL of 10% HCl in 724.13 mL of distilled water. Different hydrochloric acid (HCl) concentrations (0.5, 0.75, 1, 1.25, 1.5, and 2%) were prepared from stock solution and pretreatment was done at temperatures (50, 100, and 121 °C). Ten percent solid loading (10 g potato peel in 100 mL of aqueous phase) was done in each combination. Hydrolysate obtained after hydrolysis was cooled at room temperature and subjected to detoxification. In the process of detoxification, dry calcium carbonate was added to acidic hydrolysate, and the pH of the sample was monitored throughout the time to reach 6.8–7 [3]. After neutralization, centrifugation was done to get clear hydrolysate and followed by filtration to obtain the sugar-rich extract. Sugar analysis was performed by the DNS method [35].

## Analytical Methods

### Sugar Analysis, PHA Production, Quantification, and Characterization

Sugar analysis of liquid hydrolysate was done by di-nitro salicylic acid given by Miller [35] briefly, 1 mL of DNS reagent was added to 1 mL pretreated liquid hydrolysate and heated at 100 °C for 10 min, and reducing sugar concentration was determined by UV visible spectrophotometer (GENESYS 50) at 540 nm.

After optimization of acid hydrolysis conditions, the reducing sugar (10 g/L), from the highest yielding condition (2% HCl at autoclave condition), was applied as a carbon source for the cultivation of *Bacillus circulans*. Hydrolysate was filtered through a 0.22-μm pore-sized syringe filter diameter of 25 mm made up of Avantor material limited. Sodium chloride (5 g/L) and organic nitrogen source tryptone (1 g/L) were added after autoclaving in sterilized hydrolysate. Bacterial culture was prepared by inoculation of a single colony of bacteria in a culture

medium in a 250-mL flask at 37 °C in a shaking incubator at 180 rpm. The culture was taken from the mid-exponential phase for 1% inoculum for the hydrolysate fermentation. The growth pattern of *Bacillus circulans* was observed in terms of cell dry weight, PHA production, and sugar consumption at different time intervals such as 24, 36, and 48 h [15]. Residual sugar was analyzed at 540 nm using a thermo-scientific UV visible spectrophotometer (GENESYS 50).

After fermentation, the culture suspensions were subjected to centrifugation (microprocessor-based laboratory centrifuge) at 6000 rpm for 15 min at room temperature. Cell pellets were washed with deionized water and then dried in the oven (Matrix Scientific) at 55 °C until the constant weight of cell dry weight in each condition was recorded. The experiment was done in a triplicate manner. The extraction of PHA was done with 1:20 ratios of sample and chloroform by heating at 60 °C for 120 min in a water bath. After cooling the extractives, precipitation was done with a ten-fold volume of ice-chilled methanol as described by Shah [43].

The polymer obtained from the above method was estimated by the crotonic acid method. PHA is heated with  $H_2SO_4$ , it is converted into crotonic acid, and its concentration is determined by spectrophotometric assay. The stock solution of crotonic acid was prepared by dissolving 0.1 mg of crotonic acid into 1 mL of sulfuric acid. Different concentrations of 2, 5, 7, 8, 9, and 10  $\mu\text{g/mL}$  for crotonic acid were prepared from the stock solution with a working volume of 1 mL. Sulfuric acid was used as blank, and absorbance was recorded at 235 nm. The standard graph was plotted for concentration vs. absorbance. From the standard curve concentration of the test sample (PHA) was determined [23].

Biopolymer (PHA) was analyzed by acidic methanolysis. The 20 mg dried biomass was subjected to methanolysis by the addition of 2 mL  $CHCl_3$ , 1.7 mL  $CH_3OH$ , and 0.3 mL  $H_2SO_4$ . The mixture was heated at 100 °C for 2 h and 20 min of duration. A 2- $\mu\text{L}$  sample from the bottom organic layer was used for GCMS analysis (Shimadzu QP 2010) and the injection temperature of the column was maintained at 260 °C. Monomers were analyzed by the NIST library as previously done by Mahato et al. [31].

The functional group of PHA from *Bacillus circulans* was analyzed by FT-IR with a spectral range of 4000–400  $\text{cm}^{-1}$  (Perkin Elmer (Frontier Shelton CT08484)) as described in the literature [32]. The thermal degradation behavior of extracted PHA powder was checked using a differential scanning calorimeter (DSC) with an instrument make of Perkin Elmer model 8000. Nitrogen gas was used for purging with a flow rate of 10 mL/min. 10 mg PHA was subjected to a thermal profile from –10 to 300 °C and the heating rate was 10 °C/min [19]. To check chemical properties  $^1\text{H}$  NMR was performed. Briefly, 6 mg of pure PHA was dissolved in 1 mL  $CDCl_3$  (denatured chloroform) and spectra were recorded on 500 MHz Bruker spectrophotometer [32].

## Result and Discussion

### Composition Analysis

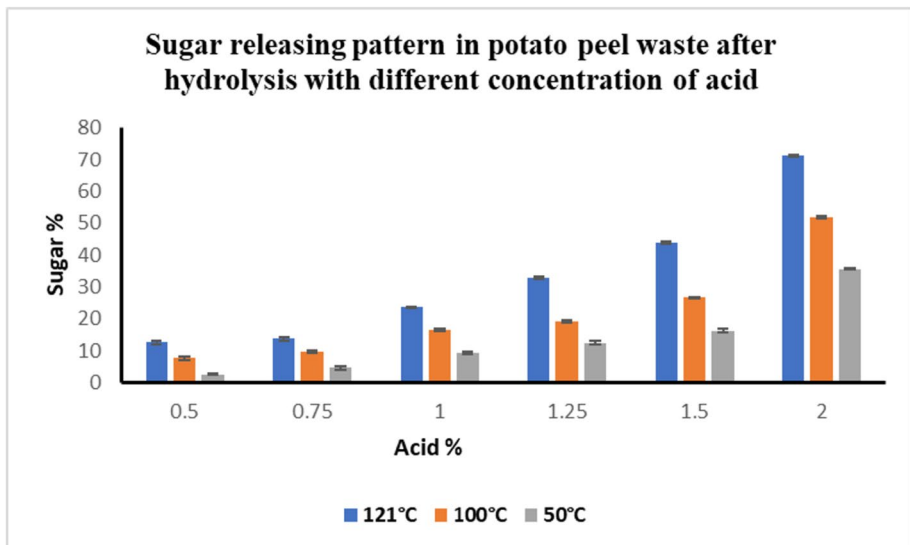
The results of the composition analysis of potato peel waste are shown in Table 1. It was observed that a high amount of starch (63.34%), followed by protein (14%), lignin (7.12%), and nitrogen (2.27%), is present in potato peel waste. In a study by Remedios and Domingues [41], 46% starch, 10% protein, ash 7%, and lignin 4% were found in potato peel samples. In a literature Lima et al. [30] reported 32.4% cellulose and 10% hemicellulose content from potato peel. However, the composition of potato peel varies in different varieties of potatoes and peeling methods [42].

**Table 1** Composition of potato peel waste

Parameters	% (w/w)
Protein	14 ± 0.40
lignin	7.12 ± 0.22
Starch	63.34 ± 0.20
Cellulose	9.42 ± 0.22
Hemicellulose	2.18 ± 0.22
Others	3.94 ± 0.35

## Acid Hydrolysis

As a result of acid hydrolysis, the complexity of peel biomass decreases, and the amount of fermentable sugar increases. Different concentrations of dilute acid (HCl) and time variation for hydrolysate development were applied in this study which resulted in a varied amount of reducing sugar. Based on the sugar yields comparative results are shown in Fig. 1 and Tables 2 and 3. In the acid hydrolysis process, a water molecule is added through a nucleophilic reaction due to the breakdown of chemical bonds of starch, and a change in porosity occurs which leads to the release of sugar upon hydrolysis [10]. Experimental conditions used in this study such as 50 °C and 100 °C at different acid ranges were found insufficient for maximum reducing sugar extraction because sugar yield is directly proportional to hydrolysis temperature at a certain limit [46]. In an autoclave (at 121 °C) with 2% acid concentration, 71.23 ± 0.46% of sugar was extracted, which is the highest in comparison to other hydrolysis conditions (0.5, 0.75, 1, 1.25, and 1.5% HCl concentration). The reason behind this observation is that soft biomass liberates high sugar at mild acidic conditions and high temperatures which are previously observed in the literature [47]. Similarly,

**Fig. 1** Sugar releasing pattern in potato peel waste after hydrolysis with different concentrations of acid

**Table 2** Reducing sugar yield (%) after acid hydrolysis

Acid %	Sugar % at 121 °C	Sugar % at 100 °C	Sugar % at 50 °C
0.5	12.47 ± 0.40	7.61 ± 0.52	2.58 ± 0.15
0.75	13.71 ± 0.54	9.69 ± 0.30	4.55 ± 0.46
1	23.58 ± 0.17	16.44 ± 0.29	9.28 ± 0.51
1.25	32.77 ± 0.36	19.06 ± 0.38	12.41 ± 0.53
1.5	43.66 ± 0.34	26.59 ± 0.14	16.22 ± 0.46
2	71.23 ± 0.46	51.84 ± 0.29	35.50 ± 0.20

in a report, 23 g/L of sugars using 1% sulfuric acid pretreatment in autoclave condition was obtained [33]. In another study [18], 1–5% HCl was used to hydrolyze potato peel and they reported 62 g/L glucose from 5% acid concentration at 90 °C for 120 min reaction time. Martín et al. [34] reported almost similar amount (72%) conversion of glucose from cassava stems where in the first step, cassava stems are subjected to acid pretreatment with 0.6% H<sub>2</sub>SO<sub>4</sub> and then enzymatic hydrolysis was with a combination of amylase and glucoamylase. Though high temperatures of more than 121 °C and increased acid concentration lead to the accumulation of toxic material such as phenolics due to the degradation of sugars [46]. Though the goal of this study was to extract maximum sugar for PHA production and therefore 2% acid was used which leads to liberating 71.23 ± 0.46% reducing sugar.

### PHA Production by *Bacillus circulans* MTCC 867 in Potato Peel Media

The PHA production by *Bacillus circulans* in potato peel hydrolysate-containing media was observed in three different time intervals (24, 36, and 48 h). The physical growth parameters play a vital role in cell dry weight accumulation and PHA production. At 24 h of incubation, biomass was minimum (0.926 ± 0.01), as actively dividing cells utilize sugar only for growth purposes. Cell biomass was found at maximum (1.36 ± 0.02) up to 36 h as secondary metabolites (PHA) produced during the late stationary phase of the growth cycle when nutrients are in the limit. The stationary phase is signal when PHA synthesis enzymes such as 3-keto thiolase, acetoacetyl-CoA reductase, and PHA synthase start converting PHA from acetyl-CoA which is commonly known as inclusion bodies [12]. A depletion of cell biomass (1.09 ± 0.06) until 48 h was observed as storage granules are utilized as a source of energy by the bacteria [32].

Table 4 and Fig. 2 show representation of cell dry weight and PHA accumulation pattern in different time intervals. All three-production media (24, 36, and 48 h) are further subjected to extraction by solvent extraction method. Different studies have shown the PHA accumulation

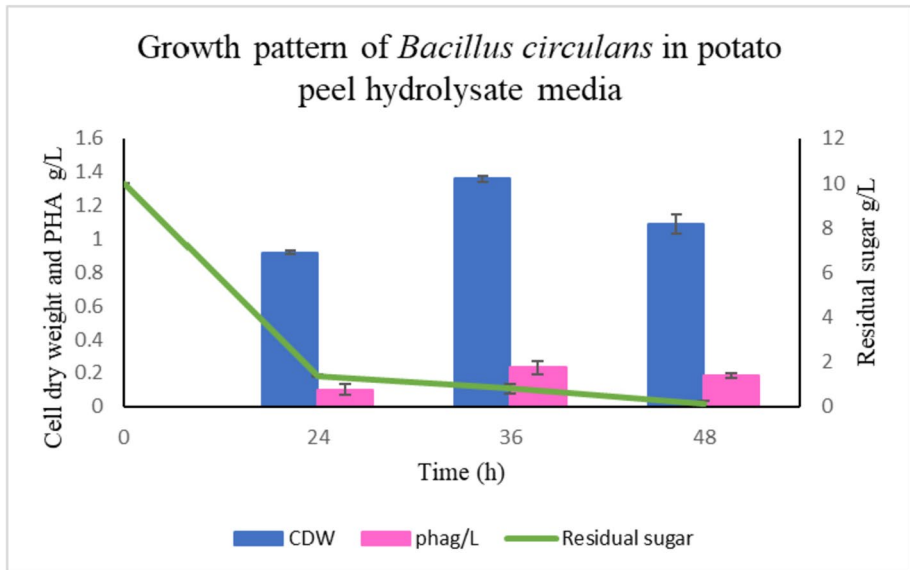
**Table 3** Reducing sugar yield (mg/gram of potato peel waste) after acid hydrolysis

Acid %	Sugar mg/g of potato peel at 121 °C	Sugar mg/g of potato peel at 100 °C	Sugar mg/g of potato peel at 50 °C
0.5	125.55 ± 0.31	76.11 ± 0.22	25.84 ± 0.16
0.75	137.15 ± 0.18	96.94 ± 0.13	45.57 ± 0.40
1	235.84 ± 0.2	164.48 ± 0.29	92.89 ± 0.15
1.25	327.77 ± 0.23	190.64 ± 0.17	124.11 ± 0.21
1.5	436.61 ± 0.11	265.97 ± 0.30	162.28 ± 0.35
2	712.33 ± 0.30	518.46 ± 0.22	355.01 ± 0.20



**Table 4** Cell dry weight (g/L) at different time intervals in potato peel hydrolysate media

Time (h)	Cell dry weight (g/L)	PHA g/L	Residual sugar (g/L)
24	$0.926 \pm 0.01$	$0.1 \pm 0.031$	$1.38 \pm 0.05$
36	$1.36 \pm 0.02$	$0.232 \pm 0.041$	$0.8 \pm 0.21$
48	$1.09 \pm 0.06$	$0.184 \pm 0.015$	$0.11 \pm 0.14$

**Fig. 2** Growth pattern of *Bacillus circulans* in potato peel hydrolysate media

ability of a variety of *Bacillus* species by utilizing different organic wastes as carbon sources [5, 23, 49]. Subsequently, these carbon sources were applied in discrete combinations of other nutrient components, which showed the high PHA accumulation in the bacterial cytoplasm. For instance, in an optimization study [40], high yield of PHA (46.57%) was observed in *Bacillus endophyticus*; however, bacteria were cultivated in media containing salts such as  $\text{Na}_2\text{HPO}_4$ ,  $\text{KH}_2\text{PO}_4$ , and sucrose 40 g/L was used as a carbon source. As several studies observed, PHA enhancement by optimizing the desirable C:N ratio [4, 48, 52] or by use of genetically engineered microbes [21]. But in the present study, wild-type *Bacillus circulans* was taken and carbon was 10 g/L used from waste potato peels, and other growth-promoting micronutrients were not added in production media. A comparison of PHA production by different strains of *Bacillus* based on different parameters including polymer type, yield, and utilized carbon source is shown in Table 5. From the comparative studies, explained in above-mentioned table, it is observed that potato peel waste can be a good carbon source for an adequate amount of PHA production by *Bacillus circulans*.

**Table 5** PHA production by different species of *Bacillus*

Bacteria	Type of polymer	Carbon source	Yield (%)	Reference
<i>Bacillus megaterium</i>	PHB	DSMZ media containing glycerol as a carbon source	14.11	Hiremath and Sura [23]
<i>Bacillus cereus</i>	PHB	Grape residue	18.79	Andler et al. [5]
<i>Bacillus sp.</i>	PHB	Mineral salt media with 2% glucose	20	Hassan et al. [22]
<i>Bacillus megaterium</i>	PHA	Food waste-derived volatile fatty acid	10	Vu et al. [49]
<i>Bacillus circulans</i>	PHA	Potato peel waste	23	Present study

## GC–MS Analysis

By GC–MS analysis medium chain length, PHA monomers such as hexa decanoic acid 3-hydroxy, methyl ester, pentadecanoic acid 14- methyl -esters, and tetra decanoic acid, 12- methyl esters were identified with area % 4.25, 1.94, and 7.43%, respectively, using potato peel waste as a carbon source. However, scl PHA was reported by *B. cereus* SS105 [32]. Moreover Choonut et al. [14] reported five different types of mcl PHA monomers reported by *Bacillus theroaylovorans*-related strains. The type of monomer accumulation inside the bacterial cell depends on the type of carbon source [9].

## Quantification of PHA

Cell dry weight (1 mg) was transformed into crotonic acid by adding 10 mL  $\text{H}_2\text{SO}_4$  and absorbance was recorded at 235 nm against  $\text{H}_2\text{SO}_4$  blank. Obtained PHA was found highest at 36 h of incubation  $0.232 \pm 0.04$  g/1000 mL of production media. However, at 24 h and 48 h, it was found  $0.1 \pm 0.041$  and  $0.184 \pm 0.015$  g/1000 mL of media, respectively. In a study 0.5 g/L of PHA was extracted from the *Bacillus marcorestrictum* and mineral salt media containing  $\text{NH}_2\text{PO}_4$ ,  $\text{KH}_2\text{PO}_4$ ,  $\text{MgSO}_4 \cdot (\text{NH}_4)_2\text{SO}_4$ ,  $\text{CaCl}_2$ ,  $\text{NaCl}$ ,  $\text{NH}_4\text{Fe}$ , citrate, mixture of trace elements, 2% cane molasses along with 10 g/L glucose were used for PHA production [36]. In literature Choonut et al. [14] reported  $0.41 \pm 0.01$  g/L of PHA by thermotolerant bacteria *Bacillus thermoamylovorans*, isolated from palm mill effluent where the MS media (mineral salt) enriched with  $\text{Na}_2\text{HPO}_4 \cdot 12\text{H}_2\text{O}$  (9 g/L),  $\text{KH}_2\text{PO}_4$  (1.5 g/L), nitrogen stress in the form of 0.1 g/L  $\text{NH}_4\text{Cl}$ , and  $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$  were taken as the culture media. However, in the present study, 1% sugar from potato peel waste,  $\text{NaCl}$ , and 0.1 g/L tryptone were used in production media which is very economical as compared to above-mentioned study.

## Characterization of PHA

### FTIR

Characterization of extracted PHA from *Bacillus circulans* was done with FTIR analysis. The distribution of functional groups present in FTIR spectra was performed based on existing literature [31, 43]. PHA shows the band at  $3437.13 \text{ cm}^{-1}$  represents the hydroxyl group, the band at  $2925.31 \text{ cm}^{-1}$  represents C–H which shows the presence of methane and methylene groups, and sharp band at position  $1724.25 \text{ cm}^{-1}$  indicates

the presence of ester bond ( $C=O$ ) which is a characteristic feature of polyhydroxyalkanoates [25]. The  $1460.02\text{ cm}^{-1}$  represents  $CH_2$ , band position at  $1280.21$  corresponds to  $C-O$ , and  $1184.10\text{ cm}^{-1}$  corresponds to  $C-O-C$  (Fig. 3). Similar pattern of FTIR spectroscopy was obtained by *Bacillus megaterium* BBST<sub>4</sub> in a study done by Porras et al. [38]. Almost similar pattern of FTIR peaks was found in PHA accumulated by *Bacillus cereus* SS105 in which the spectral peak at  $3264\text{ cm}^{-1}$  represents the  $O-H$  group of polymer; for  $CH_3$  group peak, found at  $2973\text{ cm}^{-1}$ , an outcome of changes of crystalline structure and  $1736\text{ cm}^{-1}$  band was observed due to the presence of ester bond [32]. Hence, it is confirmed that the deposited polymer inside *Bacillus circulans* MTCC 8167 is PHA.

## DSC

A differential scanning calorimetry technique is done to check the melting temperature ( $T_m$ ) of extracted PHA. The peak obtained in the DSC curve at  $165.24\text{ }^\circ\text{C}$  indicates the crystallization of the PHA polymer (Fig. 4). Thermal characterization of PHA provides detail about the self-life of PHA and its stability. DSC of standard PHA ranges between  $165$  and  $170\text{ }^\circ\text{C}$ . Polymer degradation at low temperatures may be because of impurities. Although, the rate of degradation of polymer is species specific and may change upon using different extraction methods [25]. Cueva-Almendras [17] checked the DSC curve of PHA from isolated *Bacillus thuringiensis* and they found  $T_m$  at  $166.92\text{ }^\circ\text{C}$  close to the curve observed in

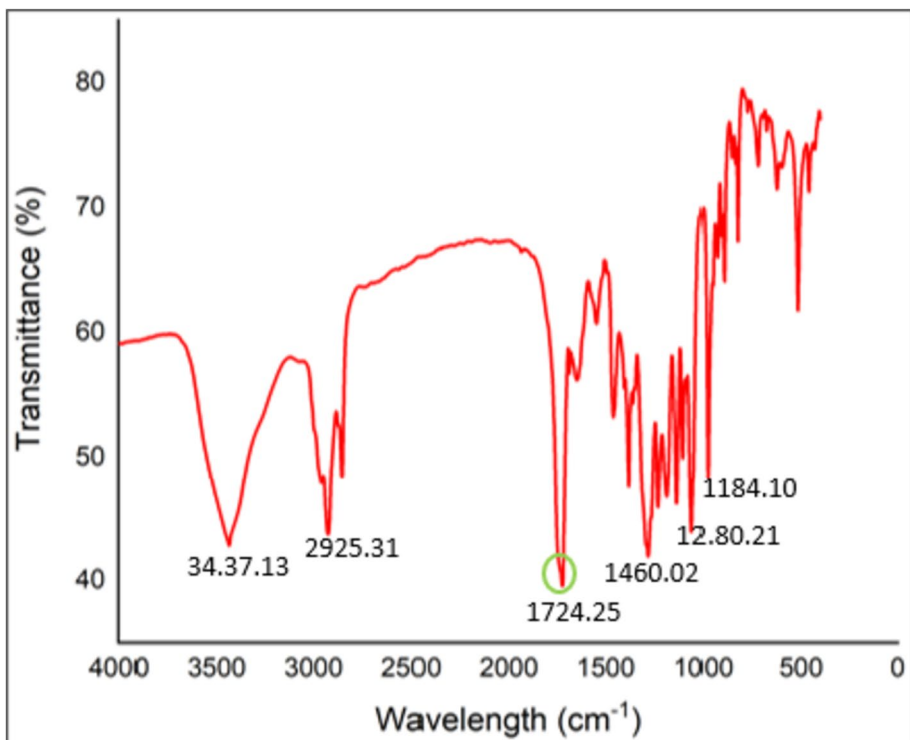


Fig. 3 FTIR spectra of purified PHA polymer

this study. Vu et al. [49] reported 145.43 °C  $T_m$ , which is lower than the present report from polymer accumulated by *Bacillus megaterium* using volatile fatty acid as a carbon source. Hassan et al. [22] observed the  $T_m$  at 175.9 °C of a polymer obtained from the *Bacillus sp.* which is higher from the present study.

## NMR

The presence of methyl and methylene groups is characteristic feature of PHA monomer. The intensity of signals obtained from proton NMR spectra of extracted biopolymer indicated the presence of PHA polymer. The doublet peak at 1.37 is an indication of the presence of the  $\text{CH}_3$  group, while the doublet quadruplet at the signal at 2.5 ppm is a characteristic feature of the presence of the  $\text{CH}_2$  group in the sample; moreover, peak intensity between 5.28 and 5.3 ppm is an indication of CH group. The peak at 7.2 ppm is attributed to solvents (denatured chloroform). By comparing the peak intensities with previous work, a similar pattern of proton NMR spectra was reported by Raghu and Divyashree [40] in PHA accumulated inside *Bacillus flexus* when it is grown on castor oil as a carbon source. Ensifer [26] identified the same type of peak pattern of proton NMR of PHA polymer accumulated by a soil isolate *Ensifer sp.* Moreover, the pattern of NMR obtained by the studies [22, 45] confirms the polymer obtained from *Bacillus circulans* using potato peel waste in this study is PHA (Fig. 5).

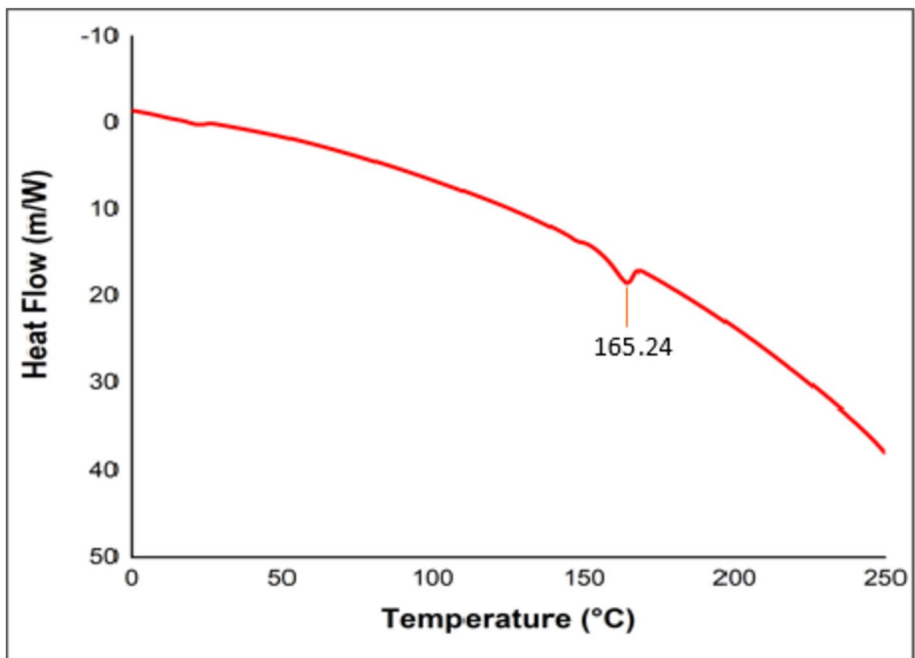
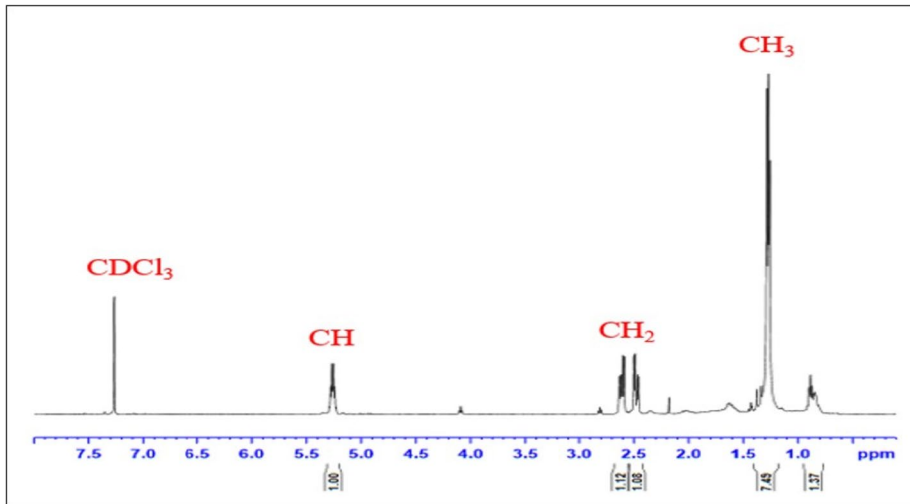


Fig. 4 DSC thermogram of purified PHA polymer



**Fig. 5** NMR spectroscopy of purified PHA polymer

## Conclusion and Future Prospective

PHA is a promising biopolymer for replacing traditional plastics. The extent of PHA deposition by microbes depends on several factors, including the microbial species' potency and its ability to utilize available carbon sources. Potato peel waste is utilized to liberate sugars, and the resulting reducing sugars serve as a carbon source for PHA accumulation by *Bacillus circulans*. During a 36-h incubation period, a substantial amount of PHA ( $0.232 \pm 0.04$  g/L) was extracted, and its characteristics were further analyzed using FTIR, DSC, and NMR. While achieving a desirable amount of CDW (cell dry weight) remains a challenge, this obstacle can potentially be overcome by optimizing growth parameters, such as the ideal quantity of carbon and nitrogen, to achieve a higher dry cell weight. Hence, potato peel waste shows promise as a cost-effective alternative carbon source for PHA production.

**Acknowledgements** The authors would like to thank the funding agencies Department of Biotechnology (DBT) grant no. BT/RLF/Re-entry/40/2017 and Science and Engineering Research Board (SERB) file no. EEQ/2020/000614, Govt. of India for providing the research grant to carry out this work. Also, special thanks to Council of Scientific & Industrial Research (CSIR), Govt. of India, for fellowship of Ms. Sonika. Authors would also like to thank Department of Chemistry, Delhi Technological University (DTU) for the support and providing testing facilities.

**Author Contribution** All authors read and approved the final manuscript, also confirm the contribution to the review article as follows: Experiment design and conceptualization of article: Sonika Kag and Rashmi Kataria; laboratory experiments and analysis: Sonika Kag; manuscript writing: Sonika Kag; Critically revised the work: Rashmi Kataria, Pravir Kumar, and Sonika Kag; Editing: Rashmi Kataria; Supervision: Pravir Kumar and Rashmi Kataria.

**Funding** Present work is funded by DBT (SAN No. 102/IFD/SAN/1276/2019-20) and SERB (File No. EEQ/2020/000614), Govt. of India.

**Data Availability** Not applicable.

## Declarations

**Ethical Approval** Not applicable.

**Consent to Participate** All authors are agreed to participate in the present manuscript.

**Consent for Publication** All authors have seen submitted version of manuscript and approved to publish.

**Competing Interests** The authors declare no competing interests.

## References

1. Abdelraof, M., Hasanin, M. S., & El-Saied, H. (2019). Ecofriendly green conversion of potato peel wastes to high productivity bacterial cellulose. *Carbohydrate Polymers*, 211, 75–83. <https://doi.org/10.1016/j.carbpol.2019.01.095>
2. Abedini, A., Amiri, H., & Karimi, K. (2020). Efficient biobutanol production from potato peel wastes by separate and simultaneous inhibitors removal and pretreatment, Renewable Energy. Elsevier Ltd. <https://doi.org/10.1016/j.renene.2020.06.112>
3. Ahmed, F., Yan, Z., & Bao, J. (2019). Dry biodegradation of acid pretreated wheat straw for cellulosic ethanol fermentation. *Bioresources and Bioprocessing*, 6. <https://doi.org/10.1186/s40643-019-0260-x>
4. Ahn, J., Jho, E. H., & Nam, K. (2015). Cupriavidus necator and its implication on the use of rice straw hydrolysates Effect of C / N ratio on polyhydroxyalkanoates ( PHA ) accumulation by Cupriavidus necator and its implication on the use of rice straw hydrolysates. <https://doi.org/10.4491/eeer.2015.055>
5. Andler, R., Pino, V., Moya, F., Soto, E., Valdés, C., & Andreeßen, C. (2021). Synthesis of poly-3-hydroxybutyrate (PHB) by *Bacillus cereus* using grape residues as sole carbon source. *International Journal of Biobased Plastics*, 3, 98–111. <https://doi.org/10.1080/24759651.2021.1882049>
6. Arapoglou, D., Varzakas, T., Vlyssides, A., & Israilides, C. (2010). Ethanol production from potato peel waste (PPW). *Waste Management*, 30, 1898–1902. <https://doi.org/10.1016/j.wasman.2010.04.017>
7. Ben Atitallah, I., Antonopoulou, G., Ntaikou, I., Alexandropoulou, M., Nasri, M., Mechichi, T., & Lyberatos, G. (2019). On the evaluation of different saccharification schemes for enhanced bioethanol production from potato peels waste via a newly isolated yeast strain of *Wickerhamomyces anomalus*. *Bioresource Technology*, 289, 121614. <https://doi.org/10.1016/j.biortech.2019.121614>
8. BezirhanArikan, E., & Bilgen, H. D. (2019). Production of bioplastic from potato peel waste and investigation of its biodegradability. *International Advanced Researches and Engineering Journal*, 03, 93–97. <https://doi.org/10.35860/iarej.420633>
9. Catherine, M., Guwy, A., & Massanet-nicolau, J. (2022). Bioresource technology reports effect of acetate concentration, temperature, pH and nutrient concentration on polyhydroxyalkanoates ( PHA ) production by glycogen accumulating organisms. *Bioresource Technology Reports*, 20, 101226. <https://doi.org/10.1016/j.biteb.2022.101226>
10. Celikci, N., Dolaz, M., & Colakoglu, A. S. (2020). An environmentally friendly carton adhesive from acidic hydrolysis of waste potato starch. *International Journal of Polymer Analysis and Characterization*, 0, 1–16. <https://doi.org/10.1080/1023666X.2020.1855047>
11. Chanasit, W., Hodgson, B., Sudesh, K., & Umsakul, K. (2016). Efficient production of polyhydroxyalkanoates (PHAs) from *Pseudomonas mendocina* PSU using a biodiesel liquid waste (BLW) as the sole carbon source. *Bioscience, Biotechnology, and Biochemistry*, 80, 1440–1450. <https://doi.org/10.1080/09168451.2016.1158628>
12. Chen, Y. J., Huang, Y. C., & Lee, C. Y. (2014). Production and characterization of medium-chain-length polyhydroxyalkanoates by *Pseudomonas mosselii* TO7. *Journal of Bioscience and Bioengineering*, 118, 145–152. <https://doi.org/10.1016/j.jbiosc.2014.01.012>
13. Chintagunta, A. D., Jacob, S., & Banerjee, R. (2016). Integrated bioethanol and biomanure production from potato waste. *Waste Management*, 49, 320–325. <https://doi.org/10.1016/j.wasman.2015.08.010>
14. Choonut, A., Prasertsan, P., Klomklao, S., & Sangkharak, K. (2020). *Bacillus thermoamylovorans* - related strain isolated from high temperature sites as potential producers of medium - chain - length polyhydroxyalkanoate ( mcl - PHA ). *Current Microbiology*, 77, 3044–3056. <https://doi.org/10.1007/s00284-020-02118-9>



15. Ciesielska, J. M., Dabrowska, D., Palasz, A. S., & Ciesielski, S. (2017). Medium - chain - length polyhydroxyalkanoates synthesis by *Pseudomonas putida* KT2440 relA / spoT mutant : Bioprocess characterization and transcriptome analysis. *AMB Express*. <https://doi.org/10.1186/s13568-017-0396-z>
16. Colombo, B., Favini, F., Scaglia, B., Sciarria, T. P., Imporzano, G. D., Pognani, M., Alekseeva, A., Eisele, G., Cosentino, C., & Adani, F. (2017). Biotechnology for biofuels enhanced polyhydroxyalkanoate ( PHA ) production from the organic fraction of municipal solid waste by using mixed microbial culture. *Biotechnology for Biofuels and Bioproducts*, 1–15. <https://doi.org/10.1186/s13068-017-0888-8>
17. Cueva-almendras, L. C. (2022). Production of polyhydroxyalkanoate by *Bacillus thuringiensis* Isolated from agricultural soils of Cascas-Peru. *Brazilian Archives of Biology and Technology*, 65, e22220107. <https://doi.org/10.1590/1678-4324-202220107>
18. Deshmukh, M., & Pande, A. (2022). Comparative Study for Production of biofuel from potato peel waste as feedstock by different enzymes. *II*, 1–6. <https://doi.org/10.35841/2329-6674.22.11.1000175>
19. Dinh, P., Minh, L., Trang, V., Minh, H., Thi, L., Phung, K., & Feng, D. (2022). Enrichment of hydrogen in product gas from a pilot-scale rice husk updraft gasification system. *Carbon Resources Conversion*, 5, 231–239. <https://doi.org/10.1016/j.crcon.2022.07.003>
20. Evangeline, S., & Sridharan, T. B. (2019). Biosynthesis and statistical optimization of polyhydroxyalkanoate ( PHA ) produced by *Bacillus cereus* VIT-SSR1 and fabrication of biopolymer films for sustained drug release. *International Journal of Biological Macromolecules*, 135, 945–958. <https://doi.org/10.1016/j.ijbiomac.2019.05.163>
21. Gao, C., Qi, Q., Madzak, C., Sze, C., & Lin, K. (2015). Exploring medium - chain - length polyhydroxyalkanoates production in the engineered yeast *Yarrowia lipolytica*. *Journal of Industrial Microbiology and Biotechnology*, 42, 1255–1262. <https://doi.org/10.1007/s10295-015-1649-y>
22. Hassan, M. A., Bakhiet, E. K., Ali, S. G., & Hussien, H. R. (2016). Production and characterization of polyhydroxybutyrate (PHB) produced by *Bacillus* sp. isolated from Egypt. *Journal of Applied Pharmaceutical Science*, 6, 46–51. <https://doi.org/10.7324/JAPS.2016.60406>
23. Hiremath, L., Sura, N., & Angadi, S. (2015). Design, screening and microbial synthesis of biopolymers of Poly-Hydroxy-Butyrate (PHB) from low cost carbons. *International Journal of Advanced Research*, 3(2), 420–425. <http://www.journalijar.com>
24. Hong, Z., Fen, X. U., Yu, W. U., Hong-hai, H. U., & Xiao-Feng, D. A. I. (2017). Progress of potato staple food research and industry development in China. *16*, 2924–2932. [https://doi.org/10.1016/S2095-3119\(17\)61736-2](https://doi.org/10.1016/S2095-3119(17)61736-2)
25. Joyline, M. (2019). Research Article Production and characterization of polyhydroxyalkanoates ( PHA ) by bacillus Megaterium strain JHA using inexpensive agro-industrial wastes Mascarenhas Joyline and Aruna K \* 10, 33359–33374. <https://doi.org/10.24327/IJRSR>
26. Khamkong, T., Penkhru, W., Lumyong, S. (2022). Optimization of Production of Polyhydroxyalkanoates (PHAs) from Newly Isolated Ensifer sp. Strain HD34 by Response Surface Methodology. *Processes* 2022;10, 1632. <https://doi.org/10.3390/pr10081632>
27. Khawla, B. J., Sameh, M., Imen, G., Donyes, F., Dhouch, G., Raoudha, E. G., & Oumèma, N. E. (2014). Potato peel as feedstock for bioethanol production: A comparison of acidic and enzymatic hydrolysis. *Industrial Crops and Products*, 52, 144–149. <https://doi.org/10.1016/j.indcrop.2013.10.025>
28. Landhäusser, S. M., Chow, P. S., Turin Dickman, L., Furze, M. E., Kuhlman, I., Schmid, S., Wiesenbauer, J., Wild, B., Gleixner, G., Hartmann, H., Hoch, G., McDowell, N. G., Richardson, A. D., Richter, A., & Adams, H. D. (2018). Standardized protocols and procedures can precisely and accurately quantify non-structural carbohydrates. *Tree Physiology*, 38, 1764–1778. <https://doi.org/10.1093/treephys/tpy118>
29. Liang, S., McDonald, A. G., & Coats, E. R. (2014). Lactic acid production from potato peel waste by anaerobic sequencing batch fermentation using undefined mixed culture. *Waste Management*, 45, 51–56. <https://doi.org/10.1016/j.wasman.2015.02.004>
30. Lima, M. de A., Andreou, R., Charalampopoulos, D., & Chatzifragkou, A. (2021). Supercritical carbon dioxide extraction of phenolic compounds from potato (*Solanum tuberosum*) peels. *Applied Sciences*, 11. <https://doi.org/10.3390/app11083410>
31. Mahato, R. P., Kumar, S., & Singh, P. (2021). Optimization of growth conditions to produce sustainable polyhydroxyalkanoate bioplastic by *pseudomonas aeruginosa* EO1. *Frontiers in Microbiology*, 12. <https://doi.org/10.3389/fmicb.2021.711588>
32. Maheshwari, N., Kumar, M., Thakur, I. S., & Srivastava, S. (2018). Production, process optimization and molecular characterization of polyhydroxyalkanoate (PHA) by CO2 sequestering *B. cereus* SS105. *Bioresource Technology*, 254, 75–82. <https://doi.org/10.1016/j.biortech.2018.01.002>
33. Malakar, B., Das, D., & Mohanty, K. (2020). Optimization of glucose yield from potato and sweet lime peel waste through different pre-treatment techniques along with enzyme assisted hydrolysis

- towards liquid biofuel. *Renewable Energy*, 145, 2723–2732. <https://doi.org/10.1016/j.renene.2019.08.037>
34. Martín, C., Christoph, J., Wei, M., Stagge, S., Xiong, S., & Jönsson, L. J. (2019). Industrial crops & products dilute-sulfuric acid pretreatment of de-starched cassava stems for enhancing the enzymatic convertibility and total glucan recovery. *Industrial Crops and Products*, 132, 301–310. <https://doi.org/10.1016/j.indcrop.2019.02.037>
  35. Miller, G. L. (1959). Use of dinitrosalicylic acid reagent for determination of reducing sugar. *Analytical Chemistry*, 31, 426–428. <https://doi.org/10.1021/ac60147a030>
  36. Narayankumar, S., Industries, K., & Krishnaswamy, V. G. (2020). Polyhydroxybutyrate production by *Bacillus marcorestinum* using polyhydroxybutyrate production by *Bacillus marcorestinum* using a cheaper substrate and its electrospun blends with polymer. <https://www.researchgate.net/publication/344167030>
  37. Pan, L., Li, J., Wang, R., Wang, Yu., Lin, Q., Li, C., & Wang, Y. (2021). Biosynthesis of polyhydroxyalkanoate from food waste oil by *Pseudomonas alcaligenes* with simultaneous energy recovery from fermentation wastewater. *Waste Management*, 131, 268–276. <https://doi.org/10.1016/j.wasman.2021.06.008>
  38. Porras, M. A., Cubitto, M. A., & Villar, M. A. (2014). Quantitative determination of intracellular PHA in *Bacillus megaterium* BBST4 strain Using Mid FTIR Spectroscopy. 1–4. <https://doi.org/10.13140/RG.2.1.3920.2407>
  39. Prakash, P., Lee, W.-H., Loo, C.-Y., Wong, H. S. J., & Parumasivam T. (2022). Advances in polyhydroxyalkanoate nanocarriers for effective drug delivery: An overview and challenges. *Nanomaterials*, 12, 175. <https://doi.org/10.3390/nano12010175>
  40. Raghu, M. G. H., & Divyashree, C. M. S. (2021). Statistical optimisation of polyhydroxyalkanoate production in *Bacillus endophyticus* using sucrose as sole source of carbon. *Archives of Microbiology*, 203, 5993–6005. <https://doi.org/10.1007/s00203-021-02554-6>
  41. Remedios, Y., & Domingues, L. (2023). Potato peels waste as a sustainable source for biotechnological production of biofuels: Process optimization. 155, 320–328. <https://doi.org/10.1016/j.wasman.2022.11.007>
  42. Sampaio, S. L., Petropoulos, S. A., Alexopoulos, A., Heleno, S. A., Santos-buelga, C., Barros, L., & Ferreira, I. C. F. R. (2020). Trends in Food Science & Technology Potato peels as sources of functional compounds for the food industry : A review. *Trends in Food Science & Technology*, 103, 118–129. <https://doi.org/10.1016/j.tifs.2020.07.015>
  43. Shah, K. R. (2012). FTIR analysis of polyhydroxyalkanoates by novel *Bacillus* sp. AS 3–2 from soil of Kadi region, North Gujarat, India. *Journal Of Biochemical Technology*, 3, 380–383.
  44. Sluiter, J. B., Ruiz, R. O., Scarlata, C. J., Sluiter, A. D., & Templeton, D. W. (2010). Compositional analysis of lignocellulosic feedstocks. Review and description of methods. *Journal of Agricultural and Food Chemistry*, 58(16), 9043–9053. <https://doi.org/10.1021/jf1008023>
  45. Sriyapai, T., Chuarung, T., Kimbara, K., Samosorn, S., & Sriyapai, P. (2022). Production and optimization of polyhydroxyalkanoates (PHAs) from *Paraburkholderia* sp. PFN 29 under submerged fermentation. *Electronic Journal of Biotechnology*, 56, 1–11. <https://doi.org/10.1016/j.ejbt.2021.12.003>
  46. Tesfaw, A. A., & Tizazu, B. Z. (2021). Reducingsugarproductionfromteffstrawbiomassusingdilute sulfuric acid hydrolysis: Characterization and optimization using response surface methodology. *International Journal of Biomaterials*, 2021. <https://doi.org/10.1155/2021/2857764>
  47. Timung, R., Naik Deshavath, N., Goud, V. V., & Dasu, V. V. (2016). Effect of subsequent dilute acid and enzymatic hydrolysis on reducing sugar production from sugarcane bagasse and spent citrionella biomass. *Journal of Energy*, 2016, 1–12. <https://doi.org/10.1155/2016/8506214>
  48. Valencia, A. I. S., Rojas, U., & Fajardo, C. (2021). Effect of C / N ratio on the PHA accumulation capability of microbial mixed culture fed with leachates from the organic fraction of municipal solid waste ( OFMSW ). *Journal of Water Process Engineering*, 40. <https://doi.org/10.1016/j.jwpe.2021.101975>
  49. Vu, D. H., Wainaina, S., Taherzadeh, M. J., Åkesson, D., & Ferreira, J. A. (2021). Production of polyhydroxyalkanoates (PHAs) by *Bacillus megaterium* using food waste acidogenic fermentation-derived volatile fatty acids. *Bioengineered*, 12, 2480–2498. <https://doi.org/10.1080/21655979.2021.1935524>
  50. Wang, J., Liu, S., Huang, J., Cui, R., Xu, Y., & Song, Z. (2023). Environmental Technology & Innovation Genetic engineering strategies for sustainable polyhydroxyalkanoate ( PHA ) production from carbon-rich wastes. *Environmental Technology and Innovation*, 30, 103069. <https://doi.org/10.1016/j.eti.2023.103069>
  51. Zhou, C., Jiang, W., Via, B. K., Fasina, O., & Han, G. (2015). Prediction of mixed hardwood lignin and carbohydrate content using. *Carbohydrate Polymers*, 121, 336–341. <https://doi.org/10.1016/j.carbpol.2014.11.062>

52. Zhou, W., Irene, D., Geurkink, B., Euverink, G. W., & Krooneman, J. (2022). Science of the Total Environment The impact of carbon to nitrogen ratios and pH on the microbial prevalence and polyhydroxybutyrate production levels using a mixed microbial starter culture. *Science of the Total Environment*, 811, 152341. <https://doi.org/10.1016/j.scitotenv.2021.152341>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

# Power Flow Management of Solar PV fed Switched Boost Inverter

Srishti Singh  
(Student Member IEEE)

CoE for EVRT  
Electrical Engineering Department  
Delhi Technological University  
Delhi, India  
srishhtisingh@gmail.com

Vansh Aggarwal  
(Student Member IEEE)

CoE for EVRT  
Electrical Engineering Department  
Delhi Technological University  
Delhi, India  
aggarwalvansh14@gmail.com

Mayank Kumar  
(Senior Member, IEEE)

CoE for EVRT  
Electrical Engineering Department  
Delhi Technological University  
Delhi, India  
mayankkumar@dtu.ac.in

**Abstract**— Photovoltaic (PV) power generation is vastly dependent on solar irradiance and temperature conditions. In this paper, the temperature is chosen at standard test condition (STC) i.e.,  $T = 25^{\circ}\text{C}$ , whereas variable irradiance is considered for power flow management. The maximum power point technique (MPPT) algorithm is implemented to concur optimal power from solar photovoltaic array, which is fed via the dc-dc boost converter to a variable dc loading system. A single stage switched boost inverter (SBI) is used for high gain dc-ac power conversion. The SBI is used to enable the ac loads for the operation at high voltage gain with 72 V dc bus. A battery energy storage system (BESS) is integrated and the performance of battery is measured in terms of state of charge (SOC) under different irradiance conditions. The bidirectional converter serves to regulate dc bus voltage whilst maintaining the power balance during deficit and surplus PV power generation conditions. Simulation results are provided to verify proposed microgrid network and power flow management. The results are in good agreement with theoretical analysis.

**Keywords**— Bidirectional converter, maximum power point technique (MPPT), perturb and observe (P&O), solar PV, state of charge (SOC), switched boost inverter (SBI).

## I. INTRODUCTION

The advancement in renewable energy sources (RES) is a fairly accepted alternative in contemporary times as compared to the conventional energy sources. Solar photovoltaic array promises to be the leading source of energy. This has become possible broadly based on the ability to produce solar energy locally with more advanced versions of solar modules that can be installed at grassroot level, reducing the dependence on imported and fumigating fuels thereby increasing energy security. This has been encouraged in the form of multiple government incentives promoting use of renewable energy through subsidies, tax credits etc. bringing down the overall cost associated with solar energy in recent years.

A microgrid is a cumulative circuit, representing power sources and attached loads. The microgrid has a distinguishing feature that it operates connected to a traditional centralized grid, but subsystems have the capacity to disconnect and function as an autonomous unit. Solar energy has huge potential which is reflected by the value of solar energy constant i.e.  $1373 \text{ W/m}^2$  which is a quantitative measure of sun's radiated power density calculated at the outer atmospheric layer. This radiated power density is subjected to scattering and absorption and finally a tropical-zone solar irradiation peak density of  $1000 \text{ W/m}^2$  is achieved at the surface of earth. This value has been chosen as the case for constant irradiance. Utilising a PV array involves connecting

multiple low power profile solar PV cells to obtain an aggregated output of desired current and voltage levels. The PV based microgrid needs to be integrated to other energy sources or ac grid since the PV power output is intermittent in nature.

This paper considers a 72V dc bus to which a PV array of 1340 W is integrated via a boost dc-dc converter. The duty cycle of boost converter is controlled using perturb and observe (P&O) algorithm as a maximum power point technique (MPPT) algorithm so that the PV arrays operate at their maximum power point. A lithium-ion battery of rating 35 Ah, 48V and having initial state of charge (SOC) 50% is also integrated to the system. A bidirectional converter is used to integrate this Li-ion battery to the considered system and regulate the SOC of battery [1]. A switched boost inverter (SBI) which is used as high gain inverter is employed, which can supply both ac as well as dc loads simultaneously [2], [3].

This paper broadly covers design and analysis of three controllers viz dc-dc boost controller; bidirectional dc-dc controller and SBI control. The MPPT controller is designed to extract the maximum power from rated PV array and deliver to the dc bus [4]. The SBI control provides improved power quality at the user end. Simultaneously, the bidirectional controller allows adequate energy management and provides user flexibility in terms of charging or discharging ability. The paper thoroughly discusses all three controllers and overall system to provide user access to higher power quality and control. The novelty of this paper is the use of high gain inverter which enable the ac loads for the operation with 72 V dc bus.

## II. SWITCHED BOOST INVERTER BASED MICROGRID FOR AC/DC LOADS

The system schematic in Fig.1 is proposed in this paper. PV array is composed of 4 modules each having a capacity of 335 W is connected in parallel therefore the installed capacity of 1340 W. PV based power generation is considered under two assumed conditions. First condition describes a hypothetical case of mean irradiance  $1000 \text{ W/m}^2$  incident over the solar array, while a second case is modelled to imitate reality therefore irradiance is considered to be varying over 24 hours span and is incident upon SunPower SPR-X21-335-BLK module at  $25^{\circ}\text{C}$  specified temperatures. The plot for PV array suggests that the maximum power that can be obtained using the module is 1340 W at 57.3 V and it is capable to deliver a current of magnitude 23.4 A. In order to obtain these desired values, P&O MPPT algorithm is used. The code for the P&O algorithm is appended to the function block. Its working principle is based upon the observation that a change

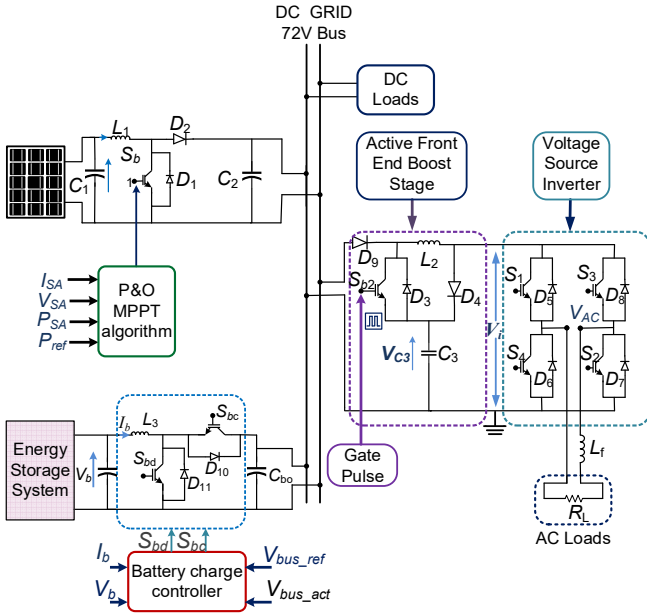


Fig. 1. Proposed system layout

in operating voltage produces a corresponding change in power output. Following this algorithm, the operating voltage can be adjusted again until the MPP is found. The MPP is the point at which the solar panel and the designed system operates at maximum power. The major advantages drawn out of employing P&O technique are that, since it does not require prior knowledge of the solar panel's characteristic curve, it acts as a simple, robust, cost-effective and efficient solution for tracking the maximum power point in photovoltaic systems. Battery energy storage system (BESS) controlling mechanism is implemented to the amount and timing of energy released or absorbed from the electrical grid, improving grid stability and reliability. In Fig. 1, The difference of  $V_{bus\_ref}$  (i.e., reference voltage assumed to be 72V) and  $V_{bus\_act}$  (i.e., actual voltage) is fed to a PID controller to obtain the value of reference current. The BESS commences alternate cycles of charging the batteries when excess energy is available from source to the grid and discharging when demand at the load ends is high. This process reduces the probability of fault occurring due to

TABLE I.  
SYSTEM PARAMETERS

System Components	Parameters	Values
PV Array	Output Power at MPP**	1340 W
	Open Circuit Voltage $V_{oc}$	67.90 V
	Short Circuit Current $I_{sc}$	6.23 A
	Peak Power Voltage $V_{MPP}$	57.30 V
	Peak Power Current $I_{MPP}$	23.4 A
Switched Boost Inverter	Input Voltage ( $V_{dc}$ )	72 V
	Fundamental Frequency	50 Hz
	Switching Frequency	10 kHz
	Shoot through Duty Ratio	0.4
	Modulation Index ( $M$ )	0.5
	Inductance( $L_2$ )	5.6 mH
	Capacitance( $C_3$ )	100 $\mu$ F
	Outer Filter Inductor ( $L_F$ )	4.6 mH
Bidirectional Converter	Load Resistance ( $R_L$ )	100 $\Omega$
	Battery Nominal Voltage	48 V
	Rated Capacity	35 Ah
	Initial SOC of Battery	50%
	Inductance ( $L_3$ )	5 mH
	Capacitance ( $C_{bo}$ )	220 $\mu$ F

\*\* at STC (standard test conditions) i.e., 25°C and 1000 W/m<sup>2</sup>

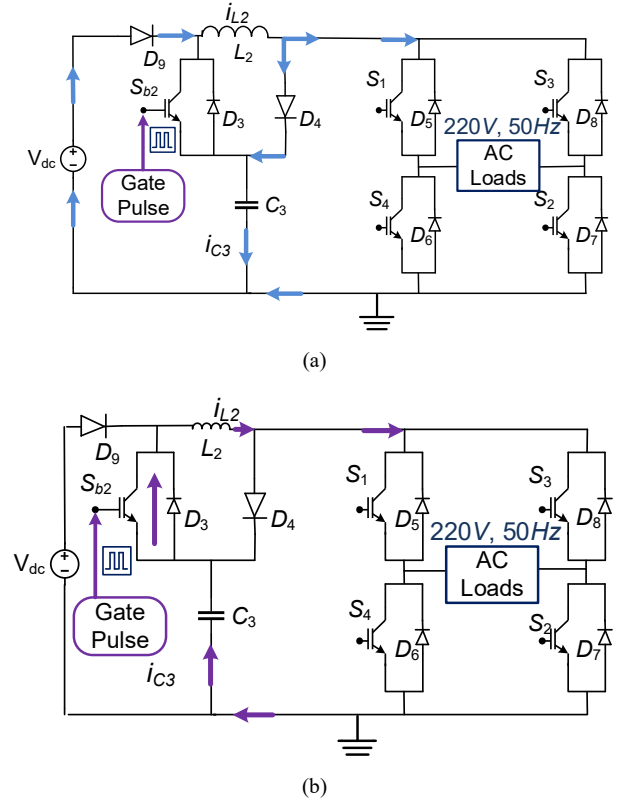


Fig. 2. Operation of SBI during (a) charging cycle,  $(1-D) \times T_s$  time period; and (b) discharging cycle,  $D \times T_s$  period.

regulated power flow from source to end and provides an excellent alternative in situations where source supply failure occurs so that pressure on load is reduced. The BESS presented in Fig.1; Li-ion battery pack is considered which is connected with the bidirectional converter. Charging and discharging operations take place between 72V dc bus and BESS via the bidirectional converter based upon the SOC of the battery.

SBI can be understood as a combination of two sub-circuits, a switched boost network and a voltage source inverter (VSI). As shown in Fig. 2 the switched boost network consists of a switch ( $S_{b2}$ ), two diodes ( $D_9$ ,  $D_4$ ), inductor ( $L_2$ ) and one capacitor ( $C_3$ ) [5]. SBI is assumed to be in shoot through state for the duration  $D \times T_s$  within the switching cycle  $T_s$ . Fig. 2. (b) shows that when switch  $S_{b2}$  is turned on, it makes path for current to flow from the  $C_3$  to  $L_2$  via switch  $S_{b2}$  since the inverter gets shorted due to shoot through state.  $L_2$  gets charged during this time by discharging of capacitor  $C_3$  and the inductor current ( $i_{L2}$ ) equals the capacitor discharging current ( $i_{C3}$ ).  $D_4$  and  $D_9$  are reverse biased during this time since the steady state shoot through voltage across capacitor ( $V_{C3}$ ) is greater than  $V_{dc}$ . Fig. 2. (a) represents the remaining duration of the cycle i.e.  $(1-D) \times T_s$  where the inverter is not in shoot through state and switch  $S_{b2}$  remains turned off [6]. For simplification, the VSI can be represented as a current source during this time. The  $V_{dc}$  and  $L_2$  (which was charged during  $D \times T_s$ ) supply power to both inverter and  $C_3$  through  $D_9$  and  $D_4$ . During this time  $i_{L2}$  is equal to the sum of capacitor charging current ( $i_{C3}$ ) and inverter input current, being fed to ac loads. It can be inferred that the conversion ratio ( $V_{C3}/V_{dc}$ ) is unity at  $D = 0$  because shoot through state is not attained but  $D$  is very high when duty ratio approaches 0.5 and cannot be further exceeded, like the case of Z source

inverter. [7]. Relation between  $D$  and  $M$  is derived in [3]. Applying the voltage second balance,

$$(V_{C3} \times D) + (V_{DC} - V_{C3}) \times (1 - D) = 0 \quad (1a)$$

$$\frac{V_{C3}}{V_{DC}} = \frac{(1 - D)}{(1 - 2D)} \quad (1b)$$

### III. PWM GENERATION AND POWER FLOW MANAGEMENT

The overall system in Fig. 1 can be understood as an aggregate of three subsystems, each with their respective control. This provides us with three subsystem block diagrams consisting of the boost converter, the bidirectional converter and the inverter as shown in Fig. 3. Of these, the boost converter control has been coded using the function block implementing the P&O MPPT algorithm.

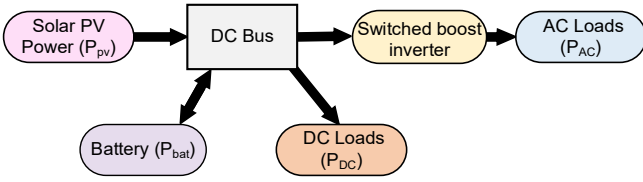


Fig. 3. Power flow control in the proposed system.

Energy management of a PV integrated dc microgrid with a BESS involves controlling and coordinating the flow of energy between the battery and energy source in a way to allow maximum effective utilization of power, so as to meet current load demands as well as make viable storage options to support the system during non-generative hours. The scheme displayed underneath explains the mechanism through which control of BESS subsystem is achieved.  $I_{ref}$  is evaluated by comparing the values of  $V_{bus\_ref}$  which is the reference voltage assumed to be 72V and  $V_{bus\_act}$  as the varying voltage considered at an instant and feeding the result obtained to a PID controller. This value of  $I_{ref}$  obtained is then compared with the battery current value and fed to a PI controller. The output is subjected to a saturation block that is fed to the PWM generator from which boost and buck operation control can be achieved using a NOT gate for the buck output. The aforementioned systems are represented in the form of exclusive block diagrams as shown in Fig. 4 (a).

A switched boost inverter is a type of dc-ac power inverter that uses switching devices such as IGBT is used to regulate the input dc voltage and convert it to ac power. The term "boost" in the name refers to the fact that the inverter increases the voltage of the input dc power source before converting it to ac power. This allows for a higher output power than traditional inverters, but also requires a higher input voltage and a larger circuit. Due to the high frequency of the switching devices used in the SBI, stress on the components of the inverter is reduced which further leads to increased lifespan of the inverter.

SBI also faces many challenges in its usage. One of the major challenges observed are harmonic distortion and electromagnetic interference (EMI). The switching operation of the inverter can cause harmonic distortion in the output ac waveform, which can lead to poor power quality and damage to other electrical equipment. The switching operation of the inverter can also generate EMI, which can cause problems with other electronic devices and equipment. Due to this, we

can control the efficiency and reliability of the SBI using certain control strategies. One such control strategy is discussed in this paper.

SBI uses shoot-through state (both switches of the voltage converter are ON at the same time) to invoke boost operation. Since traditional PWM techniques do not permit shoot through state for a VSI, the PWM control strategy used in this paper is based on sine-triangle pulse width modulation with unipolar voltage switching as shown in Fig. 4 (b) [8]-[9].

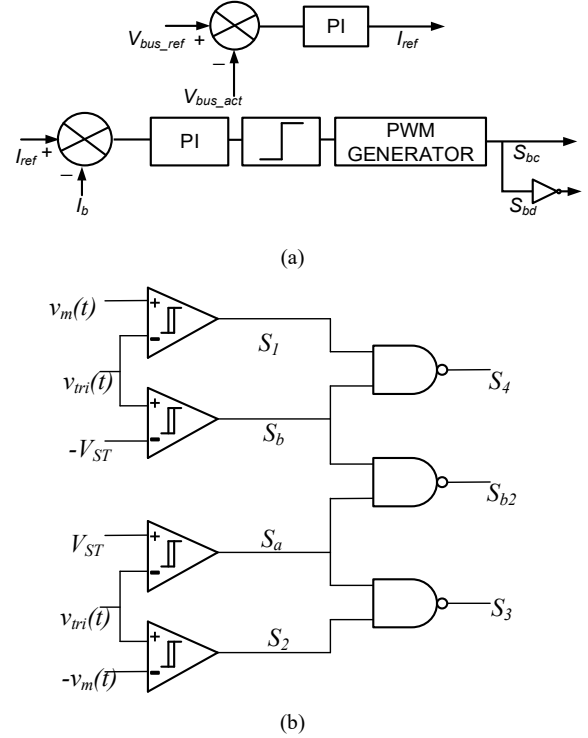


Fig. 4. (a) Bidirectional converter control; and (b) schematic of PWM control circuit of SBI.

Gate signals for  $S_1$  and  $S_2$  are generated by comparing the reference sinusoidal signal  $V_m(t)$  with a high frequency switching / triangular carrier wave  $V_{tri}(t)$ . Frequency of the carrier wave is chosen such that it is much greater than the frequency of the modulation signal, thus,  $V_m(t)$  is assumed to be nearly constant. The signals  $S_a$  and  $S_b$  are generated by comparing  $V_{tri}(t)$  with two constant voltages  $V_{ST}$  and  $-V_{ST}$ , respectively.  $S_{b2}$ ,  $S_3$  and  $S_4$  are generated using these two signals.

$$S_3 = \overline{S_2} \wedge S_a; S_4 = \overline{S_1} \wedge S_b; S_{b2} = \overline{S_a} \wedge S_b \quad (2)$$

The output voltage of the inverter side of the SBI has nine switching intervals in each switching cycle during which different conducting devices are active as can be seen from Table II. From Fig. 5. we can observe that the duty ratio  $D$  can be varied by varying  $V_{ST}$ . Using Fig. 5 it can be derived as follows:

$$V_{tri}(t) = \begin{cases} \frac{-\hat{V}_{tri}}{(T_s / 4)} \left( t - \frac{T_s}{4} \right); & \text{if } 0 < t < \frac{T_s}{2} \\ \frac{\hat{V}_{tri}}{(T_s / 4)} \left( t - \frac{3T_s}{4} \right); & \text{if } \frac{T_s}{2} < t < T_s \end{cases} \quad (3)$$



$$V_{tri}(t_1) = V_{tri}(t_2) = -V_{ST}; \text{ and } t_2 - t_1 = \frac{DT_s}{2} \quad (4)$$

Using (3) and (4),  $t_1$  and  $t_2$  can be calculated as follows:

$$t_1 = \frac{T_s}{4} \left( 1 + \frac{V_{ST}}{\hat{V}_{tri}} \right) \text{ and } t_2 = \frac{T_s}{4} \left( 3 - \frac{V_{ST}}{\hat{V}_{tri}} \right) \quad (5)$$

The duty ratio  $D$  is chosen such that the shoot through interval does not disturb the power interval of  $V_{ac}$  i.e.,  $D$  is chosen such that the total available width of zero interval in any switching cycle should be greater than total width of shoot through interval.

TABLE II.  
CONDUCTING DEVICES WITHIN RESPECTIVE INTERVALS

Interval Number	Conducting Devices of the Converter
(1), (9)	$S_2, S_3, S_4, S_{b2}$
(2), (8)	$S_2, S_4$
(3), (7)	$S_1, S_2$ ( $S_3, S_4$ in negative half cycle of $V_m(t)$ )
(4), (6)	$S_1, S_3$
(5)	$S_1, S_3, S_4, S_{b2}$

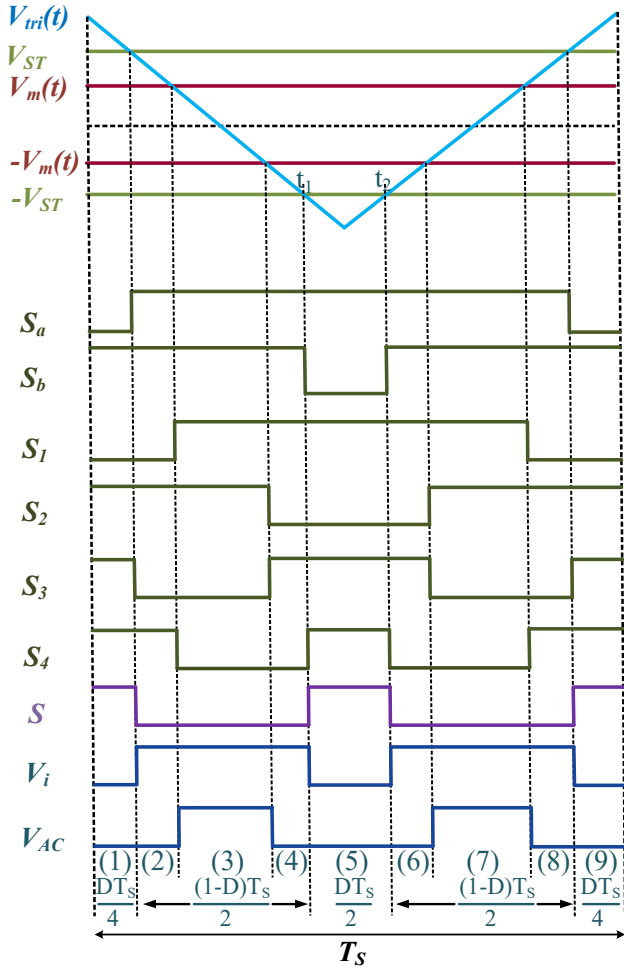


Fig. 5. PWM control signals generated for switched boost inverter during positive half cycle of  $V_m(t)$ .

#### IV. RESULTS AND DISCUSSION

The energy storage system integrated with the solar PV array with isolated dc microgrid system has been analyzed using MATLAB software. The results have been discussed under variable irradiance and constant irradiance conditions.

A. *Under variable irradiance condition:* The irradiance condition is considered for 24 hours for a day scenario.

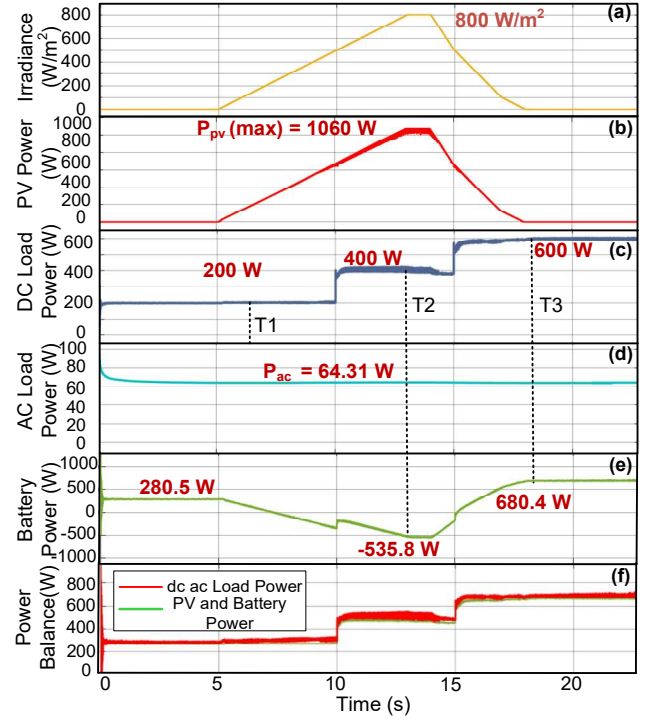


Fig. 6. a) Variable Irradiance for 24 hours b) PV Power c) dc power across load d) ac Power across load e) battery Power f) load demand and supply.

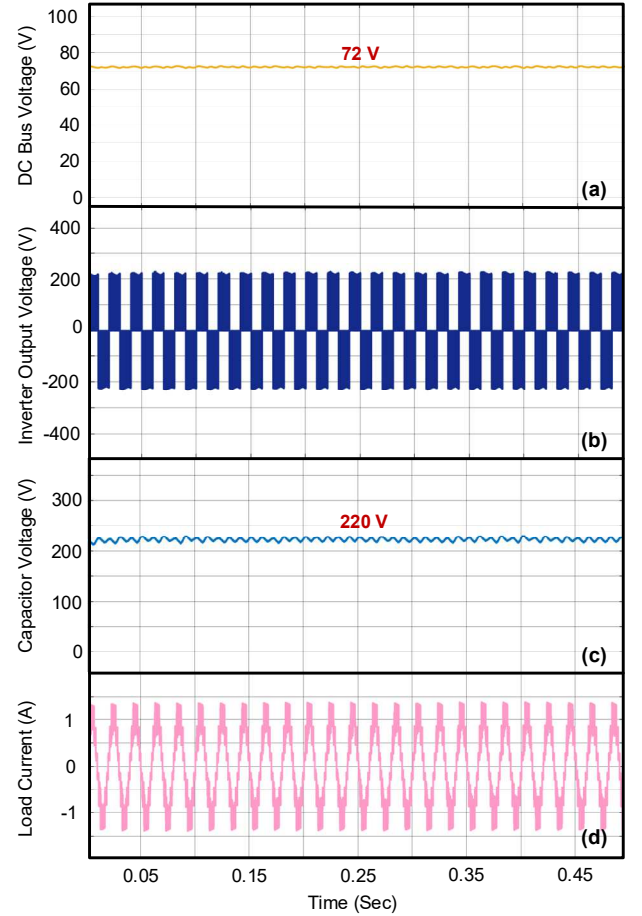


Fig. 7. Switched boost inverter waveforms, a) DC bus voltage, b) inverter output voltage, c) capacitor voltage, and d) load current.

Fig. 6. (a) represents the variable irradiance provided to the PV array across an entire day. The power provided by the PV array is represented by Fig. 6. (b) the maximum power obtained is equivalent to 1060 W. Fig. 6. (c) The voltage gets boosted by a boost converter which is controlled using P&O MPPT algorithm so that the PV arrays operate at their maximum power point, then power is supplied to the loads. The dc load is fluctuating, the load is increased at 10 and 15 hours for which the dc powers are 200 W (T1), 400 W (T2), 600 W (T3) respectively. This has been plotted as dc load power diagram. Fig. 6. (d) represents the ac power drawn across constant ac load which is maintained at 64.31 W. Fig. 6. (e) represents the power supplied by the battery. Battery power is both negative (battery charging mode) and positive (battery discharging mode) i.e., 280.5 W(T1), -535.8 W(T2) and 680.4 W(T3). Fig. 6. (f) represents the power flow management of the system. The total load power (ac and dc) is approximately equal to the power supplied by the PV array and the BESS.

The SBI derives voltage from 72 V dc bus shown in Fig. 7. (a). This voltage gets boosted and subsequently inverted by SBI and it is obtained as Fig. 7. (b) across the ac loads ( $L_f$ ,  $R_L$ ) as inverter output voltage. When the SBI is suitably controlled by applying PWM techniques Fig. 7. (c) 220V is achieved across the capacitor ( $C_3$ ) and Fig. 7. (d) ac current / load current of the SBI.

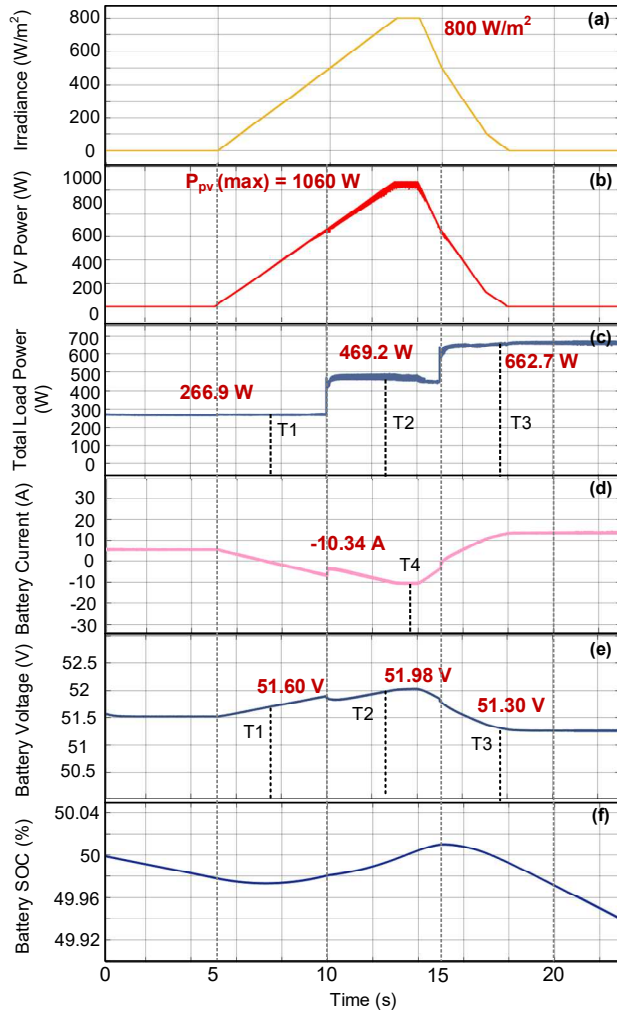


Fig. 8. Characteristics of BESS, a) variable irradiance for 24 hours b) PV power c) total power d) battery current e) battery voltage f) battery SOC.

Fig. 8. (a) and (b) represents the irradiance and the PV power respectively obtained. The results presented show the BESS characteristics. Fig. 8. (c) represents the total load power 266.9 W (T1), 469.2 W (T2) and 662.7W (T3). It is the summation of the dc load power and the ac load power for which the dc loads are fluctuating at intervals of 10 and 15 hours. Fig. 8. (d) represents the current provided by the battery. Battery current reaches a maximum value of -10.34 A at time T4 (negative sign represents that the current is flowing into the battery i.e., battery is getting charged). Fig. 8. (e) shows the battery voltage at different times: 51.60 V (T1), 51.98 V (T2) and 51.30 V (T3). Fig. 8. (f) represents the battery SOC. The increasing and decreasing characteristics of SOC depends on the availability of solar irradiance with respect to total load demand.

B. *Under Constant Irradiance Condition:* Fig. 9. (a) represents a constant irradiance of 1000W/m<sup>2</sup> provided to the PV array. Fig. 9. (b) is the power provided by the PV array under constant irradiance i.e., 1330 W. Fig. 9. (c) represents the dc load power which is fluctuating due to the fluctuating loads connected at 10 and 15 hours. DC power varies as 200 W at T1, 400W at T2 and 600 W at T3. Fig. 9. (d) shows ac power across the load, which is considered constant throughout the simulation study at the value of 64.20. Fig. 9. (e) represents the power supplied by the battery. Battery power is negative (battery charging mode) due to constant irradiance at all times and varies as -981.8 W (T1), -793.5 W(T2) and -582.5 W (T3) due to variable loading. Fig. 9. (f) displays the power flow management of the system as total load ac and dc load power is equal to the total power supplied by the PV array and the BESS.

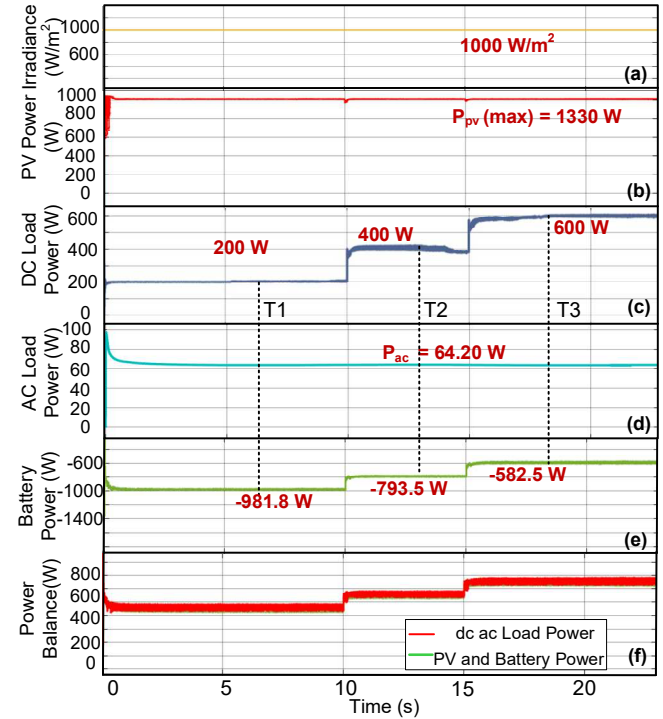


Fig. 9. Power flow management under constant radiance condition, a) constant irradiance for 24 hours, b) PV power, c) dc power across load, d) ac load power, e) battery power, f) total load demand and supply.

Fig. 10. (a) and (b) show the irradiance and the PV power respectively Fig. 10. (c) represents the variation of total load power at different points in time when dc load is varied at an

interval of 10 and 15 hours respectively whereas ac load remains constant throughout. Total load power obtained is 261.3 W till the first 10 hours of the day (T1), it escalates to 475.9 W at a point midway between 10 and 15 hours (T2) and reaches a level of 685.4 W after 15 hours (T3). In that duration, battery gets charged by the PV array as the current is supplied to it. Fig. 10. (d) Battery Current varies as the load is varied: -18.66 A (T1), -15.00 A (T2), -11.40 A (T3). Fig. 10. (e) The battery voltage is maintained at 52.2 V. Fig. 10. (f) Since the battery is only getting charged under constant irradiance, the battery state of charge should keep on increasing. This is verified by the simulation results obtained.

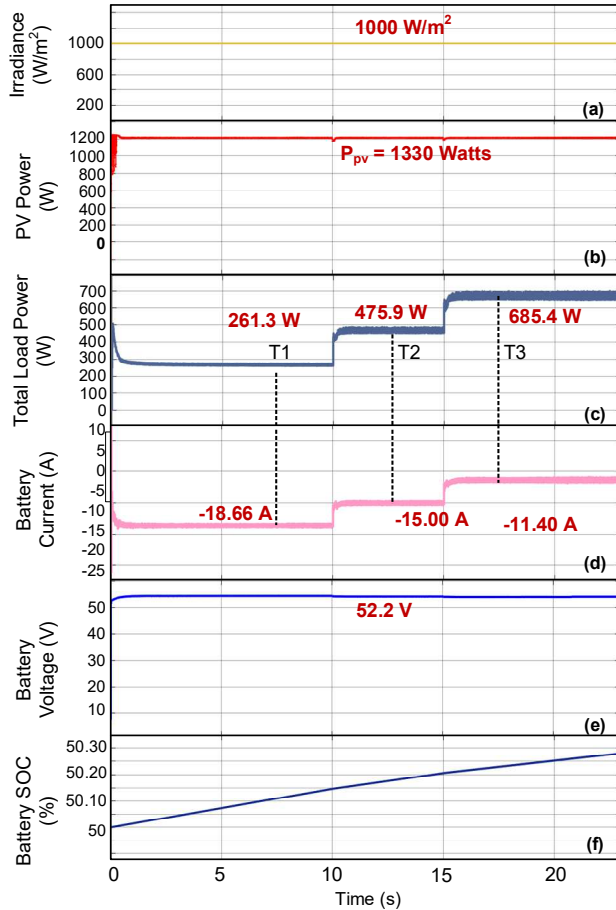


Fig. 10. Characteristics of BESS under constant irradiance condition, a) constant irradiance, b) PV power, c) total load power, d) battery current, e) battery voltage, f) battery SOC.

## V. CONCLUSION

The proposed Solar PV switched boost inverter based microgrid system allows renewable solar energy integration, which is considered at 25°C and at variable irradiance. It is

observed that the SOC of the battery increases when the load demand is lower than the total PV generated power and therefore the bidirectional converter is operating under buck-mode (charging mode). However, the bidirectional converter is employed to provide energy during hours when solar irradiance is insufficient to compensate the load demand (i.e., discharging mode). The SBI obtains voltage from dc bus at 72 V and boosts it to 220 V by charging of capacitor. This boosted voltage is inverted by the voltage source inverter component of SBI and used to provide to ac loads. The active front end boost stage is used to improve the bus voltage and attain higher value of rms output voltage that can be used to supply to ac loads. Therefore, the SBI enables the ac loads for the operation with high voltage gain with the input of 72 V dc bus. The performance of BESS is evaluated in terms of SOC under different irradiance conditions.

## ACKNOWLEDGMENT

This research supported by the Science and Engineering Research Board (SERB), Department of Science & Technology, Government of India, under the SERB sanction order number SRG/2021/001640.

## REFERENCES

- [1] S. Upadhyay, R. Adda, S. Mishra and A. Joshi, "A switched-boost topology for renewable power application," *2010 Conference Proceedings IPEC, Singapore*, 2010, pp. 758-762, doi: 10.1109/IPEC.2010.5697026.
- [2] S. Acharya and S. K. Mishra, "A Review of High Gain Inverters for Smart-grid Applications," *2020 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES)*, Jaipur, India, 2020, pp. 1-6, doi: 10.1109/PEDES49360.2020.9379522.
- [3] R. Adda, S. Mishra and A. Joshi, "A PWM control strategy for switched boost inverter," *2011 IEEE Energy Conversion Congress and Exposition, Phoenix, AZ, USA*, 2011, pp. 991-996, doi: 10.1109/ECCE.2011.6063880.
- [4] M. Kumar, "Solar PV Based DC Microgrid under Partial Shading Condition with Battery-Part 2: Energy Management System," *2018 8th IEEE India International Conference on Power Electronics (IICPE)*, Jaipur, India, 2018, pp. 1-6, doi: 10.1109/IICPE.2018.8709437.
- [5] V. Anusree and P. Saifunnisa, "Closed loop control of switched boost inverter," *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, Chennai, India, 2016, pp. 3040-3044, doi: 10.1109/ICEEOT.2016.7755259.
- [6] S. S. Nag and S. Mishra, "Current-Fed DC/DC topology based inverter," *2013 IEEE Energy Conversion Congress and Exposition, Denver, CO, USA*, 2013, pp. 2751-2756, doi: 10.1109/ECCE.2013.6647057.
- [7] Fang Zheng Peng, "Z-source inverter," *IEEE Trans. Ind. Appl.*, vol. 39, no. 2, pp. 504-510, March-April 2003, doi: 10.1109/TIA.2003.808920.
- [8] A. Ravindranath, S. K. Mishra and A. Joshi, "Analysis and PWM Control of Switched Boost Inverter," *IEEE Trans. Ind. Electron.*, vol. 60, no. 12, pp. 5593-5602, Dec. 2013, doi: 10.1109/TIE.2012.2230595.
- [9] R. Gautam, R. V. John and M. Kumar, "Cascaded H-Bridge Multilevel Inverter Based Solar PV Power Conversion System," *2022 IEEE Students Conference on Engineering and Systems (SCES)*, Prayagraj, India, 2022, pp. 1-6, doi: 10.1109/SCES55490.2022.9887731.

# Pressure Induced Surface States and Wannier Charge Centers in Ytterbium Monoarsenide

Ramesh Kumar, Rajesh Kumar, Sangeeta & Mukhtiyar Singh\*

Department of Applied Physics, Delhi Technological University, New Delhi, Delhi 110 042, India

Received 28 June 2023; accepted 14 August 2023

We demonstrate that the XMR material ytterbium monoarsenide (YbAs) shows transitions from a trivial to a non-trivial topological phase with hydrostatic pressure of 20 GPa and maintains its topological character up to structural phase transition pressure. We observed band inversions close to the Fermi level at the X high symmetry point at 20 GPa and band parities are used to confirm the same with consideration of Spin-orbit coupling (SOC) effect. The evolution of the surface states and the bulk band structure in YbAs are discussed.

**Keywords:** Ytterbium monoarsenide; Topological phase; Fermi level; Spin-orbit coupling

## 1 Introduction

The  $Z_2$  topological semimetals are a subclass of topological materials. These can be distinguished from trivial insulator via  $Z_2$  topological invariant and requires time-reversal symmetry (TRS) to protect their nontrivial topological characteristics. These topological systems do not exhibit a gap in the bulk band structure *e.g.*, rare earth mononpnictide  $\text{LnPn}$  ( $\text{Ln} = \text{Ce}, \text{Pr}, \text{Sm}, \text{Gd}, \text{Yb}; \text{Pn} = \text{Sb}, \text{Bi}$ )<sup>1</sup>. These systems have shown the  $Z_2$  topological character at ambient pressure. However,  $\text{LaAs}^2$ ,  $\text{LaSb}^3$ ,  $\text{TmSb}^4$ ,  $\text{PbTe}$ ,  $\text{PbS}$ ,  $\text{PbSe}$ ,  $\text{GeTe}^5$  exhibited inversion when external pressure is applied. Similarly, the rare earth mononpnictide family also includes YbAs, which was experimentally reported to be a topologically trivial semimetal under ambient pressure<sup>6</sup> and theoretically demonstrated to show band inversion under applied pressure of 6 GPa<sup>7</sup>, turning it into a  $Z_2$  topological insulator. In this study, we discussed the effect of pressure on topological phase of YbAs by implementing a more accurate hybrid functional with density functional theory (DFT). The invariants, calculated from the parity table of wave functions on high symmetry points, and Wannier Charge Centres (WCCs) along with existence of odd number of gapless topological surface states confirm the topological phase in YbAs.

## 2 Computational Details

Our calculations were based on the projector augmented wave (PAW) approach<sup>8</sup> as implemented

in the VASP code<sup>10</sup>. The screened hybrid functional of HSE06<sup>11</sup> was used to calculate the exchange-correlation potential. The plane wave basis set had a kinetic energy cutoff of 380 eV and  $7 \times 7 \times 7$  k-mesh applied to sample the Brillouin zone (BZ). The maximally localised Wannier functions (MLWFs)<sup>12</sup> were used to construct the TB Hamiltonian and surface band structure. The Wannier charge centers (WCCs) were obtained using the Wannier Tools code<sup>13</sup>.

## 3 Result & Discussions

At ambient pressure, the NaCl-type (space group  $Fm\bar{3}m$ ) structure of YbAs have (0, 0, 0) and (1/2, 1/2, 1/2) position coordinates for As and Yb, respectively, as illustrated in Fig. 1(a). The optimized lattice parameter (5.722 Å), structural phase transition (SPT) and dynamical stability of NaCl-type structure of YbAs with applied hydrostatic pressure is discussed in our previous report<sup>7</sup>. The band structure of YbAs at normal pressure is shown in Fig. 1(c). K-path for band structure includes the Time Reversal Invariant Momenta (TRIM) points in the BZ (*i.e.*, X,  $\Gamma$  and L). A small overlap between the valance and conduction bands around the Fermi level (Fig. 1(b)), in the orbital projected density of states (PDOS), demonstrated that YbAs is semimetallic in nature which is fully agree with previous experimental study<sup>6</sup>. The band structure, Fig. 1(c), indicates that YbAs is a topologically trivial and the *p*-orbital of As dominate in valance band and the *d*-orbital of Yb dominate in

\*Corresponding author: (E-mail: mukhtiyarsingh@dtu.ac.in)



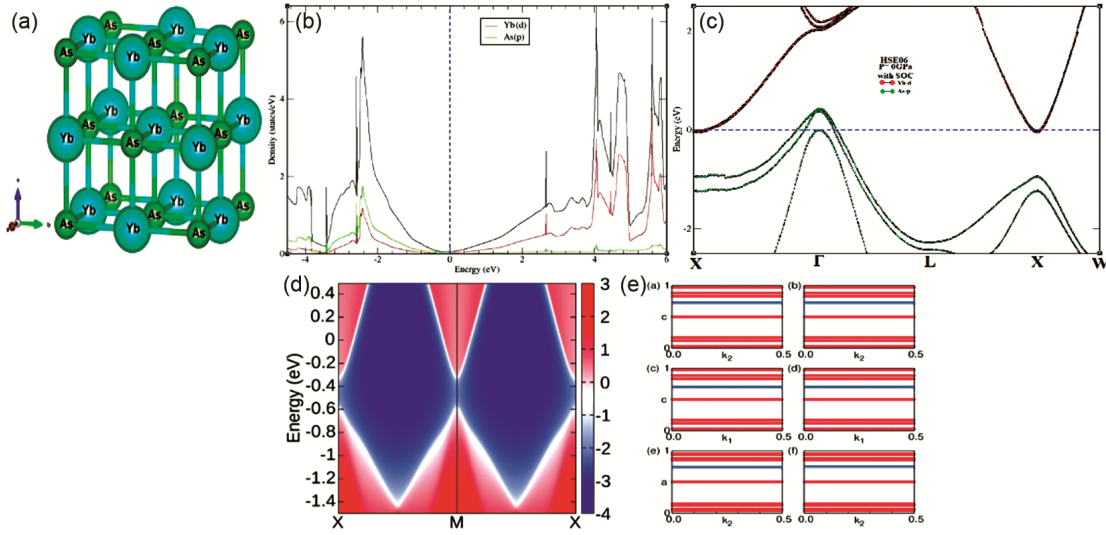


Fig.1 — (a) NaCl-type structure, (b) Orbital projected density of states (PDOS), (c) Orbital projected band structure, (d) Surface density of states (SDOS), (e) wannier center centers (WCCs), of YbAs at ambient pressure ( $P=0$  GPa). The Fermi energy is set at 0 eV.

conduction band near the Fermi level. There is no evidence of a band inversion near Fermi level, at ambient pressure, as it was reported experimentally using angle-resolved photoemission spectroscopy (ARPES) measurement<sup>6</sup>. So, the true nature of YbAs at ambient pressure has been anticipated by our investigation which also confirmed by SDOS and WCCs in Fig. 1(d) & (e). We raised the external hydrostatic pressure inside the SPT limit to investigate the topological quantum phase transition in semimetal YbAs.

External pressure leads to change in lattice parameter, which influences the energy width between bands without changing charge neutrality of the YbAs. We examined the band inversion at each TRIM point to look for the sign of a topological quantum phase transition. Unlike to earlier findings<sup>7</sup>, we observed that the band structure of YbAs does not change adiabatically from 0 to 19.5 GPa and retains its trivial state. As we increased the pressure from 19.5 to 20 GPa, we discovered a band inversion at the X point (shown by the bold arrow in Fig. 2(a)). With applied hydrostatic pressure, the elevation in SOC takes place which results this inversion at X point. The presence of the  $C_{4v}$  double group<sup>3</sup> at the X point in the rock salt structure of YbAs indicates that it contains both time and space inversion symmetries. To further confirm the quantum phase transition, we have examined the parity of the bands close to the Fermi level at the X TRIM point. The band parities at ambient pressure and 20 GPa are listed in Table 1.

Applied hydrostatic pressure of 20 GPa switch the parities at X TRIM point near Fermi level which verifies the band inversion. It appears that YbAs is not a Dirac semimetal due to the opening of the band gap at the X point with inversion (inset of Fig. 2(a)). Table 1 shows that all three X point in BZ have opposite parities, implying that YbAs may be a topological insulator described on curved Fermi surface<sup>3</sup>. However, at 20 GPa, an inverted contribution of the  $d$ -orbital of Yb in valance band and  $p$ -orbital of As in conduction band can be observed at X point near the Fermi level (Fig. 2(a)). Additionally, we derived  $Z_2$  topological invariants as described by Kane and Mele<sup>14</sup> to confirm the non-trivial topological phase of YbAs under applied hydrostatic pressure of 20 GPa. There are eight distinct TRIM points for three-dimensional systems that may be expressed as:

$$G_{i=(m_1 m_2 m_3)} = (m_1 a_1 + m_2 a_2 + m_3 a_3) / 2 \quad \dots (1)$$

where  $m_j = 0, 1$  and the reciprocal primitive lattice vectors are  $a_1, a_2$ , and  $a_3$ . Four  $Z_2$  indices ( $v_0; v_1, v_2, v_3$ ) can be examined with the help of change in the sign of parity to identify adiabatic change in the band structure. The following relationship between parity and  $Z_2$  indices can be used to determine  $v_0$ :

$$(-1)^{v_0} = \prod_{m_j=0,1} \delta m_1 m_2 m_3 \quad \dots (2)$$

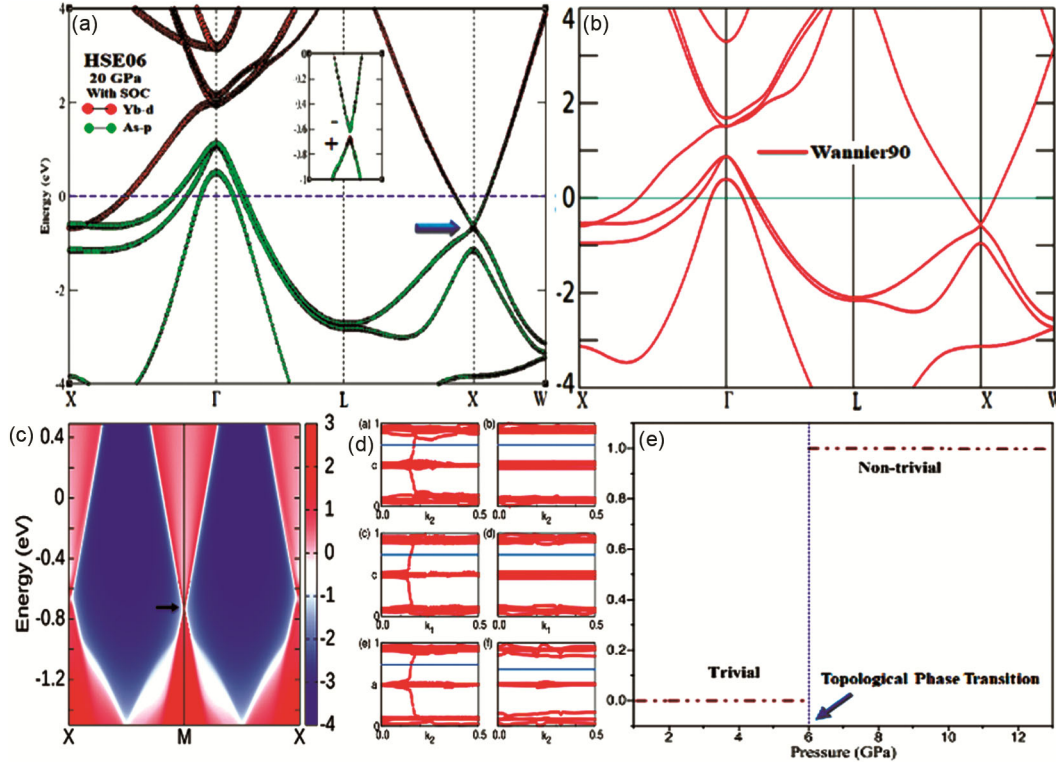


Fig.2 — (a)Orbital projected band structure,(b)wannierized band structure,(c) Surface density of states (SDOS),(d) wannier charge centers (WCCs),(e) first  $Z_2$  topological index,w.r.t. hydrostatic pressure of YbAs at P=20 GPa. The Fermi energy is set at 0 eV.

Table 1 — Parity table of YbAs at 0 GPa and 20 GPa.

TRIMs Pressure	4L	$\Gamma$	3X	$\prod \delta_m$	$Z_2$
0 GPa	-	+	+	+	0
20 GPa	-	+	-	-	1

where  $\delta_i$  signifies the parities of all the filled bands at all TRIM points. YbAs is topologically trivial since there is no band inversion at ambient pressure and which can be verified with  $v_0$  value 0 calculated from at the X point when we apply 20 GPa pressure and it turns  $v_0$  into 1, indicating a topologically non-trivial phase. Fig. 2(e) shows the change in the first  $Z_2$  topological index as a function of pressure. We computed the (001) surface band structure to further illustrate the topologically nontrivial nature of YbAs under hydrostatic pressure. YbAs exhibits bulk band inversion at three X TRIM points, showing three Dirac cones on the surface states associated with the inversions. These Dirac cones related to band inversion are projected onto the surface Brillouin zone. These three X TRIM points are projected on M point in surface Brillouin zone (SBZ). In Fig. 2(c), surface Dirac cone is shown along X-M-X k-path. To examine the topological phase and  $Z_2$  topological invariants of YbAs, we calculate WCCs on six  $k_i = 0, \pi$  ( $i = x, y, z$ ) TRIM

planes. The behaviour of WCCs for six in variant planes along given k-path is shown in Fig. 1(e) & 2(d).

Fig. 2(d) shows odd number of crossings of WCCs with horizontal reference line on  $k_i = 0$  planes in half Brillouin zone (BZ) which verifies the non-trivial topological phase with the invariant  $Z_2 = 1$ . Even number of crossings of WCCs with horizontal reference line in Fig. 1(e) with  $k_i = 0, \pi$  planes and in Fig. 2(d) with  $k_i = \pi$  planes show a trivial phase with the invariant  $Z_2 = 0$ . Four distinct 3D topological invariants ( $v_0; v_1v_2v_3$ ) = (1;000) have been identified with the help of the parity table (Table 1) and WCCs (Fig. 2(b) & Fig. 4(b)), which confirms the non-trivial phase of YbAs protected by TRS at a hydrostatic pressure of 20 GPa.

#### 4 Conclusion

We have discovered the non-trivial topological phase of experimentally synthesized YbAs under applied hydrostatic pressure. Our First-principles calculations have shown that YbAs is topologically trivial at ambient pressure with SOC. The strength of SOC in system enhanced due to applied hydrostatic pressure resulting in a topological phase change from trivial to a non-trivial at 20 GPa. We have investigated the (001) surface states where the



characteristic crossing can be observed. We have found that YbAs hosts three Dirac cones along three X TRIM points which are projected to M point in surface Brillouin zone. Band inversion has been confirmed by an exchange of orbital contribution in bands close to the Fermi level under pressure. The calculation of  $Z_2$  indices or  $Z_2$  topological invariants has established the topological phase transition. Product of parities and WCCs shows the change from zero to non-zero in first  $Z_2$  topological invariant at 20 GPa pressure. This has shown that over the mentioned pressure value, a topologically non-trivial state of YbAs can be achieved.

### Acknowledgement

One of the authors (Ramesh Kumar) would like to thank Council of Scientific and Industrial Research (CSIR), Delhi, for financial support. The authors also acknowledge the National Supercomputing Mission (NSM) for providing computing resources of 'PARAM SEVA' at IIT, Hyderabad, which is implemented by C-DAC and supported by the Ministry of Electronics and Information Technology (MeitY) and Department of Science and Technology (DST), Government of India.

### References

- 1 Duan X, Wu F, Chen J, Zhang P, Liu Y, Yuan H & Cao C, *Commun Phys*, 1 (2018) 71.
- 2 Khalid S, Sabino F P & Janotti A, *Phys Rev B*, 98 (2018) 220102(R).
- 3 Guo P J, Yang H C, Liu K & Lu Z Y, *Phys Rev B*, 96 (2017) 081112(R).
- 4 Wadhwa P, Kumar S, Shukla A & Kumar R, *J Phys: Condens Matter*, 31 (2019) 335401.
- 5 Barone P, Rauch T, Sante D D, Henk J, Mertig I & Picozzi S, *Phys Rev B*, 88 (2013) 045207.
- 6 Xie W, Wu Y, Du F, Wang A, Su H, Chen Y, Nie Z Y, Mo S K, Smidman M, Cao C, Liu Y, Takabatake T, Yuan H Q, *Phys Rev B*, 101 (2020) 085132.
- 7 Singh M, Kumar R & Bibiyan R K, *Eur Phys J Plus*, 137 (2022) 633.
- 8 Kresse G, Joubert D, *Phys Rev B*, 59 (1999) 1758.
- 9 Kohn W & Sham L J, *Phys Rev*, 140 (1965) 1133.
- 10 Kresse G & Furthmüller J, *Phys Rev B*, 54 (1996) 11169.
- 11 Heyd J, Scuseria G E & Ernzerhof M, *J Chem Phys*, 118 (2003) 8207.
- 12 Mostofi A A, Yates J R, Pizzi G, Lee Y S, Souza I, Vanderbilt D & Marzari N, *Comput Phys Commun*, 185 (2014) 2309.
- 13 Wu Q, Zhang S, Song H F, Troyer M & Soluyanov A A, *Comput Phys Commun*, 224 (2018) 405.
- 14 Fu L, Kane C L & Mele E J, *Phys Rev Lett*, 98 (2007) 106803.



# Rainfall Assessment and Water Harvesting Potential in an Urban area for Artificial Groundwater Recharge with Land Use and Land Cover Approach

Ali Reza Noori<sup>1,2</sup> · S.K. Singh<sup>1</sup>

Received: 13 March 2023 / Accepted: 30 August 2023  
© The Author(s), under exclusive licence to Springer Nature B.V. 2023

## Abstract

Cities in arid and semiarid regions face the dual challenges of managing urban floods and water shortages, threatening their sustainability. Urban areas are particularly vulnerable to flooding despite minimal rainfall and are prone to drought. This is evident in the capital of Afghanistan, Kabul, where groundwater decline and urban floods pose severe challenges. This study investigates the possibility of utilizing rainwater harvesting (RWH) to manage urban floods and recharge groundwater. The research examines various aspects of rainfall patterns, such as variability, rainy days, seasonality, probability, and maximum daily precipitation. The analysis of precipitation statistics reveals that rainfall exceeding 30 mm occurs approximately every 3–4 years. Rainfall in Kabul follows a seasonal pattern, with a coefficient of variation of 127% in October and 46% in February during the wet period. The study then assesses the potential of RWH in Kabul City as a solution for stormwater management and groundwater recharge. Based on the typology of land use and land cover, implementing a rainwater harvesting and recharge system (RWHRs) could increase mean annual infiltration from 4.86 million cubic meters (MCM) to 11.33 MCM. A weighted Curve Number (CN) of 90.5% indicates impervious surfaces' dominance. The study identifies a rainfall threshold of 5.3 mm for runoff generation. Two approaches for collecting rainwater for groundwater recharge are considered: RWHRs for a residential house with an area of 300m<sup>2</sup>, which yields approximately 88m<sup>3</sup>/year, and RWHRs for a street sidewalk to collect water from streets and sidewalks. These findings highlight the potential of RWHRs as an effective strategy for managing urban floods and recharging groundwater artificially.

**Keywords** Kabul · Groundwater · Rainwater Harvesting · Groundwater Recharge

---

✉ Ali Reza Noori  
a.noori@kpu.edu.af  
S.K. Singh  
sksinghdce@gmail.com

<sup>1</sup> Department of Environmental Engineering, Delhi Technological University, Delhi, India

<sup>2</sup> Department of Water Supply and Environmental Engineering, Faculty of Water Resources and Environmental Engineering, Kabul Polytechnic University, Kabul, Afghanistan

# 1 Introduction

Water, vital for life, faces major challenges in quantity and quality. These challenges are due to climate change, population growth, and urban and industrial expansion in cities worldwide. In arid and semiarid regions, people rely mainly on groundwater as their primary water source (Noori and Singh 2021a). Unfortunately, excessive usage and continuous extraction often lead to the depletion and scarcity of this valuable resource. Approximately 2.5 billion people rely solely on groundwater for their daily needs, accounting for 50% of the global drinking water supply. The growing dependence on these resources has resulted in a global dilemma regarding access to clean groundwater (Sarma and Singh 2021). Roughly 26% of the world's renewable freshwater supplies come from.

Kabul, a city with a population of over four million, relies heavily on groundwater resources to meet its water needs. However, a recent trend analysis conducted by (Noori and Singh 2021b) suggests that the water levels in the aquifers of the Kabul Plain have significantly decreased. The study highlighted that certain observational wells in Kabul experienced a decline rate of more than 2 m per year. Findings of previous studies indicate that the quantity and quality of Kabul's groundwater are seriously threatened (Mack et al. 2009, 2013; JICA 2011; Taher et al. 2013; Zaryab et al. 2017; Brati et al. 2019; DACAAR 2019; Jawadi et al. 2020, 2022). Groundwater challenges in the Kabul Plain stem from population growth, excessive water consumption, urbanization, soil impermeability, and climate change. A pilot project called the Kabul Managed Aquifer Recharge Project (KMARP) was implemented in Kabul, specifically at four designated sites (Noori and Singh 2021a). This project aimed to evaluate the feasibility and effectiveness of managing groundwater recharge (MAR) as a potential solution to combat water scarcity in the city.

Rainwater harvesting is a traditional technique used to tackle water scarcity in arid and semiarid regions. Ancient civilizations have employed it to meet their drinking and agricultural water needs (Mahmoud et al. 2014).

Rainwater harvesting has become popular for increasing surface and groundwater supplies to meet water resource needs. Numerous research studies have recently been conducted regarding rainwater collection techniques and their applications (Dhakate et al. 2013; Gwenzi and Nyamadzawo 2014; Jung et al. 2015; Sayl et al. 2016; Tiwari et al. 2018; Wu et al. 2018; Sadeghi et al. 2019; Matomela et al. 2020). In modern times, rainwater harvesting is recognized as a sustainable adaptation strategy in urban areas, helping to mitigate water scarcity and manage flooding problems (Gado and El-Agha 2020; Zabidi et al. 2020; Krishna et al. 2021; Ranaee et al. 2021). In urban areas, hydrological issues arise from changes in surface runoff and river flow, reduced infiltration and groundwater recharge, and the presence of impermeable surfaces (Nachshon et al. 2016). Traditional rainwater harvesting involves collecting and using rainwater for various purposes. However, a new approach to collecting rainwater to replenish underground aquifers has gained attention (Ghazavi et al. 2018; Hussain et al. 2019; Qi et al. 2019; B R and Lokeshwari 2021; Huang et al. 2021).

Inadequate management of rainfall and surface runoff can result in severe urban flooding. These floods can have numerous detrimental effects, including disruptions to daily life, damage to infrastructure, erosion of riverbanks and riverbeds, contamination of water resources, and loss of life. Urban floods also have significant economic and environmental consequences, impacting traffic systems, water and electricity supply, and telephone lines and causing socio-cultural disturbances. Urban flooding poses a significant problem

in Kabul City. According to a study by (Manawi et al. 2020) there have been substantial changes in land use and land cover pattern in North Kabul from 1964 to 2009. The study reveals a 15% reduction in green areas, a 27% decrease in bare soil, and a 51% increase in impervious surfaces. The primary factors contributing to urban floods in Kabul are unsustainable urbanization, inadequate drainage systems, and significant alterations in land cover.

Implementing a water harvesting strategy can help mitigate water loss from surface runoff and enhance water resource availability in various settings, including watershed systems, metropolitan areas, and regions with unequal water distribution. As urban populations grow and develop, land surfaces become less porous, leading to floods even with minimal rainfall. Rainwater harvesting (RWH) can serve as a supplementary water source in arid and semi-arid urban areas, helping to address water scarcity challenges. The current study will concentrate on rainwater harvesting strategies incorporating infiltration of the captured water to groundwater, referred to as a “rainwater harvesting and recharge system” (RWHRs). This study aimed to exhibit the precipitation variability and concentrations and look for appropriate management strategies for urban flooding and groundwater recharge using RWHRs approaches in the Kabul basin. The study’s discoveries provide additional understanding of the rainfall patterns in Kabul city and the use of rainwater harvesting strategy, particularly the RWHRs, to tackle water scarcity and control urban flooding in urban areas. The RWHRs methods can be adapted to support sustainable water resources management in arid and semiarid areas to overcome water scarcity, MAR, and urban flooding problems.

## 2 Materials and Methods

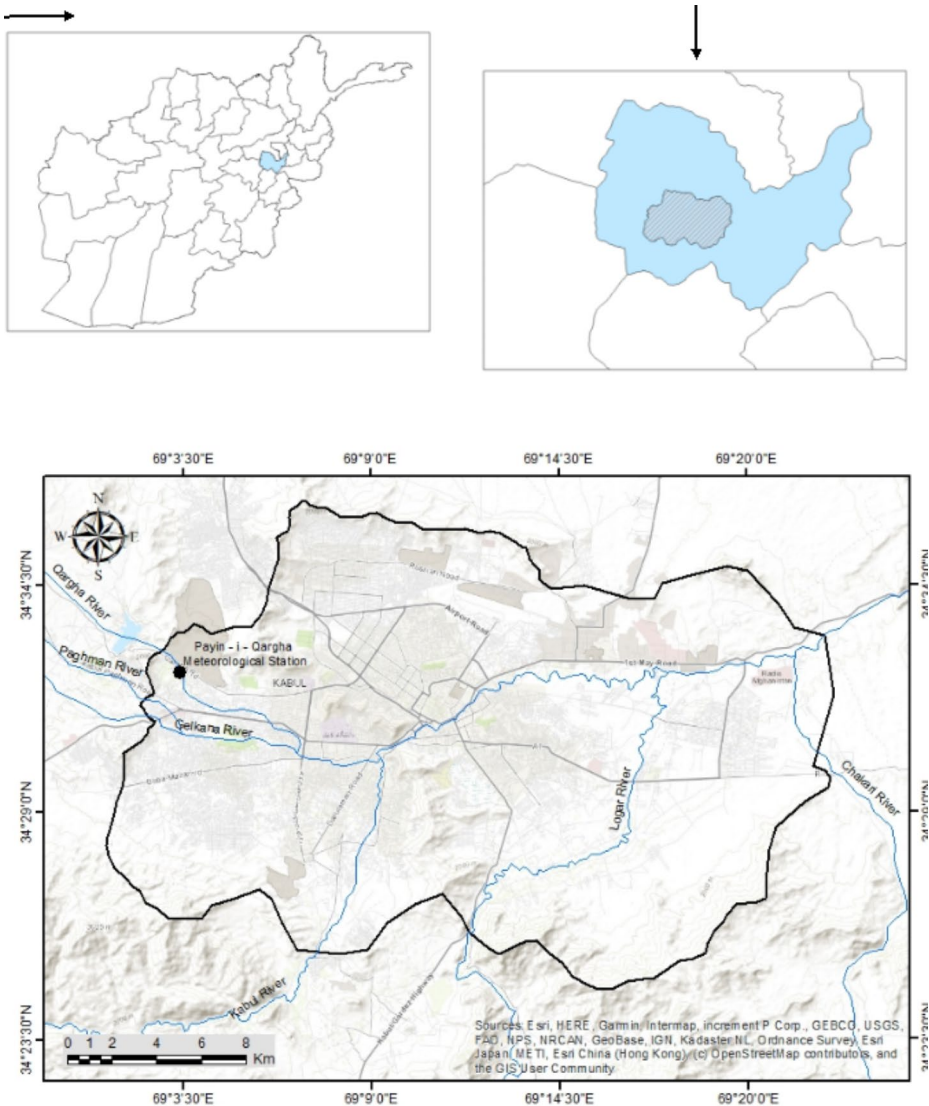
### 2.1 Study Area

The study area is the Kabul Basin (JICA 2011), which encompasses most of Afghanistan’s capital city of Kabul. The study area is situated between 34°36’30” and 34°24’40” N latitude and 69°01’25” and 69°22’30” E longitude in the country’s central-east (Fig. 1). The basin has a total area of 487 km<sup>2</sup>. The research region has arid to semiarid climatic conditions. The basin’s highest and lowest average temperature ranges were 32°C in July and −7°C in January (Zaryab et al. 2017). Built-up is the most prevalent land use type in the study area. It is the country’s leading national commercial base, and many refugees returned to their homeland after 2001, mostly settling in Kabul city.

Additionally, many neighboring provinces’ residents migrated to Kabul for employment. Mountain ranges that are low yet relatively steep surround the basin. The elevation of the basin ranges from 1763 to 2823 m. Generally, the mountains in the southern and southwestern boundaries are higher than the others.

Three rivers are entering the city of Kabul. Paghman River spills out from east to west. The Maidan River (Kabul River) arrives at the study area from the south and flows 21 km before joining the Paghman River. The Logar River, an enormous tributary of the Kabul River, flows south-north and joins the Kabul River around 17 km downstream of the Paghman waterways mouth.

The common surficial geological forms confirmed in the Kabul basin are conglomerate and sandstone, loess, metamorphic rocks, limestone, fan alluvium and colluvium, gneiss, limestone and dolomite, sandstone and siltstone, ultramafic intrusions and river channel



**Fig. 1** Location map of the study area

alluvium. A collection of terrestrial and lacustrine deposits, primarily uncemented and semi-consolidated lacustrine, fluvial, and aeolian sedimentary rocks, including sand, gravel, and silt from the Quaternary and Neogene eras, is present across the Kabul Basin (Tünnermeier and Houben 2005; Mack et al. 2009; Zaryab et al. 2017).

According to (JICA 2011), there are three main aquifer clusters in the study region: The shallow aquifer is located within the alluvial deposits and a deeper aquifer in Neogene layers (“the upper Neogene aquifer”), and a deep aquifer (“the lower Neogene Aquifer”). According to some available literature, the city encompasses four major interconnected aquifers (Uhl and Tahiri 2003; Pell Frischmann 2012; Zaryab et al. 2017). Meanwhile, shal-

low aquifers, also known as “Alluvial Aquifers” or “Quaternary Aquifers,” may be found across the city at different depths, albeit the deposits that store such groundwater are denser and have a larger groundwater potential near river systems.

## 2.2 Data Acquisition

Precipitation figures of six meteorological stations in Kabul province were obtained from the Department of Meteorology, General Directorate of Water Resources, Ministry of Energy and Water of Afghanistan for 2008–2020. The gathered data were in the form of daily logs from 2008 to 2020, which were then transformed into monthly and yearly totals for study. Only one meteorological station (Payin-i-Qargha), located within the research region, is considered out of the six stations for the current study. Since the data for 2008 and 2020 were incomplete, they were omitted from the analysis, and it was intended to consider the records from 2009 to 2019 only. Landsat imagery was downloaded to create land use and land cover (LULC) identity maps for 2020. The required satellite data were obtained from USGS Portal (<https://earthexplorer.usgs.gov/>) using the address “path 153 and rows 36.”

## 2.3 LULC Development

The availability and sustainability of groundwater are impacted by several variables, including land use and land cover (Machiwal et al. 2011; Martin et al. 2017). Expanding impermeable land surfaces, such as asphalt, concrete roads, streets, and waterproof roof materials, would hinder groundwater recharge. To create the LULC map of the study area, remotely sensed Landsat-8 satellite data was analyzed. “Supervised classification approach with maximum likelihood algorithm” in ENVI 5.3. was applied to create the LULC of the study area (Fig. 2). The research area has bare land and rock, cropland and vegetation, settlements (built-up area), water bodies, and marshland. Built-up areas reduce the effect of groundwater recharge, whereas vegetation-covered regions provide better prospects for groundwater recharges.

## 2.4 Rainfall Analysis

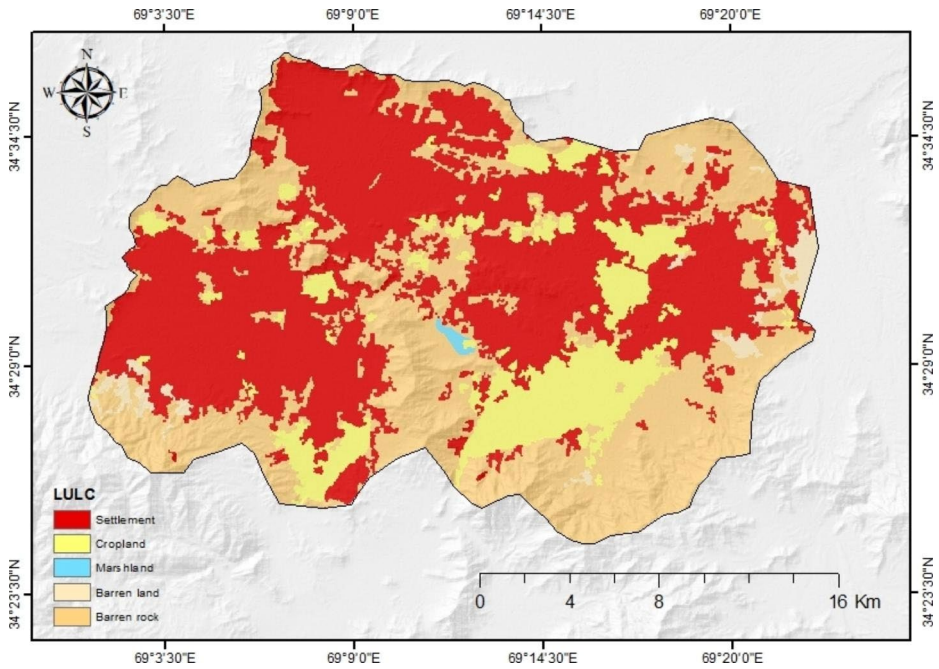
In this study, many precipitation characteristics have been examined: variability, the number of rainy days, their distribution across the season, the likelihood of daily precipitation, and the highest amount of rainfall ever recorded in a day.

The precipitation concentration index (PCI) and seasonal index (SI) were introduced by (Oliver 1980; Walsh and Lawler 1981), and the contribution index (CI) (Mahmoud et al. 2014) were utilized to identify the irregularity and seasonal distribution of precipitation over the year.

Equation 1 was used to calculate the PCI using RStudio, while Eq. 2 was used to get the SI.

$$PCI = 100 \times \sum_{n=1}^{12} \left( \frac{X_n^2}{R^2} \right) \quad (1)$$





**Fig. 2** LULC map of the study area

$$SI = \frac{1}{R} \sum_{n=1}^{12} \left| \frac{X_n - R}{12} \right| \quad (2)$$

R displays the yearly precipitation, and  $X_n$  indicates the rain in month  $n$ . The daily, monthly, or seasonal rainfall is calculated as a percentage of the annual total precipitation for the contribution index. By employing the Weibull method (Mansell 2003), daily precipitation data collected from Payin-i-Qargha meteorological station from 2009 to 2019 were statistically evaluated:

$$P = \frac{m}{N+1} \times 100 \quad (3)$$

Where  $P$  shows the probability of the precipitation (%),  $N$  is the data size, and  $m$  is the rank given to the data when sorted in descending order.

The probability of maximum daily precipitation and its return period were analyzed employing the Gumbel distribution approach. The method is the limiting form of a large number of uniformly sized samples with an exponential starting distribution. The cumulative distribution is used to calculate the likelihood (percent) that a rainfall depth  $X$  (mm) will be greater than a specified rainfall depth  $x_0$  (mm). (Mahmoud et al. 2014):

$$P(X \geq x_0) = 1 - e^{-e^{-y}} \quad (4)$$

Where  $y$  is a dimensionless variable and calculated as follows:

$$y = \frac{1.286(x - \bar{x})}{\sigma_x} + 0.577 \text{ where } x \text{ is variate} = \frac{\sigma_x(y_T - 0.577)}{1.2825} + \bar{x} \quad (5)$$

$\bar{x}$  is the mean value and  $\sigma_x$  is the standard deviation of variate  $x$ , as well as  $y_T$  is the reduced variate for a given  $T$

$$y_T = - \left[ \ln \ln \frac{T}{T-1} \right] \quad (6)$$

By determining the reduced mean  $y_n$  and reduced standard deviation  $S_n$  using tables based on the sample size, the frequency factor  $K$  can be calculated as follow:

$$K = \frac{(y_t - \bar{y}_n)}{S_n} \quad (7)$$

The following equation may be used to compute  $X_t$  by providing the value of  $K$ .

$$x_t = \bar{x} + K\sigma_{n-1} \quad (8)$$

## 2.5 Rainwater Harvesting Potential Based on LULC

Land use is a fundamental factor determining the likelihood of surface runoff, which directly affects how quickly rain and other precipitation permeates the ground. More porous surfaces become impenetrable as cities expand. Different sources have reported the permeability coefficients of various materials. (Nachshon et al. 2016), quoted from (Pauleit and Duhme 2000), illustrated the infiltration coefficients for the built-up and asphalt areas at 5%, pavement areas at 20%, woody and vegetation area at 25%, meadow, and pastures at 35%, arable land 40%, and bare soil 50%. Also, according to the argument by (Nachshon et al. 2016), the permeability coefficient for the built-up areas with (RWHRs) increases up to 80% and the remaining 20% they consider as the reason for evapotranspiration. As aforementioned, contrary to conventional rainwater harvesting (RWH) systems, which store the water on the land for individual use by the property owners, some researchers recently addressed the potential of rainwater for recharging the nearby aquifers through infiltration wells.

The site's hydrological, climatic, and surface area characteristics must be considered while building the infiltration well structure tools for the rainfall. Considering the hydraulic conductivity of the medium at the specific site of infiltration, it is necessary to create a deep enough infiltration well with a long filter length to allow sufficient water flow from the well into the ground to ensure the effective infiltration of collected rainwater into the aquifer without flooding the infiltration well system.

## 2.6 Groundwater Recharge

Estimating the rainwater infiltrating the underground water and the vadose zone depends on the target area's infiltration coefficients. And the infiltration coefficient is related to various factors such as soil hydraulic characteristics, topography, land surface coverage, etc. (Nach-

shon et al. 2016). For the different types of substrates mentioned in Table 1, (Nachshon et al. 2016) used an illustration from (Pauleit and Duhme 2000) to show the infiltration coefficients ( $I_c$ ) values in percent, which indicate the portion of yearly rainfall that infiltrated underground.

Here, it is projected that 80% of the water is seeping into groundwater for RWHRs that route the gathered water from the collecting sites straight into the subsurface, either into the vadose zone or the aquifer. In other words, instead of the 5% shown in Table 1 for non-RWHRs situations, the  $I_c$  of constructed areas where RWHRs is applied is 80%. This cautious estimate permits a 20% water loss due to evaporation and retention along the RWHRs system. This percentage is most likely significantly lower than 20%.

By using the weighted arithmetic mean of various infiltration coefficients from regions with different land cover qualities ( $I_{c(i)}$ ) and taking into account the associated surface areas ( $A(i)$ ) of each LULC (e.g., built-up, bare soil, cropland, etc.), the effective infiltration coefficient ( $I_{c(\text{eff})}$ ) can be calculated.

$$I_{c(\text{eff})} = \frac{\sum (A_i \times I_{c(i)})}{\sum A_i} \quad (9)$$

The exact amount of infiltrated water into groundwater  $I$  ( $\text{m}^3$ ) is determined as:

$$I = A \cdot R \cdot I_{c(\text{eff})} \quad (10)$$

Where  $A$  is the surface area ( $\text{m}^2$ ) through which infiltration is occurring, and  $R$  is the yearly rainfall ( $\text{m}$ ). By assumption of three main land cover components in urban contexts (i.e., built-up, bare soil, and arable land), it is straightforward to estimate  $I_{c(\text{eff})}$  for any given area with any combination of the three components by knowing the  $I_c$  values for each of these constituents. According to Table 1, the infiltration coefficients ( $I_c$ ) of built-up areas without RWHRs are equivalent to 5% and 80% for regions developed with RWHRs, 50% for bare soil, and 40% for arable land.

## 2.7 Surface Runoff

Due to drainage systems failing during extreme rain, extreme flooding events occur more frequently in urban contexts due to the combined effects of global climate change and the impervious nature of modern cities. Flooding threatens structures, additional public and private infrastructure, and people's lives. Surface runoff must be decreased in urban settings to lessen the risk of flooding and the cost of drainage systems (Nachshon et al. 2016). RWHRs

**Table 1** Infiltration coefficient of different surfaces (Nachshon et al. 2016)

Land cover	$I_c$ (%)
Built up (without RWHRs)	5
Asphalt	2
Pavement	20
Woody Vegetation	25
Meadow and pastures	35
Arable lands	40
Bare soil	50
Built up (with RWHRs)	80

may be beneficial since it increases the quantity of water that permeates the subsurface instead of flowing as surface runoff.

A variety of circumstances influence runoff and rainfall relationships. Some refer to meteorological properties such as precipitation intensity, duration, and evapotranspiration. In contrast, others relate to physical factors of the surfaces receiving the precipitations, like their permeabilities and slopes. These elements function in how much of the rainfall depth is absorbed by the atmosphere, the surface of the earth, or both. Runoff coefficients were established to calculate the potential runoff from a given depth of rainfall. These coefficients show how much of the precipitation depth should be subtracted and accounted for as a loss to runoff.

The “Natural Resources Conservation Services (NRCS)” equation for rainfall runoff has been employed to determine the possible runoff from a rainstorm. The equation, formerly known as the “Soil Conservation Service (SCS)” approach for estimating direct runoff from rainstorms, was created by the “United States Department of Agriculture (USDA) in 1972” (Mahmoud et al. 2014). It is seen as either a probabilistic or deterministic model.

Since the only relevant rainfall data for this study were daily rainfall time series, applying this approach is ideal for the study region because it eliminates rainfall intensity and removes time as a component. The correlation between the land cover, the “Hydrologic Soil Group (HSG),” and the “Curve Number (CN)” is included in the model. A soil class with high CN values is impermeable and will have more runoff than infiltration:

$$Q = \frac{(P - I_a)^2}{P - I_a + S} = \frac{(P - 0.2S)^2}{P + 0.8S} \quad (11)$$

Q represents the amount of daily runoff in mm, P represents the amount of daily precipitation in mm, S represents the region’s potential maximum storage (mm), and  $I_a$  represents the initial abstraction (usually taken as 0.2 S) in mm. The following equation indicates how much rainfall is directed to surface runoff by using CN, the runoff curve number of a hydrologic soil group and land cover combinations:

$$S = \frac{25,400}{CN} - 254 \quad (12)$$

According to (USDA 2009), the average CN values for impervious surfaces (built-up area) are about 98. For cropland, it is taken around 76; for bare land, it is assumed to be about 86. The following equation gives the weighted calculated curve number (CN<sub>w</sub>) considering different land-use classifications of the study area.

$$CN_w = \frac{\sum CN_{wi}}{100} \quad (13)$$

Where CN<sub>wi</sub> stands for the weighted curve number of the specific land cover.

## 2.8 Models for Rainwater Harvesting and Groundwater Recharge

Two models have been designed to use precipitation to replenish groundwater, avoid waste in terms of surface flows, and stop urban floods. The first model is used to collect and direct rainwater from residential houses to feed underground water, and the second model is employed to manage and control rainwater from the surfaces of roads and streets for groundwater recharge. The first case considers a typical residential home, where rainfall is collected and channeled to the groundwater recharge well (absorbing well) from the roof and the yard. The rainwater collecting channels collect the water from the roof and the yard and send it to the grease and oil trap basin.

Grease and oil traps allow water accumulation to be separated from grease and oil residues. It is constructed from a tank with a baffle wall in the middle. Water enters the basin from one side; solid particles sink to the bottom, while grease and oil float to the top. The clear water enters the basin's other side from under the baffle wall and exits to the sand filter.

Sand filters refine the water by passing it through fine sand to eliminate the tiniest contaminants. It comprises of a basin with a fine-sand layer and a gravel layer. The water enters the basin from above and passes through both of these layers to be purified. In order to prevent pore clogging, this filter also has to be backwashed sometimes. For backwash water that can have its overflow linked to municipal rainfall channels, a backwash water drying basin is also taken into consideration. The filtered water then enters in the recharge well which typically equipped with a casing, screen, gravel pack, and gravel bed.

Similarly, the second model is considered for collecting rainwater from roads and streets to direct it to groundwater recharge wells. The system of groundwater recharge wells can be constructed at a distance on the sides of the streets (pedestrian area). The rainwater from the road surface and sidewalk is collected through the closed channel on the side of the road equipped with screens and enters the sedimentation basin through specific chambers. The settling tank is separated into two separate parts by a buffer wall. The first part is the grease and oil separator, and the second is the settling tank. The settling tank is connected to a recharge well similar to recharge wells of residential houses based on construction. The settling tanks can be cleaned out regularly during the year.

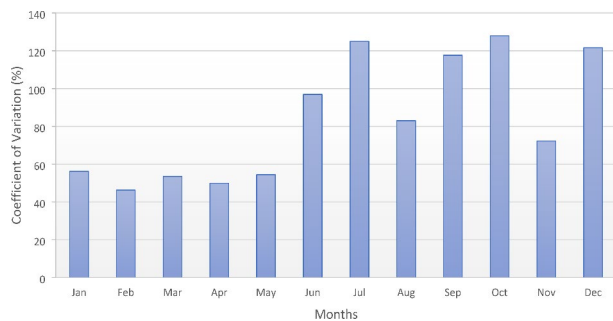
## 3 Result and Discussions

### 3.1 Rainfall Conditions

The rainfall in Kabul varies over the months, as seen in Fig. 3. The last six months of the year are the wettest according to the coefficient of variation, and the coefficient of variation for the rainfall during these months reached as high as 127% (October). In contrast, it is decreased during the first six months of the year and the coefficient of variation has a minimum value of 46% (February).

According to many indexes, Kabul has inconsistent, seasonal, and intense precipitation (Fig. 4). The most rainfall occurred during February, March, and April. In 2019, the first five months (January through May) had more than 83% yearly rainfall. In 2019, the two wettest months (February and April) alone saw up to 45% of the annual precipitation.

**Fig. 3** Variability of precipitation during normal times as determined by the coefficient of variation



Given the region's yearly rainy days, concentration feature, and low overall quantity of precipitation, understanding how much rain falls over a day is crucial. For the time series depicted in Fig. 5, the mean annual rainfall is 368 mm, with an average rainy day of about 80 days each year.

According to Figs. 5 and 2019 had the most precipitation, totaling 486.21 mm. The minimum amount of rainfall happened in 2018, with a total record of 269.4 mm. The maximum daily rain has recorded on 17 March 2014 with a rainfall amount of 80.77 mm, which caused floods with an inundation level of 60 to 80 cm (Manawi 2020). According to the maximum rainfall record of the meteorological station, the minimum (from the daily ultimate precipitation record) precipitation was observed on 28 January 2010 with a precipitation amount of 18.04 mm. According to observations of rainy days, the research region experienced a minimum of 30 rainy days in 2017 and a maximum of 102 rainy days in 2019.

The Weibull approach's probability analysis for 2009 to 2019 (Fig. 6) shows a return period of 3 to 4 years for daily rainfall of less than 30 mm. Therefore, the previously described urban issues, such as street floods, would happen every 3–4 years. The quantity of 80.77 mm, which has a return cycle of ten years, was the largest amount of rainfall recorded in 2014 thus far.

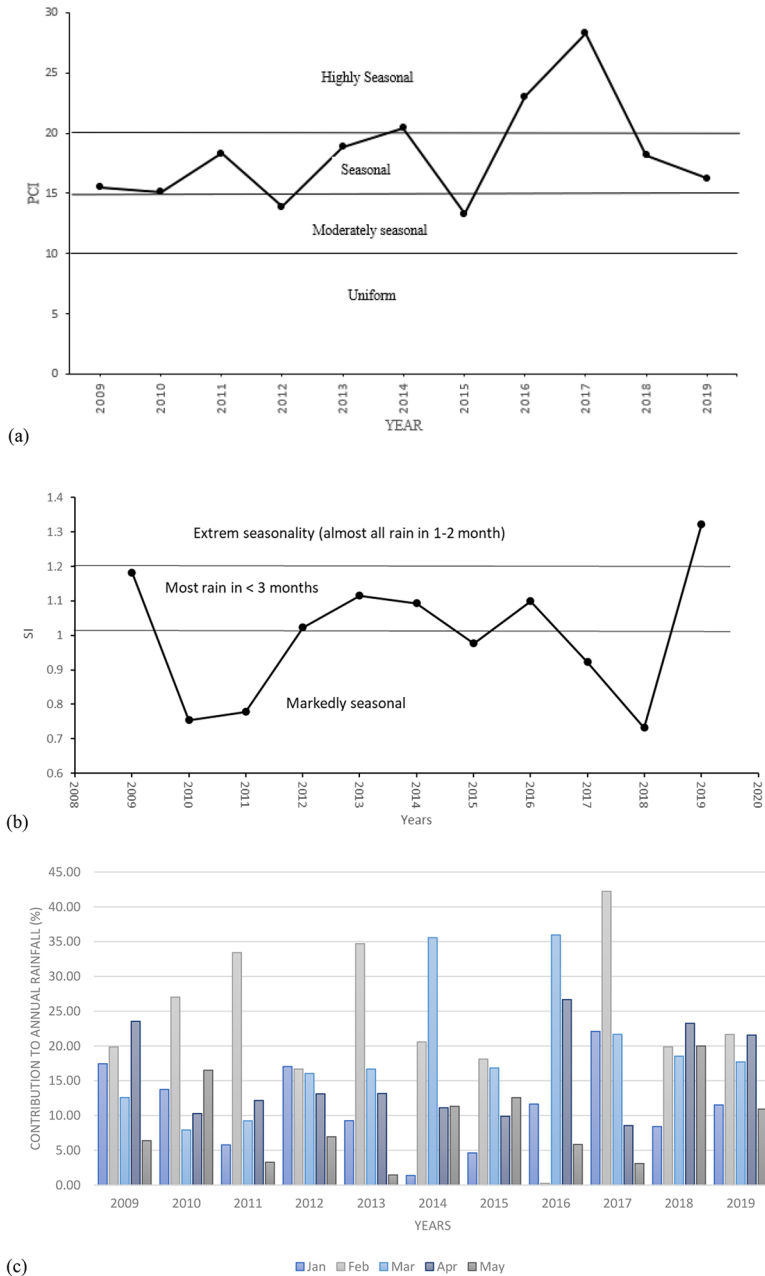
The statistical analysis used the Gumbel technique for rainfall extremes and produced mean and standard deviation values of 36.87 mm and 18.23 mm, respectively. Consequently, the constants  $y_n$  and  $S_n$  Eq. (7) had respective values of 0.4996 and 0.9676. Figure 7 depicts the probability curve for the highest daily precipitation.

Using Eq. (6), a return period of 1 to 5 years is calculated for the threshold rainfall depth of 55 mm, which is expected to enhance the danger of flooding in the city. For the 122 mm record, a return time of about 150 years has also been discovered. The findings described above demonstrate a high likelihood that the problematic runoff from the 55 mm precipitation will reoccur. As a result, there is a very high likelihood that RWH will occur in the urban region of the Kabul basin.

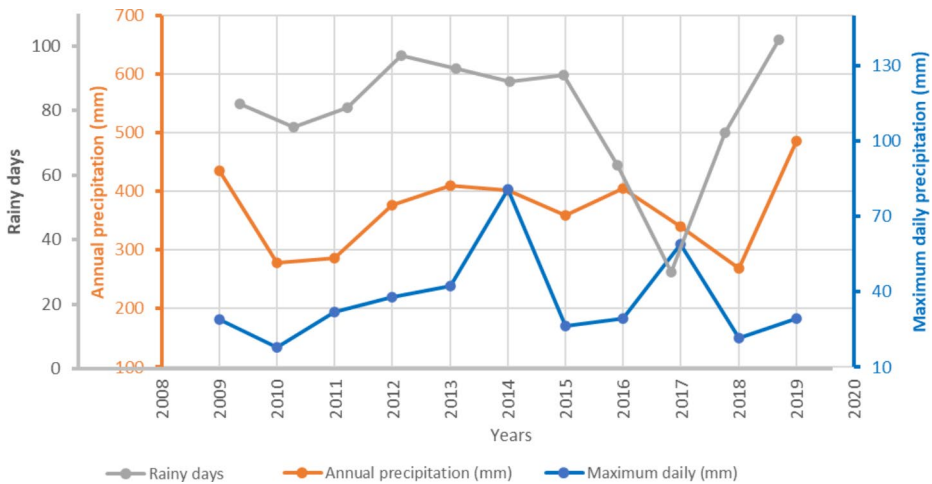
### 3.2 RWHRs vs. non-RWHRs Conditions

Based on a developed LULC map, the  $I_{c(\text{eff})}$  was estimated for the research region under the circumstances with and without RWHRs. As stated previously, it is expected that the primary land covers of the urban environment are built up having  $I_c$  of 5% without RWHRs, and 80% with RWHRs, arable land with  $I_c$  of 40%, and barren soil with  $I_c$  of 50% has been applied. After segregation of LULC of the study area, built-up, bare land, and arable land,

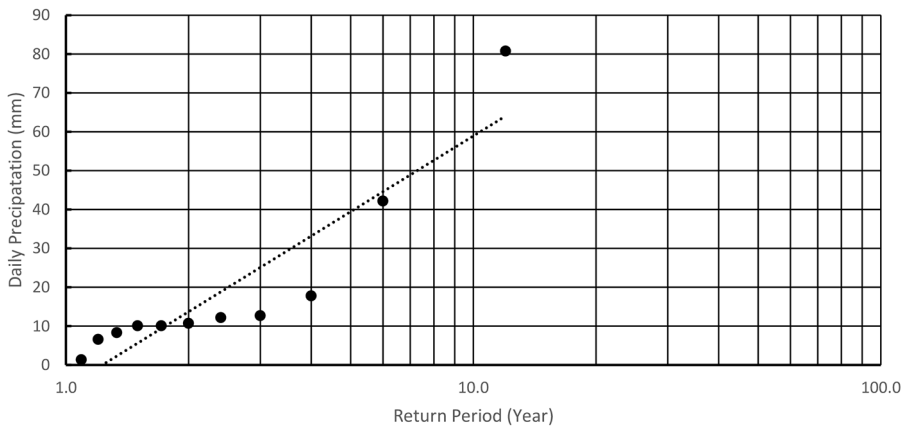




**Fig. 4** The seasonality and concentration of precipitation (a) precipitation concentration index, (b) seasonality index, and (c) average percent contribution to annual rainfall



**Fig. 5** Time series of annual precipitation, daily maximum precipitation, and yearly rainy days

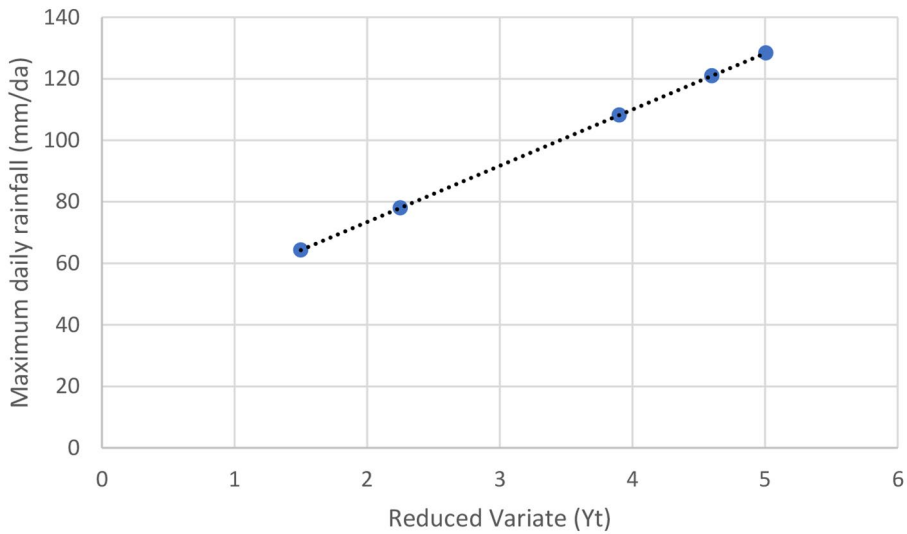


**Fig. 6** Daily precipitation probability curve from Jan 2009 to Dec 2019

as well as their  $I_c$ , it was possible to compute the total  $I_{c(eff)}$  of the entire region employing (Eq. 9).

The spatial proportion of each LULC is illustrated in Table 2, including the computed  $I_{c(eff)}$  and groundwater infiltration rates for RWHRs and non-RWHRs situations with  $R=368$  mm (average yearly rainfall at the region from 2009 to 2019 data from Payin-i-Qargha meteorological station). The significance of groundwater recharge by RWHRs for the regional and municipal water cycles is illustrated in Table 2. According to Table 2, compared to non-RWHRs circumstances, RWHRs will increase  $I_{c(eff)}$  by factors of 2.33 for the studied area.

Implementing RWHRs across the entire built-up area of the city may increase average annual infiltration from 4.86 MCM (million cubic meters) to 11.33 MCM for  $I_{c(eff)}$  values of 27.13% and 63.18%, which are calculated for study area for non-RWHRs and RWHRs con-



**Fig. 7** Maximum daily rainfall probability curve (2009–2019)

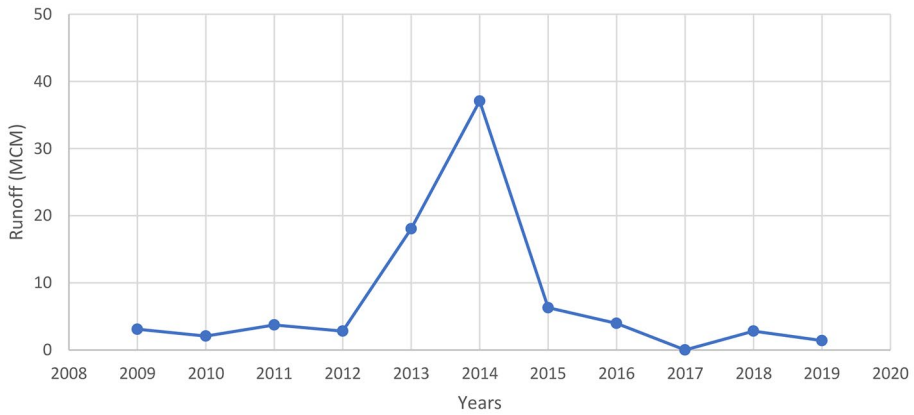
**Table 2** Calculated spatial proportion of LULC and their influence on  $I_{c(eff)}$  and groundwater recharge in the Kabul basin

LULC	Total area	Spatial fraction (%)	Calculated $I_{c(eff)}$		Calculated groundwater recharge (cu.m)	
			Without RWHS	With RWHS	Without RWHS	With RWHS
Built up	234.2079	48.06	27.13%	63.18%	4,864,510	11,328,409
Arable land	60.381	12.39				
Bare Soil	192.6495	39.53				

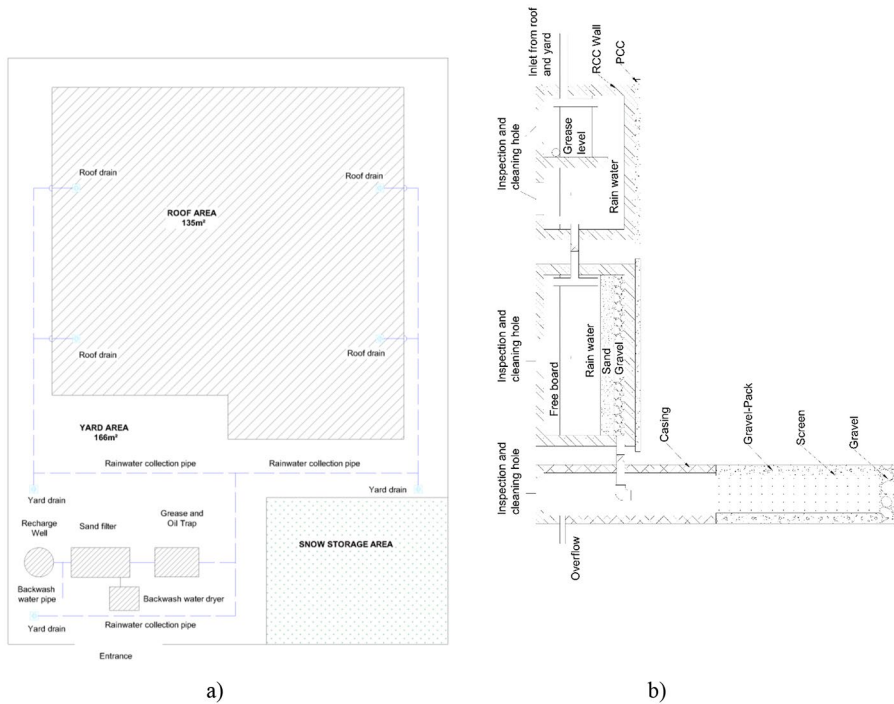
**Table 3** Curve Number values based on LULC.

No.	LULC	Area (sq. km)	Land cover area (%)	CN (%)	CNwi (%)
1	settlement (impervious area)	234.21	48.07	98	4710.707
2	cropland	60.38	12.39	76	941.8297
3	barren land	9.63	1.98	86	170.0537
4	barren rock	183.01	37.56	86	3230.306
<b>Total</b>		<b>487.2384</b>	<b>100</b>		<b>9052.896</b>

ditions, respectively. This value is obtained according to the total size of Kabul City ( $487.24 \text{ km}^2$ ) and the annual rainfall (368 mm). This straightforward calculation indicates how the Kabul Basin's groundwater recharge might rise by 6.5 MCM due to the deployment of RWHS, which is more than 200% greater than groundwater recharge under non-RWHS conditions. The computed  $I_{c(eff)}$  of 63.18% has a limited potential to create severe flood



**Fig. 8** Potential runoff time series



**Fig. 9** The proposed RWHS structure for residential houses a) location plan b) cross-section of recharge structures

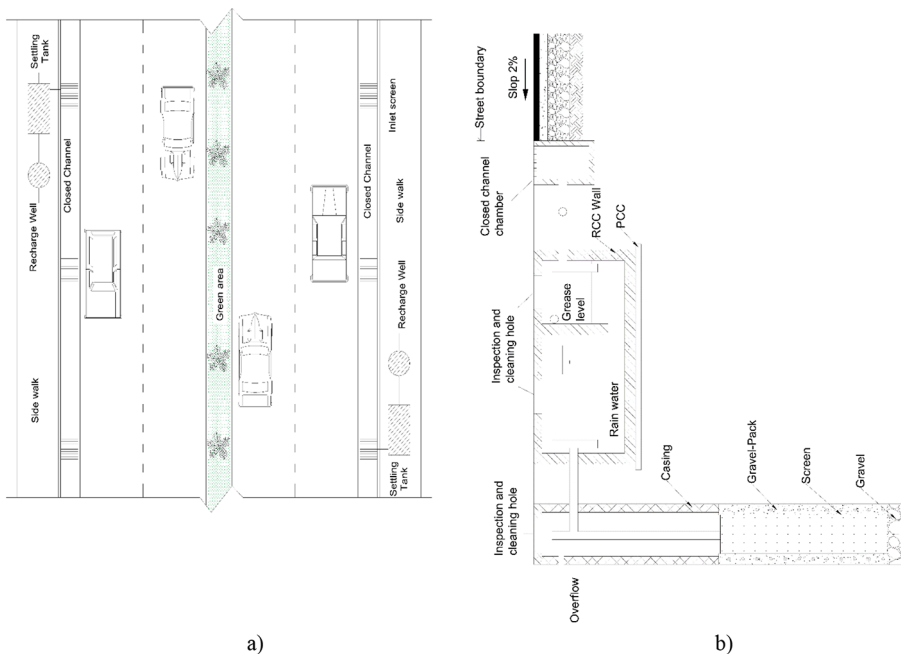
events concerning surface runoff. Based on Eqs. (3), (4), and (11), (12) it is predicted that implementing RWHS in the study region will result in a 36% reduction in surface runoff volumes for an average land cover of 48% built-up, 12% arable land, and 40% barren land. Please remove the text from here to the end of this paragraph. (11) (12),r

### 3.3 Surface Runoff Estimation

The land-use map of the Kabul Basin is depicted in Fig. 2. The correlation between rainfall and runoff is determined by the land cover, which equates to the soil's ability to retain water. The region is divided into five types: settlements, cropland, marshland, barren land, and barren rocks. According to soil characteristics studied by (Ahmadzai and Omuto 2019) the soil is deep and varies in texture from sandy loam to loam and clay. Therefore, based on soil classification by (USDA 2009) the hydrological soil group of the area is assigned to group B. Soils in this group have moderately low runoff potential when thoroughly wet.

The  $CN_{wi}$  values for each land use class are displayed in Table 3 concerning the associated region. The area's domination of impervious surfaces is confirmed by the  $CN_w$  value, which is determined to be 90.5% (Eq. 13). The initial abstraction and maximum soil retention are equal to 26.6 mm and 5.3 mm, respectively when the  $CN_w$  value is substituted in Eq. (12). This allows for the determination of the probable runoff depth for the daily precipitation. The amount of rainfall required to generate runoff is discovered to be 5.3 mm.

The highest daily precipitation for the year 2014 was 80.77 mm (Fig. 5). The accompanying runoff depth (potential) is calculated to be about 74.7 mm, resulting in a surface runoff volume of roughly 37 MCM (Fig. 10). Since many rainfall-runoff models employ a rainfall threshold of 5 mm/event for runoff production, the US-NRCS technique provides a fair approximation of the rainfall-runoff relationship in the research region (Mahmoud et al. 2014). A projected runoff volume of 18.06 MCM would result from the 42 mm rainfall that occurred in 2013 as an intense rainstorm event. This volume is far more than the capacity of the urban drainage system and has a return period of 5 years (Figs. 6 and 10). Consequently,



**Fig. 10** The proposed RWHRs structure for roads and streets a) location plan b) cross-section of recharge structures

it can be said that water harvesting in the research region has great potential to manage water deficits and mitigate drought.

### 3.4 Rooftop and Street Surface Rainwater Harvesting for Groundwater Recharge

Residential houses in Kabul city typically have an area of 200–400 m<sup>2</sup> and 2–4 stories. This study analyzed a residential home with a total area of 300 m<sup>2</sup>. About 60% of its total land area is considered for building, and the rest is considered a yard. The groundwater recharge well with its accessories, fed by roof and yard rainwater, is regarded inside the yard (Fig. 8a).

Considering the rainfall amount of 368 mm/year and taking into account the area of 300 m<sup>2</sup>, the total size of the residential house, including its yard, and taking into account 80% of the ability to collect rainwater, the total volume of rainwater that can be ordered for groundwater recharge is about 88 cubic meters. To prevent oil and silt from entering the absorption well, rainwater, after collection, enters the grease and oil trap basin, whose dimensions have been taken 150×150×100 cm, with an overflow height of 120 cm (Fig. 8b). The oil trap basin has a capacity of 1.8 m<sup>3</sup>. The sand filter basin installed after the oil trap basin has a dimension of 150×300×100 cm. Its lower part is contained filter material with a depth of 45 cm, and the upper portion includes a freeboard with a depth of 30 cm. The top layer (fine sand) has a thickness of 24 cm, and the bottom layer, which consists of gravel, has a thickness of 20 cm.

The maximum daily precipitation is about 81 mm/day with a return period of 12 years (Fig. 6). Since the soil characteristic in the Kabul basin has good hydraulic conductivity (Noori and Singh 2021a) the recharge wells have a total diameter of 1 m which its casing is about 0.8 m diameter. The recharge depth will be typically considered 20 m based on the maximum daily rainfall and the total catchment area (total residential house area).

To collect rainwater from roads, streets, and their sidewalk areas, the system of groundwater recharge wells with its accessories can be constructed at a distance of 100–150 m on the sides of the road (pedestrian area) (Fig. 9a). The distance between the wells depends on the side of the streets and the hydraulic conductivity of the soil. A settling tank has been considered before adding water to absorption wells to prevent silt and clay entry. The settling tank is separated into two parts by a buffer wall (Fig. 9b). The first part acts as the grease and oil separator, and the second is for solid particle settlement. The tank is 350×150×100 cm with a 20 cm freeboard. The settling tank is connected to a recharge well similar to recharge wells of residential houses based on construction.

## 4 Conclusions

An effort has been made to discuss the principles of rainwater harvesting systems and their potential for contributing to sustainable water management in urban areas. The study also examines the variability and concentration of precipitation in the specific study area. However, the expansion of cities and the rapid transformation of virgin lands into urban settings necessitate applying rainwater harvesting techniques in these locations to minimize the negative impact of urbanization on the local and regional water cycle. The benefits of



rainwater harvesting, such as groundwater recharge, reduction of surface runoff, and mitigation of flooding risks in metropolitan areas, are thoroughly discussed.

The rainfall pattern in Kabul is seasonal, with most rainfall occurring in February, March, and April. For instance 2019, the two wettest months (February and April) accounted for approximately 45% of the total annual precipitation. On average, the yearly rainfall is 368 mm, typically distributed over 80 wet days. Considering the land use and land cover typology, the effective imperviousness ( $I_c$ ) was calculated for the study area under two conditions: with rainwater harvesting systems (RWHRS) and without RWHRS. Utilizing RWHRS could increase yearly infiltration from 4.86 million cubic meters (MCM) to 11.33 MCM. Using the “US-NRCS” approach, a weighted CN value of 90.5% was determined, indicating the dominance of impervious surfaces. Furthermore, the threshold for runoff formation was found to be 5.3 mm of rainfall. Probability analysis using the Weibull approach predicts a return period of 3–4 years for daily rainfall below 30 mm.

Two approaches for collecting rainwater to recharge groundwater are described in the study. The first method involves implementing RWHRS for a residential house with an area of 300 m<sup>2</sup>, which can yield approximately 88 m<sup>3</sup> of water for groundwater replenishment. The second approach involves implementing RWHRS in street sidewalks to recharge the local aquifer. It is also suggested to consider implementing RWHRS for commercial buildings and public institutions. Additionally, it is recommended to prioritize the use of RWHRS in urban environments where a local aquifer is present, as it is an environmentally friendly approach that eliminates the need for complex pumping systems, reduces the space required for water storage tanks, and allows for fair distribution of the available rainwater resource to the entire urban population. The study also proposes the development of collection systems and absorption wells on a larger scale to capture rainwater in public and impervious areas.

These findings underscore the significant potential of rainwater harvesting and storage systems (RWHRS) as an effective strategy for managing urban floods and artificially recharging groundwater. By implementing RWHRS, cities can mitigate the impact of excessive rainfall and reduce the risk of flooding. The stored rainwater can also replenish groundwater levels, which is crucial for sustaining water resources in urban areas. This research emphasizes the importance of adopting RWHRS as a proactive and sustainable approach to address the challenges of urban flood management and groundwater recharge.

**Acknowledgements** The authors are grateful to the Department of Meteorology, General Directorate of Water Resources, Ministry of Energy, and Water of Afghanistan for providing precipitation data for the study.

**Author Contributions** All authors contributed to the study’s conception and design. Material preparation, data collection, data analysis, and the first draft of the manuscript were performed by Ali Reza Noori. Prof. S.K. Singh reviewed the work, revised it critically for important intellectual content, and approved the version to be published.

**Funding** The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

**Data Availability** The data used in the current research will be available through the corresponding author based on request.

## Declarations

**Ethical Approval** Not applicable.

**Consent to Participate** Not applicable.

**Consent to Publish** Not applicable.

**Competing Interests** The authors have no relevant financial or non-financial interests to disclose.

## References

- Ahmadzai H, Omuto C (2019) Afghanistan soil catalogue. FAO, Kabul, Afghanistan
- B R PB, Lokeshwari M (2021) Potential of rain water harvesting and Ground Water Improvement at RVCE. *Glob Jounal Res Eng C Chem Eng* 21
- Brati MQ, Ishihara MI, Higashi O (2019) Groundwater level reduction and pollution in relation to household water management in Kabul, Afghanistan. *Sustain Water Resour Manag* 5:1315–1325. <https://doi.org/10.1007/s40899-019-00312-7>
- DACAAR (2019) Hydro-geological Booklet Kabul Province. DACAAR, Kabul, Afghanistan
- Dhakate R, Rao VVSG, Raju BA et al (2013) Integrated Approach for identifying suitable Sites for Rainwater Harvesting Structures for Groundwater Augmentation in Basaltic Terrain. *Water Resour Manag* 27:1279–1299. <https://doi.org/10.1007/s11269-012-0238-3>
- Gado TA, El-Agha DE (2020) Feasibility of rainwater harvesting for sustainable water management in urban areas of Egypt. *Environ Sci Pollut Res* 27:32304–32317. <https://doi.org/10.1007/s11356-019-06529-5>
- Ghazavi R, Babaei S, Erfanian M (2018) Recharge Wells Site Selection for Artificial Groundwater recharge in an urban area using fuzzy logic technique. *Water Resour Manag* 32:3821–3834. <https://doi.org/10.1007/s11269-018-2020-7>
- Gwenzi W, Nyamadzawo G (2014) Hydrological impacts of urbanization and urban roof Water Harvesting in Water-limited catchments: a review. *Environ Process* 1:573–593. <https://doi.org/10.1007/s40710-014-0037-3>
- Huang Z, Nya EL, Rahman MA et al (2021) Integrated water resource management: rethinking the contribution of rainwater harvesting. *Sustain* 13:1–9. <https://doi.org/10.3390/su13158338>
- Hussain F, Hussain R, Wu RS, Abbas T (2019) Rainwater harvesting potential and utilization for artificial recharge of groundwater using recharge wells. *Processes* 7. <https://doi.org/10.3390/pr7090623>
- Jawadi HA, Sagin J, Snow DD (2020) A detailed Assessment of Groundwater Quality in Future Development. *Water* 1–19. <https://doi.org/10.3390/w12102890>
- Jawadi HA, Iqbal MW, Naseri M et al (2022) Nitrate contamination in groundwater of Kabul Province, Afghanistan : reasons behind and conceptual management framework discourse. *J Mt Sci* 19:1274–1291. <https://doi.org/10.1007/s11629-021-7002-1>
- JICA (2011) The study on groundwater resources potential in Kabul basin in the islamic republic of Afghanistan final report. JICA
- Jung K, Lee T, Choi BG, Hong S (2015) Rainwater Harvesting System for Continuous Water Supply to the regions with high Seasonal Rainfall Variations. *Water Resour Manag* 29:961–972. <https://doi.org/10.1007/s11269-014-0854-1>
- Krishna TM, Sudharsan RVK, Sudhangan B D (2021) Water Management through Rainwater Harvesting in Urban Areas. *Nat Volatiles Essent Oils* 8:6118–6124
- Machiwal D, Jha MK, Mal BC (2011) Assessment of Groundwater potential in a Semi-Arid Region of India using remote sensing, GIS and MCDM techniques. *Water Resour Manag* 25:1359–1386. <https://doi.org/10.1007/s11269-010-9749-y>
- Mack TJ, Akbari MA, Ashoor MH et al (2009) Conceptual Model of Water Resources in the Kabul Basin, Afghanistan. USGS
- Mack TJ, Chornack MP, Taher MR (2013) Groundwater-level trends and implications for sustainable water use in the Kabul Basin, Afghanistan. *Environ Syst Decis* 33:457–467. <https://doi.org/10.1007/s10669-013-9455-4>
- Mahmoud WH, Elagib NA, Gaese H, Heinrich J (2014) Rainfall conditions and rainwater harvesting potential in the urban area of Khartoum. *Resour Conserv Recycl* 91:89–99. <https://doi.org/10.1016/j.resconrec.2014.07.014>
- Manawi SMA (2020) Urban flooding and waterlogging in the northern. Part of Kabul City
- Manawi SMA, Nasir KAM, Shiru MS et al (2020) Urban flooding in the Northern Part of Kabul City: causes and mitigation. *Earth Syst Environ* 4:599–610. <https://doi.org/10.1007/s41748-020-00165-7>
- Mansell MG (2003) Rural and urban hydrology. Thomas Telford

- Martin SL, Hayes DB, Kendall AD, Hyndman DW (2017) The land-use legacy effect: towards a mechanistic understanding of time-lagged water quality responses to land use/cover. *Sci Total Environ* 579:1794–1803. <https://doi.org/10.1016/j.scitotenv.2016.11.158>
- Matomela N, Li T, Ikhumhen HO (2020) Siting of Rainwater Harvesting potential Sites in Arid or semi-arid Watersheds using GIS-based techniques. *Environ Process* 7:631–652. <https://doi.org/10.1007/s40710-020-00434-7>
- Nachshon U, Netzer L, Livshitz Y (2016) Land cover properties and rain water harvesting in urban environments. *Sustain Cities Soc* 27:398–406. <https://doi.org/10.1016/j.scs.2016.08.008>
- Noori AR, Singh SK (2021a) Status of groundwater resource potential and its quality at Kabul, Afghanistan : a review. *Environ Earth Sci* 80:1–13. <https://doi.org/10.1007/s12665-021-09954-3>
- Noori AR, Singh SK (2021b) Spatial and temporal trend analysis of groundwater levels and regional groundwater drought assessment of Kabul, Afghanistan *Environ Earth Sci* 80. <https://doi.org/10.1007/s12665-021-10005-0>
- Oliver JE (1980) Monthly precipitation distribution: a comparative index. *Prof Geogr* 32:300–309. <https://doi.org/10.1111/j.0033-0124.1980.00300.x>
- Pauleit S, Duhme F (2000) Assessing the environmental performance of land cover types for urban planning. *Landsc Urban Plan* 52:1–20. [https://doi.org/10.1016/S0169-2046\(00\)00109-2](https://doi.org/10.1016/S0169-2046(00)00109-2)
- Pell Frischmann (2012) Afghanistan Resource Corridor Development: Water Strategy Final Kabul River Basin Report Version 4.0
- Qi Q, Marwa J, Mwamila TB et al (2019) Making rainwater harvesting a key solution for water management: the universality of the Kilimanjaro Concept. *Sustain* 11:1–15. <https://doi.org/10.3390/su11205606>
- Ranaee E, Abbasi AA, Yazdi JT, Ziyade M (2021) Feasibility of rainwater harvesting and consumption in a middle eastern semiarid urban area. *Water (Switzerland)* 13:1–23. <https://doi.org/10.3390/w13152130>
- Sadeghi KM, Kharaghani S, Tam W et al (2019) Green Stormwater infrastructure (GSI) for Stormwater Management in the City of Los Angeles: Avalon Green Alleys Network. *Environ Process* 6:265–281. <https://doi.org/10.1007/s40710-019-00364-z>
- Sarma R, Singh SK (2021) Simulating contaminant transport in unsaturated and saturated groundwater zones. *Water Environ Res* 93:1496–1509. <https://doi.org/10.1002/wer.1555>
- Sayl KN, Muhammad NS, Yaseen ZM, El-shafie A (2016) Estimation the physical variables of Rainwater Harvesting System using Integrated GIS-Based remote sensing Approach. *Water Resour Manag* 30:3299–3313. <https://doi.org/10.1007/s11269-016-1350-6>
- Taher MR, Chornack MP, Mack TJ (2013) Groundwater levels in the Kabul Basin, Afghanistan, 2004–2013. USGS
- Tiwari K, Goyal R, Sarkar A (2018) GIS-based methodology for identification of suitable locations for Rainwater Harvesting Structures. *Water Resour Manag* 32:1811–1825. <https://doi.org/10.1007/s11269-018-1905-9>
- Tünnermeier T, Houben DG (2005) Hydrogeology of the Kabul Basin Part I: Geology, aquifer characteristics, climate and hydrography. BGR
- Uhl WV, Tahiri MQ (2003) An overview of groundwater resources and challenges. Vincent W. Uhl Uhl. Baron, Rana Associates, Inc., Washington Crossing, PA, USA
- USDA (2009) Chapet 7 Hydrologic Soil Groups. In: National Engineering Handbook. Washington DC
- Walsh RPD, Lawler DM (1981) Rainfall seasonality: description, spatial patterns and change through Time. *Weather* 36:201–208. <https://doi.org/10.1002/j.1477-8696.1981.tb05400.x>
- Wu RS, Molina GLL, Hussain F (2018) Optimal Sites Identification for Rainwater Harvesting in north-eastern Guatemala by Analytical Hierarchy process. *Water Resour Manag* 32:4139–4153. <https://doi.org/10.1007/s11269-018-2050-1>
- Zabidi HA, Goh HW, Chang CK et al (2020) A review of roof and pond rainwater harvesting systems for water security: the design, performance and way forward. *Water (Switzerland)* 12:1–22. <https://doi.org/10.3390/w12113163>
- Zaryab A, Noori AR, Wegerich K, Kløve B (2017) Assessment of water quality and quantity trends in Kabul aquifers with an outline for future drinking water supplies. *Cent Asian J Water Res* 3:3–11

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



# Role of Surface-Chemistry in Colloidal Processing of Ceramics: A Review

Megha Bansal<sup>1</sup> · Deenan Santhiya<sup>2</sup> · S. Subramanian<sup>3</sup>

Received: 9 March 2023 / Accepted: 4 September 2023  
© The Indian Institute of Metals - IIM 2023

**Abstract** Colloidal processing of ceramics has gained significant attention in recent years owing to its widespread application in biomedical and various industrial sectors. Polymer-assisted colloidal synthesis offers additional advantages and possibilities for development of advanced ceramic materials. This review enumerates the ancient techniques for ceramic production, the factors influencing the surface chemistry in colloidal processing of ceramics, together with the description of the interparticle forces, such as electrostatic and van der Waals interaction, steric and depletion, that contribute majorly to surface chemistry involved in colloidal processing of ceramics. The literature pertaining to the surface chemical interactions of various ceramic materials with polymeric additives are surveyed. Finally, the developments underlying major advancements in colloidal processing of ceramic materials are highlighted.

**Keywords** Ceramics · Polymeric additives · Surface chemical interaction · Interparticle forces · Bioceramics

## 1 Introduction

The archaeological pottery portrays ceramic components as one of the oldest manufactured materials [1]. Pottery artefacts were discovered at various archaeological sites worldwide, providing insights into the early stages of ceramic production. Over the centuries, ceramic technology has advanced with improvements in kiln design, glazing techniques, and the discovery of new clay types. Ceramic production expanded beyond pottery to include the manufacturing of tiles, bricks, figurines, sculptures, and other functional and artistic objects [2]. Hitherto, the typical system used to create ceramic materials was dry grinding of compacts [3]. Due to dry pressing, inhomogeneous densification could be observed that resulted in decreasing the strength and stability of the material [4]. For the manufacture of low-tech parts, dry pressing promised to be a novel approach, however, for applications that required high consistent quality ceramics, this approach proved to be disadvantageous. This is because of major reasons including: (1) de-agglomeration of dry powder or removal of defects present in the powder at the time of manufacturing was difficult, and (2) to achieve prior manufactured products, shape restrictions and disproportionate machining hindered the processing and also made it uneconomical for production of complex ceramic structures. Subsequently, dry pressing paved the way for colloidal synthesis techniques to create more meaningful shapes for widespread applications and extending their scope for both thick and penetrable ceramic creation [5]. The term “colloid” refers to the particles possessing one-dimensional size ranging from  $10^{-3}$  to  $1\ \mu\text{m}$ . All colloidal systems have a distinctive property showing large contact area between dispersion medium and the particles [6]. Colloidal synthesis enables the integration of different components or dopants into the ceramic matrix, leading to

✉ Deenan Santhiya  
deenan.santhiya@dce.ac.in

✉ S. Subramanian  
ssmani@iisc.ac.in

<sup>1</sup> Department of Biotechnology, Delhi Technological University, Delhi, India

<sup>2</sup> Department of Applied Chemistry, Delhi Technological University, Delhi, India

<sup>3</sup> Department of Materials Engineering, Indian Institute of Science, Bangalore, India

the development of composite or multifunctional materials. These systems combine the properties of different materials to achieve enhanced performance or new functionalities through regulation of basic suspension "structure" and its progression during formation [7–12].

The major strength and reliability of ceramic materials is attained by inter-particle forces that control isolation and elimination of agglomerates and defects during colloidal processing [13]. Polymer-assisted colloidal synthesis offers advantages such as the ability to control the morphology, porosity, and structure of the final ceramic material. It allows for the fabrication of complex and tailored ceramic structures. The use of polymers as templates or precursors also enables the synthesis of ceramic composites and hierarchical structures [14, 15]. The advancements in the field of ceramic material production have expanded their application in contemporary life including their usage in electronic components, aerospace, automobile, medical equipment, industrial applications and chemical processing [16].

The main focus of this paper is to review the surface chemistry involved in the colloidal processing of ceramics. The review is divided into various sections beginning with a brief history of ancient techniques adopted for the manufacture of ceramics. This is followed by a description of the factors influencing surface chemistry of ceramic materials with polymeric additives. Additionally, various inter-particle forces, such as van der Waals, electrostatic, and steric controls governing colloidal processing in the presence of polymers and electrolytes, are discussed. The surface-chemical interactions of ceramic materials with polymeric additives and surfactants are enumerated. Finally, developments and applications underlying the importance of surface-chemistry in colloidal synthesis for production of bio-ceramics, adoption of additive manufacturing techniques and molecular modelling tools as sustainable strategies for ceramic production are highlighted.

## 2 Historical Perspective

The origins of ceramic processing using colloidal science can be traced to the ancient manufacturing of earthenware products exploiting the plasticity of clays. The clay particles dispersed in water were the first colloidal dispersion employed by humans, but without a clear comprehension of the process. [17]. Intercalation of water between clay particles allowed formation of a double-layer-like force that helped to maintain plasticity and also permitted it to deform on application of shear stress. Since centuries, this mixing of ceramics with water did not only serve as paste for pottery formation, but was also used for formation of slips for processes like slip casting. At the end of the eighteenth century,

slip casting was used for manufacturing a large number of clay products [18]. One of the notable milestones in the development of colloidal processing was the work of the Scottish scientist, Thomas Graham, referred to as the "father of colloid chemistry." In the mid-1800s, Graham conducted extensive studies on the diffusion of substances in solution and recognized the presence of particles dispersed in liquids, now known as colloids. His work laid the foundation for understanding colloidal phenomena and their significance in various applications since 1845 [19]. Following this, in 1909, Ashley applied the knowledge of colloid theory to describe the properties of clay [20]. Later during 1950s, suspension-based processes for manufacturing metal-ceramics (or cermets) were developed, though the real application only started after three decades. Ceramic materials gained popularity in the industrial sector in 1980s and 90 s after development of pure materials for production of advanced ceramics having high density. In a seminal paper, it was demonstrated for the first time that by adopting colloidal processing route, high-strength sintered alumina with 1.04 GPa bend strength could be obtained [21]. Eventually, colloidal processing played a crucial role in the manufacturing of advanced ceramics with a high degree of dependability [22].

## 3 Factors Influencing the Surface Chemistry of Ceramic Materials with Polymeric Additives

The surface chemistry of ceramic materials can be influenced by several factors when polymeric additives are introduced. The surface qualities of any polymer inevitably affect the properties of ceramics including wetting, bonding, contact, and biocompatibility. Some of the key factors that can impact the surface chemistry of ceramic materials with polymeric additives are highlighted below:

- (a) *Polymer composition* The composition of the polymer additive can significantly affect the surface chemistry of ceramic materials [23]. Different polymers possess unique chemical properties, such as functional groups and molecular weight, which can interact differently with the ceramic surface [24]. The presence of specific functional groups can facilitate chemical bonding or interactions between the polymer and the ceramic material.
- (b) *Polymer concentration* The concentration of the polymer additive plays a crucial role in surface chemistry. Higher concentrations of polymers can lead to increased polymer-ceramic interactions, resulting in modified surface properties [25]. Additionally, the concentration can influence the phase separation behaviour

of the polymer within the ceramic matrix, affecting the distribution of the polymer at the surface.

- (c) *Polymer molecular weight* The molecular weight of the polymer additive can influence the surface chemistry of ceramics. Higher molecular weight polymers tend to exhibit stronger interactions with the ceramic surface due to increased chain entanglements and extended conformations. This can lead to enhanced adhesion and surface modification effects (Fig. 1) [26].
- (d) *Polymer compatibility* The compatibility between the polymer and ceramic material is essential for achieving desired surface chemistry [27]. Compatibility ensures good wetting and spreading of the polymer on the ceramic surface, leading to improved adhesion and surface modification [28]. Incompatibility can result in phase separation or poor interactions, limiting the effectiveness of the polymer additive.
- (e) *Processing conditions* The processing conditions during the incorporation of the polymer additive can affect the surface chemistry of ceramics [29]. Parameters such as temperature, pressure, and duration of processing can influence the degree of polymer-ceramic interaction and the formation of interfacial bonds [30]. Proper processing conditions are crucial to achieving desirable surface modifications.
- (f) *Surface preparation* The surface preparation of the ceramic material before incorporating the polymer additive is critical. Cleaning, roughening, or function-

alizing the surface can enhance the interaction between the polymer and ceramic surface, leading to improved adhesion and surface modification [31]. The surface roughness and chemistry play a role in determining the strength and stability of the polymer-ceramic interface.

- (g) *Environmental factors* Environmental conditions, such as humidity, temperature, and exposure to chemicals, can affect the surface chemistry of ceramic materials with polymeric additives [32]. These factors can influence the stability, degradation, or reactivity of the polymer on the ceramic surface, potentially altering the surface properties over time.

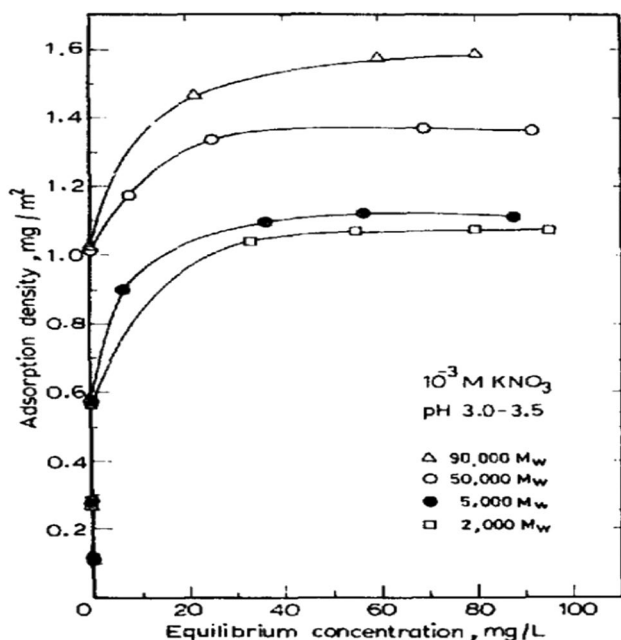
It is important to note that the specific combination of ceramic material and polymeric additive, as well as the intended application, dictate the extent and nature of the surface chemistry modifications. Different ceramic systems and polymers exhibit unique interactions and behaviour, making it essential to consider these factors when designing and engineering ceramic materials with polymeric additives.

#### 4 Forces Governing the Surface Chemistry in Colloidal Processing of Ceramics

Various surface chemical forces are involved in the colloidal processing of ceramics in the presence of polymeric additives. For efficient processing of these ceramic and polymeric additives, control of inter-particle attraction or repulsion, governed by Derjaguin–Landau–Verwey–Overbeek (DLVO) forces, is mostly studied [33] (Fig. 2).

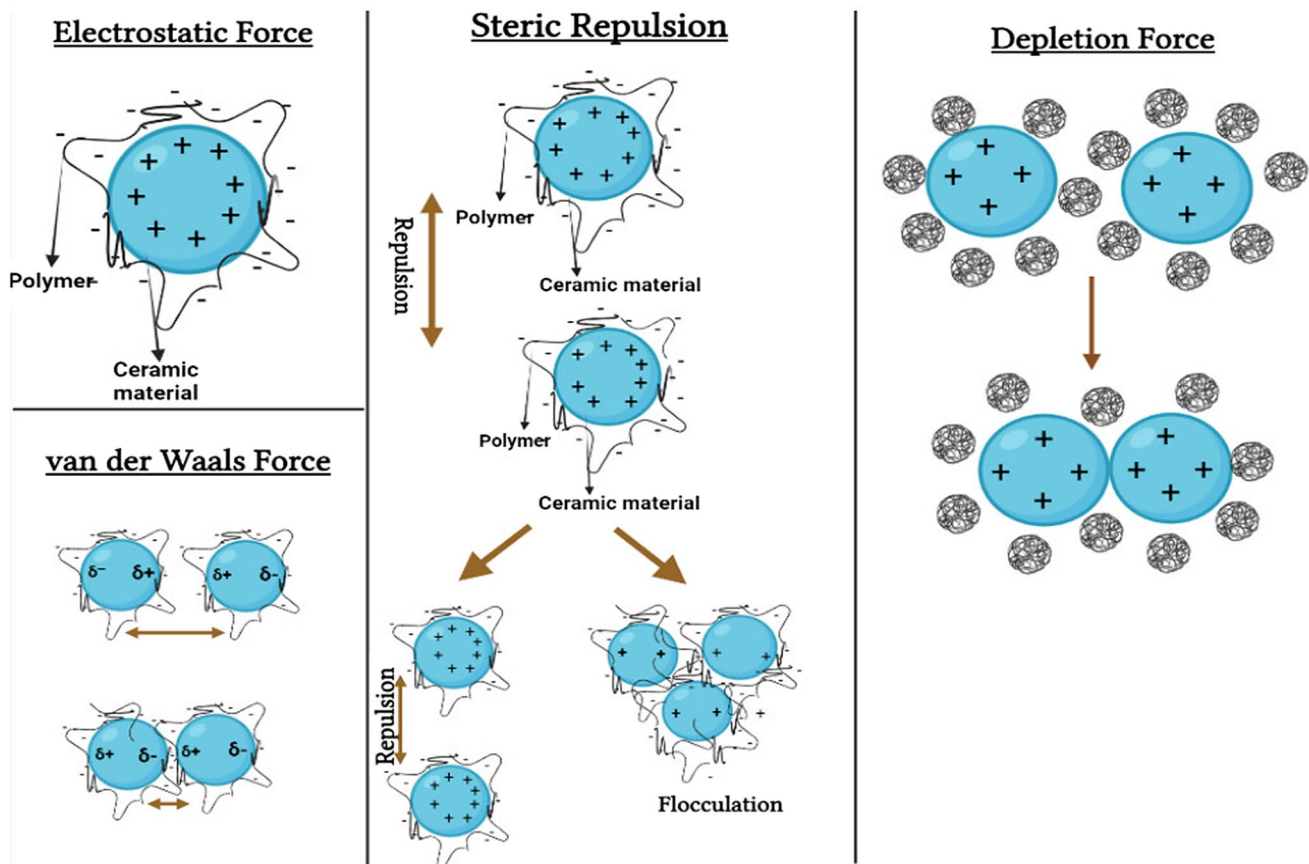
##### i. Electrostatic forces:

Colloidal particles often carry an electrical charge on their surface, which leads to electrostatic interactions between them. These forces arise due to the presence of ions in the solution or dissociation of functional groups on the particle surface. The magnitude of the electrostatic force decreases with increasing distance between the particles. A medium with high dielectric constant, like water, enables the substance to acquire a surface electric charge [34]. Stern or Helmholtz layer contains some of the immobile counter-ions rigidly bound to its surface [35]. A high electric field at the surface causes dissociated ions to induce repulsive interaction, whereas a stronger coulombic attraction is generated towards the surface [36]. As a result, dissociated ions and charged surface of particle form an electrical double layer [37]. On close proximity of two similar charged particles, an electrostatic repulsion occurs between two surfaces as the double layers overlap [38]. Thus, electrostatic stabilization prevents particles from agglomerating together [39]. The most



**Fig. 1** Adsorption isotherms of polyacrylic acid of different molecular weights on alumina (Adapted from Santhiya et al., 1998 with permission)





**Fig. 2** Forces governing the surface-chemistry involved in colloidal processing of ceramics

common additives imparting electrostatic stabilization to colloidal suspensions are polyelectrolytes. Polyelectrolytes contain an ionizable group like sulfonic or carboxylic group, that blocks copolymers containing ionizable segments [40]. Investigations on alumina and zirconia suspensions as a function of pH showed that electrical double layer forces acted as regulators in determining the behaviour of suspensions. There was a significant difference in single and mixed suspensions with respect to pH. The DLVO theory was used to study the effect of electric double layer force on yield stress [41, 42].

ii. Van der Waals forces

Van der Waals forces are a type of intermolecular force that arise due to fluctuations in electron distributions. These forces include London dispersion forces, which arise from temporary fluctuations in electron density, and Keesom forces and Debye forces, which arise due to the orientation and induced polarization of dipoles [43]. Van der Waals forces are generally attractive and act over short distances. They become significant when particles are in close proximity to each other, even in the absence of an electrical charge. All

the ceramic materials are subjected to van der Waals attractive forces. The value of the Hamaker constant is an assessment of the magnitude of interaction between ceramic materials and polymeric additives [44]. A higher Hamaker constant indicates stronger van der Waals forces, suggesting a stronger interaction between the ceramic material and the polymeric additive. Conversely, a lower Hamaker constant suggests weaker interaction. By knowing the Hamaker constant of the ceramic material and the polymeric additive, researchers and engineers can assess the potential for adhesion, compatibility, or dispersion between these materials. This information can be crucial in various applications, such as composite materials, coatings, or adhesives, where the interaction between ceramics and polymers is of interest [45].

iii. Steric forces

Steric or entropic forces arise from the repulsion between particles due to the overlap of their surface layers [46]. These forces are influenced by the size, shape, and surface chemistry of the particles, as well as the presence of polymer chains or other molecules attached to the particle surface. Steric forces can pre-

vent particles from coming too close to each other and thus stabilize colloidal dispersions [47]. Therefore, short-range structure of densely packed colloidal particles can be determined by steric force [48]. The presence of steric forces has important implications for the short-range structure of densely packed colloidal particles. Due to the repulsive nature of steric forces, they can prevent or inhibit the formation of close-packed arrangements or aggregation of particles. Instead, the particles tend to arrange themselves in a more dispersed or loosely packed manner, forming structures, such as random close-packed or disordered arrangements [49]. Electrolyte concentration does not affect steric interactions between particles and is stable at low and high concentration of particles [50]. Longer and more densely packed polymer chains tend to provide stronger steric repulsion, effectively increasing the distance at which particles can approach each other before experiencing significant repulsion [34].

iv. Depletion forces

The interaction between large colloidal particles suspended in continuous phase consisting of polymers, non-ionic surfactants, polyelectrolytes and small particles is termed as depletion interaction [51]. These forces arise from the entropic effects caused by the presence of non-adsorbing polymer molecules or large particles in the colloidal system. When non-adsorbing polymer molecules or large particles are added to a colloidal suspension, they can create an excluded volume effect. This effect leads to a decrease in the available volume for colloidal particles to move freely [52]. As a result, colloidal particles tend to aggregate or cluster together to minimize the entropic penalty associated with their restricted movement [53]. On increasing the free polymer concentration, the thickness of depletion layer decreases [54]. The depletion attraction occurs when the non-adsorbing polymers or particles create a region of lower particle density between them. This leads to an effective attraction between colloidal particles, causing them to come closer together. On the other hand, the depletion repulsion occurs when the non-adsorbing components create a region of higher particle density, resulting in an effective repulsion between colloidal particles [55]. On increasing the depletant concentration, non-adsorbing molecules induce depletion stabilization [56]. Electrostatically stabilized colloidal particle exhibits similar interaction as non-adsorbing polymer at large separation [57]. Demixing polymer segments from solvent results in repulsive interaction and inhibits particle aggregation thereby inducing depletion stabilization [58]. The flocculation of colloidal suspensions using population bal-

ance model has also been demonstrated for polymer bridging on ceramic suspensions [59].

## 5 Surface-Chemical Interactions of Ceramic Materials with Polymeric Additives

For providing excellent property to the formed ceramics, the combination of different constituents (microarchitectures) integrating ceramic and polymeric layers becomes essential [60]. Table 1 provides a brief overview of the most commonly used polymeric additives for ceramic materials. Different polyelectrolytes have been used with ceramic powders to study the electrokinetic and dispersion characteristics of ceramic suspensions [61]. Some examples of various ceramic systems are illustrated in this section.

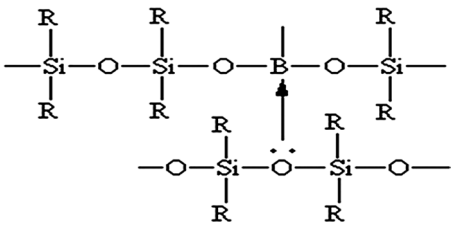
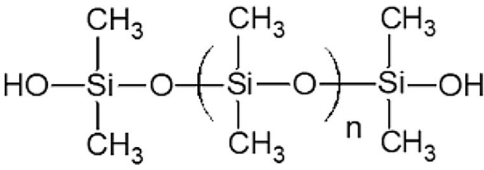
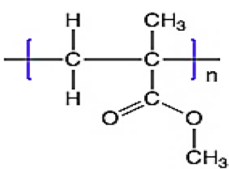
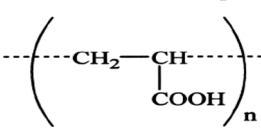
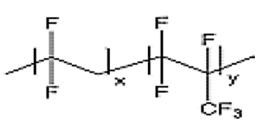

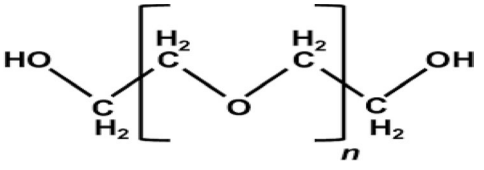
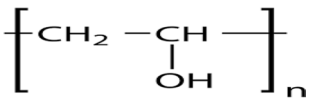
### 5.1 Single Oxide Systems

Poly (acrylic acid) (PAA) and poly (vinyl alcohol) (PVA) were adsorbed onto alumina to study their surface-chemical interactions. On increasing the pH, the adsorption density of PAA decreased, while the opposite was observed for PVA (Fig. 3). High-affinity Langmuirian trend was observed for adsorption isotherm of PAA at acidic pH range, whereas it was low for PVA. After analysing the desorption studies, it was determined that around 80% PVA was desorbed in 3–9 pH range, whereas it was 70% for PAA in the pH range of 3–8. FTIR spectroscopic studies lent support for surface chemical interaction and hydrogen bonding mechanisms for PAA-alumina, compared to hydrogen bonding only for PVA-alumina system [63].

Spurred by the above study, Saravanan and Subramanian investigated the surface properties of zirconia suspension with widely used dispersants, ammonium poly(methacrylate) (APMA) and poly (ethylene glycol). In presence of APMA, adsorption density of PEG significantly reduced, however, the reverse was not true. Dispersion studies revealed that zirconia suspension was more stable in the presence of APMA, whereas no characteristic difference could be observed in the presence of PEG. Complexation of zirconium species with APMA was revealed by co-precipitation tests and FTIR spectral studies confirmed hydrogen bonding and complexation for APMA and hydrogen bonding alone for PEG, as the interaction mechanisms with zirconia [64].

Near-net shape alumina ceramics could be produced using the hydrolysis-induced hydrogel casting (GCHAS) technique [67]. Suspensions containing 50% solids yielded approximately 95% alumina forms of about 300 mm height and 130 mm diameter using GCHAS technique. Alumina was dispersed in an aqueous solution of methylene bisacrylamide and methacrylamide with polycarboxylic acid as a dispersing agent. The sintered (1600 °C, 2 h) mechanical

**Table 1** Polymeric additives for ceramic materials

Name of polymer	Structure of polymer	References
Polyborosiloxanes		[62]
Polysiloxanes		[62]
Polyacrylic acid		[63]
Poly(methyl methacrylate)		[64]
Poly(Vinylidene fluoride) Hexafluoropropylene		[65]
Polycarbosilane		[62]
Polyethylene glycol		[64]
Polyvinyl alcohol		[66]

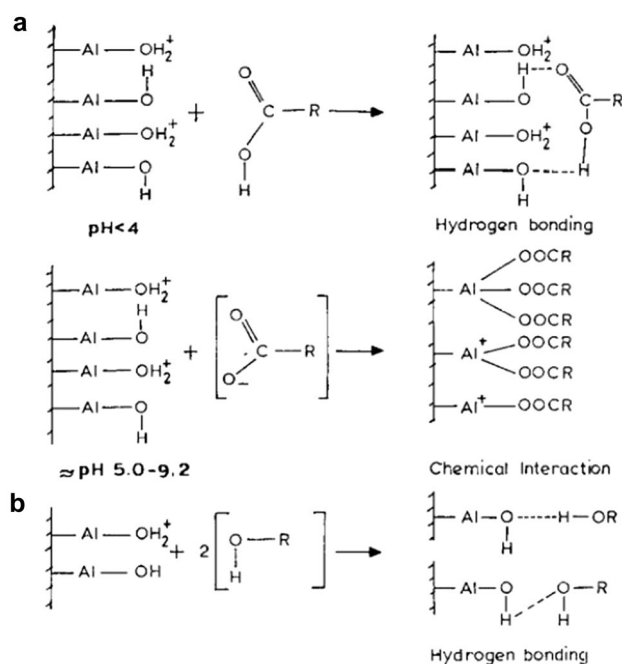
properties were essentially similar to those obtained by gel casting (GC), hydrolysis assisted (HAS) and normal dry pressing of lyophilized granules [67].

Slip casting was used to produce transparent alumina ceramics using three stabilization methods: (a) plasma treatment (b) electrostatic stabilization and (c) electrosteric stabilization [68]. Electrostatic stabilisation performed best in terms of all examined metrics (green body porosity and density, sintering temperatures, densities, and real in-line transmittance (RIT)). Plasma treatment produced similar outcomes, but exhibited low RIT, attributable to tiny cracks that reduced transparency. Despite having much bigger

pores, lower green density, and higher grain size, the standard electrosteric technique produced samples with higher transparency, compared to the plasma method [68].

## 5.2 Mixed Ceramic Oxide Systems

The surface chemical interactions using various polymeric additives was investigated in the presence of xanthan gum taking alumina and silica suspensions as ceramic materials [69]. In presence of xanthan gum, the adsorption maximum was observed at pH 2 for alumina, whereas no adsorption took place on silica. On increasing the xanthan gum



**Fig. 3** Schematic representation of the interaction between **a** Alumina-PAA **b** Alumina-PVA (Adapted from Santhiya et al., 1999 with permission)

concentration, the isoelectric point of alumina shifted to the acidic pH range. Flocculation of alumina suspensions increased to approximately 92% on addition of xanthan gum beyond 50 ppm. However, silica suspension showed no observable effect on electrokinetic and flocculation property after xanthan gum addition [69].

Zirconia-alumina suspensions were used to study the effects of powder size, coverage, solid loading and dispersion during processing [70]. A tight packing size distribution alone was not enough for attaining high-density crude compacts and hence it was necessary to simultaneously control the suspension rheology while optimizing solid loading and dispersion dosage. Furthermore, it was emphasized that high green density did not guarantee the best sintering condition, especially if it was achieved by adding coarse powder [70]. A model to generate shear from a mixture of alumina-zirconia flocculated suspensions was developed, which could provide quantitative information on solids, composition, and particle size distribution for a better understanding of fine particles [71]. Clear macroporous ceramics composed of titania and zirconia were produced by two techniques, namely, by template-assisted colloid processing and the other by heterocoagulation. In both cases, consistent and uniform macropores were produced [72]. Yttria-stabilized tetragonal zirconia ceramic suspensions (Y-TZP) were investigated using colloid route to fabricate green bodies by slip casting [73]. The hardness (H) value of 17 and Young's modulus (E) of 250 GPa were obtained by

nano-indentation. The samples were then subjected to low temperature degradation (LTD) treatment with conditions set at 130 °C, 240 h with 60-h increments. The Raman spectroscopy studies showed monoclinic phase in all samples with the Raman peaks centred at 180, 191, and 383  $\text{cm}^{-1}$ , while nano-indentation after 240 h of LTD showed that H and E values were reduced to 10 and 175 GPa, respectively, indicating a significant reduction in mechanical properties [73]. Colloidal processing of tetragonal zirconia polycrystalline (Y-TZP) suspensions stabilized with yttria along with various dispersants resulted in highly transparent Y-TZP [74].

Zirconia-alumina composites ZTA-30 (70  $\text{Al}_2\text{O}_3$ : 30  $\text{ZrO}_2$ ) and ZTA-60 (40  $\text{Al}_2\text{O}_3$ : 60  $\text{ZrO}_2$ ) of utility in orthopaedic applications were prepared by colloidal processing and agglomerated by hydrolysis, slip and gel casting of ceramic mixtures containing 50% solids by volume [75]. For evaluation, the same ceramic composite was also stabilized with freeze-dried particle inlay (FG). Green (three-point bend) strengths (B 17 MPa) of ceramics consolidated using gel casting (GC) were higher than those cemented using other procedures. Additionally, after 1 h of sintering at 1600 °C, the GC ceramics demonstrated superior fracture toughness and flexural strength qualities, when compared to those cemented by other processes, including traditional die pressing (FG) [75].

### 5.3 Electroceramics

The tape casting of lead zirconate titanate (PZT) using dispersant D3021, poly(vinyl alcohol) binder, Surfynol SE-F wetting agent, and poly(propylene glycol) plasticizer [76] was carried out through the colloidal processing route. Measurements of pH, conductivity, and zeta potential were performed to assess the suspension stability in the presence of a dispersant. The isoelectric point of PZT was located approximately at pH 6.5. The results exhibited shear thinning rheological behaviour, corresponding to a weakly flocculated system with low time-dependent viscosity effects. Rheological and mechanical properties were further studied [76]. The thermal stability and degradation of  $\text{Al}_2\text{O}_3$  and (Ba,Sr) $\text{TiO}_3$  (BST), and other organic additives for ceramic manufacturing were observed by mass spectrometry and thermal gravimetric analysis [77].

Ammonium polymethacrylate surfactant was used to disperse lead lanthanum zirconate titanate (PLZT) particles at different pH levels. The adsorption density of the polymer and slurry rheology was studied as a function of pH, to explain the stabilization mechanism. At the optimum surfactant concentration and pH, loading of suspension up to 50% solids by volume, at a relatively low viscosity was achieved [78].

## 5.4 Ultra-high Temperature Ceramics

Colloidal processing of zirconium diboride ( $\text{ZrB}_2$ ) ultrahigh temperature ceramics (UHTC) was investigated for formulating lattice forming systems. Colloidal route formed samples with higher density compared to samples produced by dry route with 99% density attained by hot pressing. It is envisaged that complex shaped ceramics can be formed from above-mentioned methods [79]. The suspension dispersion and colloidal processing of complex shaped ultra-high temperature  $\text{ZrB}_2$  ceramics has been recently reviewed [80].

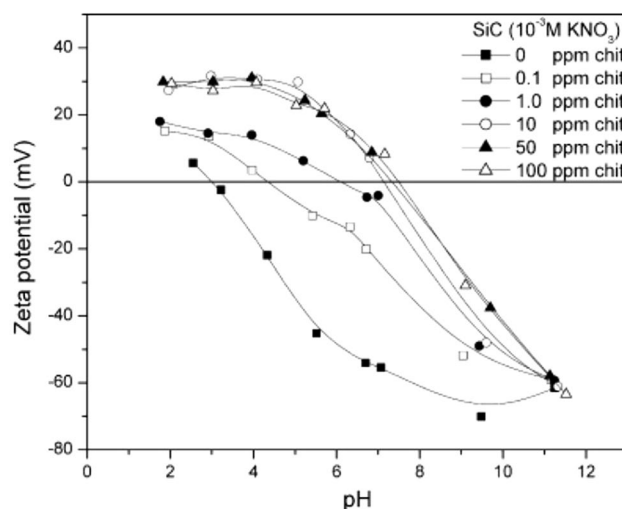
## 6 Developments and Applications Underlying Colloidal Synthesis for Ceramic Production

Colloidal processing for ceramic manufacturing imparts control over rheological properties, interparticle forces, structural evolution and drying behaviour thereby eliminating heterogeneities [81]. The major principle behind this methodology focuses on understanding interrelations between colloidal structure and its various properties [82]. Although the hypothetical comprehension of interparticle forces is fairly well established, the direct force estimations of ceramic materials having different colloidal chemistry, composition and crystallography are essential. The major consideration for processing ceramics requires stability of fabrication process to accommodate a wide range of properties.

### 6.1 Biopolymer Additives

Use of biopolymer additives for ceramic manufacturing offer great potential for colloidal processing [83]. Typical biopolymers used for ceramic applications are summarized in Table 2. For example, chitosan, a cationic bio-polymer, is considered to be biodegradable and non-toxic with a wide range of applications in pharmaceuticals, cosmetics, food processing and biomaterials [84]. Saravanan et al. (2006),

studied the surface chemical interactions of chitosan with silicon carbide (SiC) demonstrating maximum adsorption density at pH 7. On addition of chitosan to SiC suspension, zeta potential of SiC surface was +30 mV at pH 2–5, whereas it increased to +40 mV at pH 9–11. Based on these findings, it was found that chitosan served as an efficient biopolymer for SiC synthesis (Fig. 4) [85]. Saccharides (sucrose, galactose, lactobionic acid and L-ascorbic acid) are sustainable organic additives applied in colloidal synthesis [86]. They are commonly used as polymeric binders and dispersants owing to their non-toxic, and water-soluble properties [87]. Gel-casting of foams followed by use of surfactants and gelation of polysaccharides results in stable ceramic suspensions [88]. The sintering process generates a clean burnout on using saccharide-based derivatives compared to the conventionally used organic additives. Organic synthesis allows for modification of saccharides in different



**Fig. 4** Zeta potential values of SiC suspension as a function of pH, before and after interaction with chitosan (Adapted from Saravanan et al., 2006 with permission)

**Table 2** Biopolymers for ceramic applications

Sr. No	Biopolymer	Application	References
1.	Gelatin	Cosmetics, pharmaceutical industry, distillery, fingerprinting	[90]
2.	Cellulose	Biofuel, blending agent, drug delivery, coating	[91]
3.	Lignin	Photocatalyst, water treatment, immune booster, synthesizing polymers	[92]
4.	Starch	Agriculture, food industry, drug delivery	[93]
5.	Dextran	Anticoagulant, surgical treatment	[94]
6.	Curdan	Food industry, immunomodulator, therapeutics	[95]
7.	Pullulan	Adhesives, binder, vegetables, paper industry, flocculant	[96]
8.	Gellan gum	Stabilizer in food, tissue regeneration and bone repair	[97]
9.	Agar	Electrophoresis, selection of antibiotic, chromatography, purification	[98]
10.	Scleroglucan	Construction, oil engineering, pharmaceutical industry, thickener	[99]



functional groups thereby opening a new range of applications in ceramic technology [89].

## 6.2 Additive Manufacturing

Additive manufacturing of three-dimensional (3D) structures commonly scaffolds, facilitates its use in tissue engineering applications [100]. Polycarbonate, polyamide and polyether ether ketone (PEEK) are widely used thermoplastic polymers for scaffold engineering using selective laser sintering technology [101]. Tribological testing and mechanical properties confirm new estimates on strength and rupture of 3D samples including a composition of PEEK + nano titanium oxide + or hydroxyapatite ceramics, thereby proving to be useful materials for biomedical applications [102]. Casting and machining are traditional fabrication methods for manufacturing of ceramics which are now being widely replaced by 3D printing that enables preparation of highly complex structures using computer-aided design (CAD) [103]. Ceramic 3D printing can be categorized on the basis of feedstock form as slurry-based, powder-based and bulk solid-based [104]. To further enhance the functional parts of ceramic structures, 3D printing process is integrated with feedstock preparation and post-treatments [105]. 3D printing also offers an advantage of forming porous ceramics [106]. In manufacturing of porous ceramics, three-dimensional lattice is separated into numerous small units, followed by spraying of binder on the ceramic surface bed, aided under computer control. Rheological properties and liquid content of binder are adjusted prior to extrusion of binder from print head [107]. Therefore, porous ceramics are formed by combination of hollow spheres on powder bed and binder. Hollow spheres allow to design the micropore size of porous ceramics thereby attaining porosity of 97% useful for application in industry and biomedical environment [107].

## 6.3 Bioactive Ceramics

Bioactive ceramics find their application for skeletal reconstruction as a result of diseases such as osteoporosis in aged or obese, globally [108]. Areas of bone which are too large require healing by bioactive materials that replace damaged bone with glass ceramics. Bioactive glass–ceramics are polycrystalline in nature having very low contents of alkali oxides [109]. Apatite-wollastonite (A-W) is the most common glass–ceramic possessing high bending strength, fracture toughness and Young's modulus, compatible for orthopaedic applications [110]. This ceramic is majorly used for load-bearing areas of bone and is also used for vertical replacements, where high compressive strength is required [111].

## 6.4 Molecular Modelling

Molecular modeling techniques provide a powerful means to gain molecular-level insights into the interactions between ceramic powder surfaces and dispersants [112]. By understanding these interactions, researchers can develop strategies to enhance dispersion, improve the stability of ceramic suspensions, and optimize the formulation of dispersants for ceramic powder processing [113]. Molecular dynamics (MD) simulations can be used to study the behaviour of ceramic powder surfaces and dispersant molecules at the atomic level [114]. By simulating the motion and interactions of individual atoms over time, MD simulations can provide information on the structure, dynamics, and energetics of the system. The simulations can capture the adsorption and desorption of dispersant molecules onto the ceramic powder surface and explore the effects of different factors, such as pH, temperature, and concentration [115].

## 7 Summary

Colloidal processing of ceramics has been widely practised since ancient times. The traditional methodologies adopted for synthesis of ceramics like slip casting and drying processing time have limitations to form complex shapes. The surface-chemistry involved in the colloidal processing of ceramics using different organic additives strongly influences the behaviour and rheological characteristics of ceramic materials. The study of interparticle forces such as electrostatic, van der Waals, steric and depletion is important to better understand the surface chemical properties of ceramic suspensions. From a conventional perspective, ceramic applications had been limited to a narrow range. However, advancements in the field of colloidal processing using additive manufacturing (3D printing), synthesis of biopolymer composites, bioceramics, and porous ceramics along with shaping of complex ceramic structures, have paved the way for the utilization of ceramics in a wide range of applications such as tissue regeneration, bone formation, scaffold generation, injectable ceramics and also in coatings and inks. Thus, it is envisaged that the future advancements in colloidal processing of ceramics can significantly contribute in providing better opportunities for utilization of ceramic materials.

**Acknowledgements** The authors are honoured to dedicate this contribution to Professor P C Kapur, a doyen in the field of mineral processing and process metallurgy.



## References

- Owen H, *Two Centuries of Ceramic Art in Bristol*, Bell and Daldy (1873).
- Bechthold M, Kane A, and King N, *Ceramic Material Systems: In Architecture and Interior Design*, Birkhäuser (2015).
- Colombo P, *Philos Trans R Soc Math Phys Eng Sci* **364** (2006) 109. <https://doi.org/10.1098/rsta.2005.1683>
- Scharnweber J, Chekhonin P, Oertel C-G, Romberg J, Freudenberger J, Jaschinski J, and Skrotzki W, *Adv Eng Mater* **21** (2019) 1800210. <https://doi.org/10.1002/adem.201800210>
- Moreno R, *J Eur Ceram Soc* **40** (2020) 559. <https://doi.org/10.1016/j.jeurceramsoc.2019.10.014>
- Chin'as-Castillo F, and Spikes H A, *J Tribol* **125** (2003) 552. <https://doi.org/10.1115/1.1537752>
- Pugh R J, and Bergstrom L, *Surface and Colloid Chemistry in Advanced Ceramics Processing*, CRC Press (2017).
- Bell N S, Monson T C, DiAntonio C, and Wu Y, *J Ceram Sci Technol* **7** (2016) 1. <https://doi.org/10.4416/JCST2015-00025E>
- Bourgeat-Lami E, *J Nanosci Nanotechnol* **2** (2002) 1. <https://doi.org/10.1166/jnn.2002.075>
- Goodwin J, *Colloids and Interfaces with Surfactants and Polymers*, John Wiley & Sons (2009).
- Hiemenz P C, and Rajagopalan R, *Principles of Colloid and Surface Chemistry*, Revised and Expanded, 3rd Edition (2016). <https://doi.org/10.1201/9781315274287>
- Lange F F, *J Am Ceram Soc* **72** (1989) 3. <https://doi.org/10.1111/j.1151-2916.1989.tb05945.x>
- Franks G V, Tallon C, Studart A R, Sesso M L, and Leo S, *J Am Ceram Soc* **100** (2017) 458. <https://doi.org/10.1111/jace.14705>
- Blanco López M C, Rand B, and Riley F L, *J Eur Ceram Soc* **20** (2000) 1587. [https://doi.org/10.1016/S0955-2219\(99\)00241-1](https://doi.org/10.1016/S0955-2219(99)00241-1)
- Singh B, Bhattacharjee S, Besra S L, Sengupta D, and Misra V, *Trans Indian Ceram Soc* **63** (2004) 1. <https://doi.org/10.1080/0371750X.2004.11012121>
- Colombo P, Mera G, Riedel R, and Sorarù G D, *J Am Ceram Soc* **93** (2010) 1805. <https://doi.org/10.1002/9783527631940.ch57>
- Cochran L F, Patent US3360852A, 02 January 1968, *Manufacture of Ceramic Bases* (1968). <https://patents.google.com/patent/US3360852A/en>
- Glass S J, and Ewsuk K G, *MRS Bull* **22** (1997) 24. <https://doi.org/10.1557/S0883769400034709>
- Graham T, *Philos Trans R Soc Lond* **151** (1861) 183. <https://doi.org/10.1098/rstl.1861.0011>
- Ashley H E, *The Colloid Matter of Clay and Its Measurement*, USGS Numbered Series: 388, Washington, DC (1909).
- Alford N M, Birchall J D, and Kendall K, *Nature* **330** (1987) 51. <https://doi.org/10.1038/330051a0>
- Finke T, Gernsbeck M, Eisele U, Bockhorn H, Hartmann M, and Kureti S, *Ceram Forum Int* **84** (2007) 144.
- Johnson S B, Franks G V, Scales P J, Boger D V, and Healy T W, *Int J Miner Process* **58** (2000) 267. [https://doi.org/10.1016/S0301-7516\(99\)00041-1](https://doi.org/10.1016/S0301-7516(99)00041-1)
- Gibson I, Rosen D, Stucker B, and Khorasani M, in *Additive Manufacturing Technologies*, (eds) Gibson I, Rosen D, Stucker B, and Khorasani M, Springer, Cham (2021), p 379. [https://doi.org/10.1007/978-3-030-56127-7\\_14](https://doi.org/10.1007/978-3-030-56127-7_14)
- Sadat-Shojai M, and Moghaddas H, *J Appl Polym Sci* **137** (2020) 49810. <https://doi.org/10.1002/app.49810>
- Santhiya D, Nandini G, Subramanian S, Natarajan K A, and Malghan S G, *Colloids Surf Physicochem Eng Asp* **133** (1998) 157. [https://doi.org/10.1016/S0927-7757\(97\)00132-5](https://doi.org/10.1016/S0927-7757(97)00132-5)
- Nobles K P, Janorkar A V, and Williamson R S, *J Biomed Mater Res B Appl Biomater* **109** (2021) 1909. <https://doi.org/10.1002/jbm.b.34835>
- Aydemir C, Altay B N, and Akyol M, *Color Res Appl* **46** (2021) 489. <https://doi.org/10.1002/col.22579>
- Arulvel S, Mallikarjuna Reddy D, Dsilva Winfred Rufuss D, and Akinaga T, *Surf Interfaces* **27** (2021) 101449. <https://doi.org/10.1016/j.surf.2021.101449>
- Chen Z, Li Z, Li J, Liu C, Lao C, Fu Y, Liu C, Li Y, Wang P, and He Y, *J Eur Ceram Soc* **39** (2019) 661. <https://doi.org/10.1016/j.jeurceramsoc.2018.11.013>
- Chaudhary R P, Parameswaran C, Idrees M, Rasaki A S, Liu C, Chen Z, and Colombo P, *Prog Mater Sci* **128** (2022) 100969. <https://doi.org/10.1016/j.pmatsci.2022.100969>
- Ryan K R, Down M P, and Banks C E, *Chem Eng J* **403** (2021) 126162. <https://doi.org/10.1016/j.cej.2020.126162>
- Leong Y K, *Mater Des* **15** (1994) 141. [https://doi.org/10.1016/0261-3069\(94\)90113-9](https://doi.org/10.1016/0261-3069(94)90113-9)
- Israelachvili J N, *Intermolecular and Surface Forces*, Academic Press (2011).
- Barnes H A, Hutton J F, and Walters K, *An Introduction to Rheology*, Elsevier (1989).
- Ninham B W, and Nostro P L, *Molecular Forces and Self Assembly: In Colloid, Nano Sciences and Biology*, Cambridge University Press (2010).
- Hansen R S, and Smolders C A, *Colloid and Surface Chemistry in the Mainstream of Modern Chemistry*, ACS Publications (1962). <https://doi.org/10.1021/ed039p167>
- Derjaguin B, and Landau L, *Prog Surf Sci* **43** (1993) 30. [https://doi.org/10.1016/0079-6816\(93\)90013-L](https://doi.org/10.1016/0079-6816(93)90013-L)
- Schulz C, Dreizler A, Ebert V, and Wolfrum J, *Handbook of Experimental Fluid Mechanics*, Combust. Diagn. (2007).
- Sunthar P, in *Rheology of Complex Fluids*, (eds) Krishnan J M, Deshpande A P, and Kumar P B S, Springer, New York (2010), p 171. [https://doi.org/10.1007/978-1-4419-6494-6\\_8](https://doi.org/10.1007/978-1-4419-6494-6_8)
- Ramakrishnan V, Pradip, and Malghan S G, *J Am Ceram Soc* **79** (1996) 2567. <https://doi.org/10.1111/j.1151-2916-1996.tb09017.x>
- Ramakrishnan V, Pradip, and Malghan S G, *Colloids Surf Physicochem Eng Asp* **133** (1998) 135. [https://doi.org/10.1016/S0927-7757\(97\)00135-0](https://doi.org/10.1016/S0927-7757(97)00135-0)
- Ángyán J, Dobson J, Jansen G, and Gould T, *London Dispersion Forces in Molecules, Solids and Nano-structures: An Introduction to Physical Models and Computational Methods*, Royal Society of Chemistry (2020).
- Bergstrom L, in *Ceramic Transactions Vol. 51, Ceramic Processing Science and Technology* (eds) Hausner H, Messing G L, and Hirano S, American Ceramic Society, Westerville (1995), p 341. <https://www.osti.gov/biblio/99151>
- Bergstrom L, Shinozaki K, Tomiyama H, and Mizutani N, *J Am Ceram Soc* **80** (1997) 291. <https://doi.org/10.1111/j.1151-2916.1997.tb02829.x>
- Yanez J A, Shikata T, Lange F F, and Pearson D S, *J Am Ceram Soc* **79** (1996) 2917. <https://doi.org/10.1111/j.1151-2916.1996.tb08726.x>
- Leong Y K, Scales P J, Healy T W, and Boger D V, *Colloids Surf Physicochem Eng Asp* **95** (1995) 43. [https://doi.org/10.1016/0927-7757\(94\)03010-W](https://doi.org/10.1016/0927-7757(94)03010-W)
- Onogi S, and Matsumoto T, *Polym Eng Rev* **1** (1981) 45.
- Biggs S, and Healy T W, *J Chem Soc Faraday Trans* **90** (1994) 3415. <https://doi.org/10.1039/FT9949003415>
- Hunter R J, *Foundations of Colloid Science*, Oxford University Press (2001).
- Kontogeorgis G M, and Kiil S, *Introduction to Applied Colloid and Surface Chemistry*, Wiley (2016).
- Tadros T F, *Rheology of Dispersions: Principles and Applications*, Wiley (2011).
- Somasundaran P, and Runkana V, *Int J Miner Process* **72** (2003) 33. [https://doi.org/10.1016/S0301-7516\(03\)00086-3](https://doi.org/10.1016/S0301-7516(03)00086-3)

54. Stedman S J, Evans J R G, and Woodthorpe J, *J Mater Sci* **25** (1990) 1833. <https://doi.org/10.1007/BF01045394>
55. Doraiswamy D, *Rheol Bull* **71** (2002) 1.
56. Burns J L, Yan Y, Jameson G J, and Biggs S, *Colloids Surf Physicochem Eng Asp* **162** (2000) 265. [https://doi.org/10.1016/S0927-7757\(99\)00237-X](https://doi.org/10.1016/S0927-7757(99)00237-X)
57. Patel P D, and Russel W B, *J Rheol* **31** (1987) 599. <https://doi.org/10.1122/1.549938>
58. Yan Y, Burns J L, Jameson G J, and Biggs S, *Chem Eng J* **80** (2000) 23. [https://doi.org/10.1016/S1383-5866\(00\)00073-3](https://doi.org/10.1016/S1383-5866(00)00073-3)
59. Runkana V, Somasundaran P, and Kapur P C, *Chem Eng Sci* **61** (2006) 182. <https://doi.org/10.1016/j.ces.2005.01.046>
60. Moreno R, *Adv Appl Ceram* **111** (2012) 246. <https://doi.org/10.1179/1743676111Y.0000000075>
61. Pradip, Premachandran R S, and Malghan S G, *Bull Mater Sci* **17** (1994) 911. <https://doi.org/10.1007/BF02757568>
62. Barroso G, Li Q, Bordia R K, and Motz G, *J Mater Chem A* **7** (2019) 1936. <https://doi.org/10.1039/C8TA09054H>
63. Santhiya D, Subramanian S, Natarajan K A, and Malghan S G, *J Colloid Interface Sci* **216** (1999) 43. <https://doi.org/10.1006/jcis.1999.6289>
64. Saravanan L, and Subramanian S, *Colloids Surf Physicochem Eng Asp* **252** (2005) 175. <https://doi.org/10.1016/j.colsurfa.2004.10.104>
65. Dash S, Thakur V N, Kumar A, Mahaling R N, Patel S, Thomas R, Sahoo B, and Pradhan D K, *Ceram Int* **47** (2021) 33563. <https://doi.org/10.1016/j.ceramint.2021.08.265>
66. Zhao J, Yan C, Liu S, Zhang J, Li S, and Yan Y, *J Clean Prod* **268** (2020) 122329. <https://doi.org/10.1016/j.jclepro.2020.122329>
67. Ganesh I, Sundararajan G, Olhero S M, Torres P M C, and Ferreira J M F, *Ceram Int* **36** (2010) 1357.
68. Drdlikova K, Maca K, Slama M, and Drdlik D, *J Am Ceram Soc* **102** (2019) 7137.
69. Saravanan L, and Subramanian S, *Miner Eng* **98** (2016) 213. <https://doi.org/10.1016/j.mineng.2016.08.022>
70. Subbanna M, Kapur P C, and Pradip, *Ceram Int* **28** (2002) 401.
71. Subbanna M, Pradip, and Malghan S G, *Chem Eng Sci* **53** (1998) 3073. [https://doi.org/10.1016/S0009-2509\(98\)00158-4](https://doi.org/10.1016/S0009-2509(98)00158-4)
72. Sakka Y, Tang F, Fudouzi H, and Uchikoshi T, *Sci Technol Adv Mater* **6** (2005) 915.
73. Rayón E, Moreno R, Alcázar C, Salvador M D, Manjón F, Jiménez-Piqué E, and Lanes L, *J Am Ceram Soc* **96** (2013) 1070. <https://doi.org/10.1111/jace.12225>
74. Chin C H, Mughtar A, Azhari C H, Razali M, and Aboras M, *Ceram Int* **44** (15), (2018) 18641. <https://doi.org/10.1016/j.ceramint.2018.07.090>
75. Olhero S M, Ganesh I, Torres P M C, Alves F J, and Ferreira J M F, *J Am Ceram Soc* **92** (2009) 9. <https://doi.org/10.1111/j.15512916.2008.02823.x>
76. Navarro A, Alcock J R, and Whatmore R W, *J Eur Ceram Soc* **24** (2004) 1073. [https://doi.org/10.1016/S0955-2219\(03\)00460-6](https://doi.org/10.1016/S0955-2219(03)00460-6)
77. Pietrzak E, Wicinska P, Pawlikowska E, and Szafran M, *J Therm Anal Calorim* **130** (2017) 365. <https://doi.org/10.1007/s10973-017-6401-6>
78. Cho J-M, and Dogan F, *J Mater Sci* **36** (2001) 2397.
79. Tallon C, Chavara D, Gillen A, Riley D, Edwards L, Moricca S, and Franks G V, *J Am Ceram Soc* **96** (2013) 2374.
80. Liu G, Yan C, and Jin H, *Materials* (2022). <https://doi.org/10.3390/ma15082886>
81. Lewis J A, *J Am Ceram Soc* **83** (2000) 2341. <https://doi.org/10.1111/j.1151-2916.2000.tb01560.x>
82. Sigmund W M, Bell N S, and Bergström L, *J Am Ceram Soc* **83** (2000) 1557. <https://doi.org/10.1111/j.1151-2916.2000.tb01432.x>
83. Hench L L, *J Am Ceram Soc* **74** (1991) 1487. <https://doi.org/10.1111/j.1151-2916.1991.tb07132.x.107>
84. Pan J R, Huang C, Chen S, and Chung Y-C, *Colloids Surf Physicochem Eng Asp* **147** (1999) 359. [https://doi.org/10.1016/S0927-7757\(98\)00588-3](https://doi.org/10.1016/S0927-7757(98)00588-3)
85. Saravanan L, Subramanian S, Kumar A B V, and Tharanathan R N, *Ceram Int* **32** (2006) 637. <https://doi.org/10.1016/j.ceramint.2005.04.023>
86. Wicinska P, Zurawska A, Falkowski P, Jeong D-Y, and Szafran M, *J Korean Ceram Soc* **57** (2020) 231. <https://doi.org/10.1007/s43207-020-00036-x>
87. Mastalska-Popławska J, Sikora M, Izak P, and Góral Z, *J Sol-Gel Sci Technol* **96** (2020) 511. <https://doi.org/10.1007/s10971-020-05404-x>
88. Trichês E D, Dellú M, Pandolfelli V C, and Ortega F D, *Mater Sci Forum* **591–593** (2008) 498. <https://doi.org/10.4028/www.scientific.net/MSF.591-593.498>
89. Bednarek P, Szafran M, and Mizerski T, *Adv Sci Technol* **62** (2010) 169. <https://doi.org/10.4028/www.scientific.net/AST.62.169>
90. Alipal J, Mohd Pu'ad N A S, Lee T C, Nayan N H M, Sahari N, Basri H, Idris M I, and Abdullah H Z, *Mater Today Proc* **42** (2021) 240. <https://doi.org/10.1016/j.matpr.2020.12.922>
91. Pascual A R, and María E, *Cellulose*, Books on Demand (2019).
92. Yu O, and Kim K H, *Appl Sci* **10** (2020) 4626. <https://doi.org/10.3390/app10134626>
93. Gregorová E, Pabst W, and Boháček I, *J Eur Ceram Soc* **26** (2006) 1301. <https://doi.org/10.1016/j.jeurceramsoc.2005.02.015>
94. Nikpour P, Salimi-Kenari H, Fahimipour F, Rabiee S M, Imani M, Dashtimoghdam E, and Tayebi L, *Carbohydr Polym* **190** (2018) 281. <https://doi.org/10.1016/j.carbpol.2018.02.083>
95. Chaudhari V, Buttar H S, Bagwe-Parab S, Tuli H S, Vora A, and Kaur G, *Front Nutr* (2021). <https://doi.org/10.3389/fnut.2021.646988>
96. Singh R S, Saini G K, and Kennedy J F, *Carbohydr Polym* **73** (2008) 515. <https://doi.org/10.1016/j.carbpol.2008.01.003>
97. Sworn G, and Stouby L, in *Handbook of Hydrocolloids* (Third Edition), (eds) Phillips G O, and Williams P A, Woodhead Publishing Series in Food Science, Technology and Nutrition (2021), p 855. <https://doi.org/10.1016/B978-0-12820104-6.00009-7>
98. Song E-H, Shang J, and Ratner D M, *Polymer Science: A Comprehensive Reference*, (eds) Matyjaszewski K, and Möller M, Elsevier (2012), p 137. <https://doi.org/10.1016/B978-0-444-53349-4.00246-6>
99. Schmid J, Meyer V, and Sieber V, *Appl Microbiol Biotechnol* **91** (2011) 937. <https://doi.org/10.1007/s00253-011-3438-5>
100. Shishkovsky I, Sherbakov V, Ibatullin I, Volchkov V, and Volova L, *Compos Struct* **202** (2018) 651. <https://doi.org/10.1016/j.compstruct.2018.03.062>
101. Shishkovsky I, *New Trends in 3D Printing*, Books on Demand (2016).
102. Mazzoli A, *Med Biol Eng Comput* **51** (2013) 245. <https://doi.org/10.1007/s11517-012-1001-x>
103. Gibson I, Rosen D, Stucker B, and Khorsani M, *Additive Manufacturing Technologies*, Springer, Cham (2021). <https://doi.org/10.1007/978-3-030-56127-7>
104. Chen Z, Sun X, Shang Y, Xiong K, Xu Z, Guo R, Cai S, and Zheng C, *J Adv Ceram* **10** (2021) 195. <https://doi.org/10.1007/s40145-020-0444-z>
105. Sachs E, Cima M, Williams P, Brancazio D, and Cornie J, *J Eng Ind* **114** (1992) 481. <https://doi.org/10.1115/1.2900701>
106. Man Y, Ding G, Xudong L, Xue K, Qu D, and Xie Z, *J Asian Ceram Soc* **9** (2021) 1377. <https://doi.org/10.1080/21870764.2021.1981571>
107. Potdar A, Protasova L N, Thomassen L, and Kuhn S, *React Chem Eng* **2** (2017) 137. <https://doi.org/10.1039/C6RE00185H>
108. Juhasz J A, and Best S M, *J Mater Sci* **47** (2012) 610. <https://doi.org/10.1007/s10853-011-6063-x>

109. Fernandes H R, Gaddam A, Rebelo A, Brazete D, Stan G E, and Ferreira J M F, *Materials* (2018). <https://doi.org/10.3390/ma11122530>
110. Zhao S, Liu B, Ding Y, Zhang J, Wen Q, Ekberg C, and Zhang S, *J Clean Prod* **271** (2020) 122674. <https://doi.org/10.1016/j.jclepro.2020.122674>
111. Jones J R, in *Tissue Engineering Using Ceramics and Polymers*, (eds) Boccaccini A R, and Gough J E, Woodhead Publishing Series in Biomaterials, Woodhead Publishing (2007), p 52. <https://doi.org/10.1533/9781845693817.1.52>
112. Pradip, Rai B, and Sathish P, *KONA Powder Part J* **22** (2004) 151. <https://doi.org/10.14356/kona.2004018>
113. Gocmez H, *Ceram Int* **32** (2006) 521. <https://doi.org/10.1016/j.ceramint.2005.04.005>
114. Zhang S, Li Z, Yao Y, Tian L, and Yan Y, *Nano Energy* **100** (2022) 107476. <https://doi.org/10.1016/j.nanoen.2022.107476>
115. Rápó E, and Tonk S, *Molecules* (2021). <https://doi.org/10.3390/molecules26175419>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

## Socialization of the Importance of Building English Skills for Elementary School Children at SD Negeri 122395 Pematang Siantar

Vera Elisabet Siahaan<sup>1</sup> , Angel Nerin Patricia Panggabean<sup>2</sup> , Hanji Agustina Sitohang<sup>3</sup> , Theresi Theresi<sup>4</sup> , Santa Veronika Situmorang<sup>5</sup> , Fitrianti Manurung<sup>6</sup> , Herman Herman<sup>7</sup> , Yanti Kristina Sinaga<sup>8</sup> , Irene Adryani Nababan<sup>9</sup> , Prakash Puhka<sup>10</sup> 

<sup>1</sup>Universitas HKBP Nommensen Pematangsiantar, Pematang Siantar, Indonesia

<sup>2</sup>Universitas HKBP Nommensen Pematangsiantar, Pematang Siantar, Indonesia

<sup>3</sup>Universitas HKBP Nommensen Pematangsiantar, Pematang Siantar, Indonesia

<sup>4</sup>Universitas HKBP Nommensen Pematangsiantar, Pematang Siantar, Indonesia

<sup>5</sup>Universitas HKBP Nommensen Pematangsiantar, Pematang Siantar, Indonesia

<sup>6</sup>Universitas HKBP Nommensen Pematangsiantar, Pematang Siantar, Indonesia

<sup>7</sup>Universitas HKBP Nommensen Pematangsiantar, Pematang Siantar, Indonesia

<sup>8</sup>Universitas HKBP Nommensen Pematangsiantar, Pematang Siantar, Indonesia

<sup>9</sup>Universitas HKBP Nommensen Pematangsiantar, Pematang Siantar, Indonesia

<sup>10</sup>Delhi Technological University, India

### ABSTRACT

**Background.** The results of observing the abilities of students at this school, it can be seen that grade 5 students at SD Negeri 122395 Pematangsiantar have the basic skills to master the use of English well.

**Purpose.** This study aimed to investigate the importance of giving socialization to the students about basic skills that the students need to acquire in learning English well.

**Method.** This ability can be seen from the speed of students in understanding and responding to learning that has been socialized. Grade 5 students of SD Negeri 122395 Pematangsiantar consisting of 20 students are the subject of this observation. This socialization was conducted at SD Negeri 122395 Pematang Siantar. Three phases were implemented such as preparation, implementation and evaluation.

**Results.** The results showed that importance of building English skills starts from elementary school. Interesting learning methods can also build student activity and skills in the learning process that takes place in the classroom.

**Conclusion.** This is also based on the communication skills that students get in the learning process, not only limited to knowledge, students will be motivated and feel that what they have learned will be useful and used in the future..

**Citation:** Siahaan, V. E., Panggabean, A. N., P., Sitohang, H. A., Theresi, T., Situmorang, S. V., Manurung, F., Herman, Herman., Sinaga, Y. K., Nababan, I. A., & Puhka, P. (2023). Socialization of the Importance of Building English Skills for Elementary School Children at SD Negeri 122395 Pematang Siantar. *Pengabdian: Jurnal Abdimas*, 1(3), 155–163.

<https://doi.org/10.55849/abdimas.v1i3.327>

### Correspondence:

Herman Herman,  
herman@uhnnp.ac.id

**Received:** June 12, 2023

**Accepted:** July 15, 2023

**Published:** August 31, 2023



Vera Elisabet Siahaan, Angel Nerin Patricia Panggabean, Hanji Agustina Sitohang, Theresi Theresi, Santa Veronika Situmorang, Fitrianti Manurung, Herman Herman, Yanti Kristina Sinaga, Irene Adryani Nababan, Prakash Puhka



**KEYWORDS**

Basic Skills, English, Learning Method, Socialization

**INTRODUCTION**

English is the main language in most countries in the world, English is a universal language or is called an international language (Martina dkk., 2021; Oh dkk., 2023). Also, one of the most important languages to be learned or mastered by all people in the world. Many countries, especially former British colonies, consider English a second language that must be learned after the national language (Enríquez Raído & Cai, 2023; Hoque dkk., 2023; Tarp, 2021). In this increasingly advanced era of globalization, proficiency in English is an undeniable necessity. English has become an international language that is widely used in various sectors, including business communications, international trade, technology, education and tourism (Sohn dkk., 2023). Therefore, it is important for students in schools to have strong English skills in order to communicate effectively and successfully in an increasingly connected world.

In Indonesia, the government and education system have recognized the importance of building English skills from an early age (Crăciun & Bunoiiu, 2019; Fung dkk., 2023). The need for a workforce proficient in English is increasing along with the economic and industrial development in the country. In addition, success in facing global competition is also very dependent on good English skills (Pamuji & Limei, 2023). In Indonesia, English is introduced as one of the subjects in schools since the elementary education level. However, the acquisition of English language skills at school is often still limited to a basic understanding of grammar, vocabulary, and limited practice in speaking and writing (Jones, 2023; Stauss dkk., 2021). This creates a challenge for students in mastering English comprehensively and being able to use it in real-life contexts.

In facing this challenge, it is important for schools to build students' English skills with a holistic and contextual approach. The development of effective English skills involves a combination of good grammar and vocabulary, deep understanding of English language culture, fluent listening and speaking skills, and good writing skills (Al Awabdeh & Albashtawi, 2023; Faubl dkk., 2021). In addition, the application of innovative and interactive teaching methods is also needed to encourage student motivation and participation in learning English. Methods such as language games, group discussions, real-life simulations, and the use of technology can help students become actively involved and improve their ability to speak English. Apart from in the classroom, the widespread use of English in the school environment can also be an effective means of strengthening students' English skills (Deignan & Morton, 2022; Kildè, 2022). Extracurricular activities such as English clubs, debates, plays or student exchanges with schools abroad can provide practical opportunities for students to use English in real situations.

In order to build strong English skills in schools, it is also important to involve professional and well-trained teachers. A qualified English teacher can provide appropriate guidance, foster students' interest in learning English, and expose them to diverse resources, such as English literature, film and music (Blake dkk., 2019; Räisänen, 2020; "Video Component of Media Education in Direct and Reverse Acculturation at North Carolina State University and Texas Christian University," 2021). As such, building English skills in schools is an important step in supporting students' personal development and preparing them for an increasingly globally connected world. Through a holistic, innovative and interactive approach. There are many strategies for learning English that can no doubt arouse students' interest in the subject. For English teachers at SD/MI, this creates difficulties (Fadiyah dkk., 2023; Fiqih dkk., 2023; Ranal dkk., 2023). In order for students to participate actively in the learning process, teachers must continue to innovate.

Under 8% of Indonesians are proficient in using English accurately and precisely, which is still a relatively low percentage. Despite the fact that Indonesia has the potential to communicate effectively with this global language, English has helped this nation become more globally known (Auliani dkk., 2023; Mustafiyanti dkk., 2023; Wanti dkk., 2023). In addition, although this is not true, there is still an opinion that Indonesians who speak English are not considered nationalists. Because Indonesia is still developing, it must follow the rest of the world, making English a global language

## RESEARCH METHODOLOGY

This research was conducted so that students are more active in conveying or re-applying information that has been conveyed to them by the teacher, the role playing method emphasizes learning skills while playing can make students more creative, enthusiastic, and interested in the learning process (Silitonga et al., 2022). This technique is very much in demand by students because a game will make learning interesting. Games will foster enthusiasm and a sense of togetherness through fun learning.

There are several procedures carried out in carrying out this socialization activity as shown below:

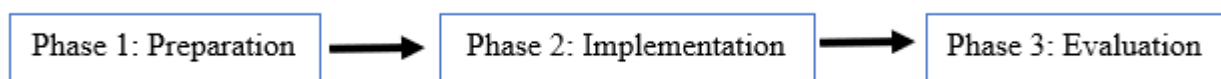


Figure 1. Procedures for socialization (Purba et al., 2022)

### Phase 1. Preparation

At this stage several activities were carried out such as:

- Conduct team discussions regarding the distribution of material to be delivered.
- Coordinate with the school to provide a place and time for socialization and determine which class will be socialized.
- Confirm the participants' readiness to take part in socialization activities.
- Prepare the media and tools needed for socialization activities.

### Phase 2: Implementation

This activity was attended by 20 fifth grade elementary school students and 1 English subject teacher as well as 6 presenters who will deliver socialization material on the topic of the importance of building English skills for elementary school children based on the lesson plan that has been prepared.



Figure 2. Participants participating in the socialization





Figure 3. Self-introduction by the Team

In this phase the activity is broken down into 2 stages, namely delivery of material or modules and a question and answer session.

### Phase 3: Evaluation

After the presenter explained the material regarding the importance of building English skills for elementary school children, then the presenter team invited the students to do ice breaking (singing, dancing, and playing games), after which the presenter team gave several questions to the students about the material that had been delivered (Mulyasari dkk., 2023; Noer dkk., 2023). After the students were able to answer these questions the presenting team gave rewards (books, pens, candy, and snacks), then the presenting team concluded the extent of their mastery of English..

## RESULT AND DISCUSSION

### 1. Phase 1.Preparation

- 1) Conducting discussions with the socialization implementation team regarding the activities to be carried out
- 2) Preparing materials and tools needed during socialization activities
- 3) Ensuring that all tools and materials are complete

### 2. Phase 2. Implementation

#### A. Submission of Material About the Importance of Building English Skills for Elementary School Children

The use of English is very important in keeping up with current technological advances. This is caused by several factors, one of which is English as an international language which demands that everyone master English.

- 1) The Concept and Importance of Learning English in Elementary Schools Students should be able to know how the concepts and the importance of the English language. This concept is a start for students to understand that English is a language that really needs to be mastered.
- 2) Strategies to Improve English Language Skills in Elementary Schools Students need to know several strategies that must be mastered to improve English language skills.
  - a) Listening to English music

This method can be tried because listening to music in English can make us speak English fluently. like listening to western songs playing on television, radio, Spotify or Joox song applications and so on (which smells of English).

b) Reading books and newspapers in English

Apart from listening, watching and regularly speaking English. So this method can improve our ability to speak English, such as reading books and newspapers in English. The more we get used to reading foreign language literature, it will help us master the foreign language and be able to hone our reading and pronunciation skills well.

3) Benefits and Purpose of Learning English, Some of the benefits obtained from learning English are:

a) Improving Language Skills

By improving English skills students will learn to understand grammar and sentence structure correctly.

b) Giving more Wider Insights

By mastering English students can find out information about the culture and life of other countries.

c) Increasing Career and Study Opportunities

By mastering English, children can increase career opportunities and study in the future.

d) Easy to Get along with and Adapt in New Environments

With this, students will more easily make friends and exchange information about the culture of their respective countries.

e) Helping to Prepare Yourself In The Future

English skills are important in the current era of globalization. By mastering English, students can prepare themselves to compete in the future.

Some of the objectives of learning English are:

a. Improving Communication Skills

Through learning English, we will be taught to communicate. In a different Language. Learn basic vocabulary (Vocabulary), simple phrases and sentences that allow them to talk about themselves, their siblings, family, friends, hobbies, and their surroundings.

b. Understanding of Culture and Diversity

As explained earlier that English is an international language by studying it, we will gain a better understanding of the culture, traditions and diversity that exist in other English-speaking countries. We can learn about cultural differences and develop tolerance and respect for diversity.

c. Preparing for Higher Education Levels

Mastery of English at the basic level helps all younger siblings build a strong foundation for language skills at the higher education level. Good English language skills will support them in understanding the material provided by teachers in secondary schools to tertiary institutions which often use English as a medium of delivery.

d. Global Skills

As an international language in many fields such as business, technology, science, and tourism. By learning English from an early age, you will have an advantage in global competition in the future and have wider opportunities in your career.

e. Development of cognitive skills

Learning English can help improve our cognitive skills.

## B. Question and Answer Session

There are findings that are illustrated after the question and answer session, namely the lack of students' interest in learning English is because the participants think that English is difficult and unpleasant (Al Maarif dkk., 2023; Hermansyah dkk., 2023; Utami dkk., 2023). Therefore, we use the role playing method in our socialization to increase students' interest in learning English.

## 3. Phase 3. Evaluation

Evaluation at this stage is obtained from presentations and question and answer sessions.

### a. Evaluation of the Presentation Session

Some of the students were less enthusiastic in listening to the material presented because some of the students were less focused, some were busy playing with their classmates, some were sleepy.

### b. Evaluation of the Question and Answer Session

When giving questions there are students who are active in conveying questions to the presenter. There are also students who actually want to ask/answer but they lack the confidence to submit questions and answers.



Figure 4. Question and Answer Session

## 4. Challenges in the Implementation of Socialization

There were several challenges faced by the team when carrying out socialization at SDN 122395 namely the lack of adequate facilities so that a media was not delivered in the form of a video, besides that the atmosphere outside the class was not conducive due to the class carrying out Physical Education (Sports) subjects

## CONCLUSION

Based on the results of the socialization about the importance of building English language skills which were carried out for fifth grade students of public elementary schools. The existence of English is actually very important for students. These things can be seen from the response that was seen during the presentation of the material carried out, the students were very interested in the material presented. During the question and answer session students could understand the questions given so they were very interested in the questions that were conveyed.

## ACKNOWLEDGEMENT

The researchers, students from Department of Economics Education, would like to express our deepest gratitude to the principal and deputy principal, as well as the teaching staff of SD Negeri 12239, especially to the English teacher who has given time and who has given permission to the presenter team to carry out this outreach activity. We also thank the class V students who participated and gave their time for us to come and support this socialization activity. The researchers also say thank you to Dr. Herman S.Pd., M.Pd. who have equipped the team and provided directions and suggestions in carrying out this socialization activity as an output of the English General Course assignment in support of the Merdeka Campus

## AUTHORS' CONTRIBUTION

Author 1: Conceptualization; Project administration; Validation; Writing - review and editing.

Author 2: Conceptualization; Data curation; In-vestigation.

Author 3: Data curation; Investigation.

Author 4: Formal analysis; Methodology; Writing - original draft.

Author 5: Supervision; Resources, Visualization.

Author 6: Validation

Author 7-10: Other contribution, Review and Editing.

## REFERENCES

- Al Awabdeh, A., & Albashtawi, A. H. (2023). Predictable Factors that Help Students Engage in Online EFL Classroom and their Relationship to Self-Management. *International Journal of Emerging Technologies in Learning (iJET)*, 18(01), 4–18. <https://doi.org/10.3991/ijet.v18i01.35257>
- Al Maarif, M. F., Afifah, R. A. N., Choirunnisa, A., Jannah, A. M., Zanuvar, M. Y., Saddhono, K., & Yingxiang, S. (2023). Integrating and Strengthening National Vision in the Community as an Effort to Prevent Radicalization and Foster Love for the Motherland. *Pengabdian: Jurnal Abdimas*, 1(1), 20–29. <https://doi.org/10.55849/abdimas.v1i1.151>
- Auliani, R., Suprawihadi, R., & Avinash, B. (2023). Application of Appropriate Technology for Clean Water. *Pengabdian: Jurnal Abdimas*, 1(1), 30–39. <https://doi.org/10.55849/abdimas.v1i1.152>
- Blake, H. L., Bennetts Kneebone, L., & McLeod, S. (2019). The impact of oral English proficiency on humanitarian migrants' experiences of settling in Australia. *International Journal of Bilingual Education and Bilingualism*, 22(6), 689–705. <https://doi.org/10.1080/13670050.2017.1294557>
- Crăciun, D., & Bunoii, M. (2019). *Learning science outside the classroom: A summer school experience*. 050002. <https://doi.org/10.1063/1.5090086>
- Deignan, T., & Morton, T. (2022). The challenges of English medium instruction for subject lecturers: A shared viewpoint. *ELT Journal*, 76(2), 208–217. <https://doi.org/10.1093/elt/ccab084>
- Enríquez Raído, V., & Cai, Y. (2023). Changes in web search query behavior of English-to-Chinese translation trainees. *Ampersand*, 11, 100137. <https://doi.org/10.1016/j.amper.2023.100137>
- Fadiyah, F., Fuadi, A., Nurjannah, N., Irmayanti, I., & Lita, W. (2023). Quizizz Application-Based Interactive Learning Media Development Workshop for Junior High School Teacher. *Pengabdian: Jurnal Abdimas*, 1(2), 59–65. <https://doi.org/10.55849/abdimas.v1i2.157>
- Faubl, N., Pótó, Z., Marek, E., Birkás, B., Füzesi, Z., & Németh, T. (2021). Kulturális különbségek elfogadása a külföldi orvostanhallgatók beilleszkedésében. *Orvosi Hetilap*, 162(25), 978–987. <https://doi.org/10.1556/650.2021.32104>

- Fiqih, M., Thaha, A., Shidiq, S., Nafis, Moch. A., & Martin, W. (2023). The Concept of Internal Quality Assurance in Madrasah Diniyah PP. Al-Hidayah Tanggulangin Sidoarjo. *Pengabdian: Jurnal Abdimas*, 1(1), 40–45. <https://doi.org/10.55849/abdimas.v1i1.150>
- Fung, H. W., Lam, S. K. K., Chien, W. T., Hung, S. L., Ling, H. W.-H., Lee, V. W. P., & Wang, E. K. (2023). Interpersonal stress mediates the relationship between childhood trauma and depressive symptoms: Findings from two culturally different samples. *Australian & New Zealand Journal of Psychiatry*, 57(7), 1052–1061. <https://doi.org/10.1177/00048674221138501>
- Hermansyah, S., Nasmilah, N., Pammu, A., Saleh, N. J., Huazheng, H., & Congzhao, H. (2023). Socialization Making Media Learning Interactive E-Module based Flippbook in Elementary School 4 Maiwa. *Pengabdian: Jurnal Abdimas*, 1(1), 1–7. <https://doi.org/10.55849/abdimas.v1i1.117>
- Hoque, M. A., Ahmad, T., Manzur, S., & Prova, T. K. (2023). Community-Based Research in Fragile Contexts: Reflections From Rohingya Refugee Camps in Cox's Bazar, Bangladesh. *Journal on Migration and Human Security*, 11(1), 89–98. <https://doi.org/10.1177/23315024231160153>
- Jones, C. (2023). Jigsaw Migration: How Mixed Citizenship LGBTQ Families (Re)Assemble Their Fragmented Citizenship. *International Migration Review*, 019791832311751. <https://doi.org/10.1177/01979183231175101>
- Kildè, L. (2022). The enhancement of self-directedness in the studies of English books for specific purposes: An analysis of ESP study books. *Journal of Education Culture and Society*, 13(1), 189–200. <https://doi.org/10.15503/jecs2022.1.189.200>
- Martina, D., Lin, C.-P., Kristanti, M. S., Bramer, W. M., Mori, M., Korfage, I. J., Van Der Heide, A., Van Der Rijt, C. C. D., & Rietjens, J. A. C. (2021). Advance Care Planning in Asia: A Systematic Narrative Review of Healthcare Professionals' Knowledge, Attitude, and Experience. *Journal of the American Medical Directors Association*, 22(2), 349.e1-349.e28. <https://doi.org/10.1016/j.jamda.2020.12.018>
- Mulyasari, D., Noer, R. M., Sari, N., Ermawaty, E., Triharyadi, F., Tampubolon, D., & Catherine, S. (2023). Improving Health Status in The Elderly Through Health Checks and Education at Nuriah Nursing Homes in Karimun. *Pengabdian: Jurnal Abdimas*, 1(2), 75–81. <https://doi.org/10.55849/abdimas.v1i2.183>
- Mustafiyanti, M., Putri, M. P., Muyassaroh, M., Noviani, D., & Dylan, M. (2023). A Form of Independent Curriculum, an Overview of Independent Learning at State Elementary School 05 Gelumbang Muaraenim. *Pengabdian: Jurnal Abdimas*, 1(2), 82–96. <https://doi.org/10.55849/abdimas.v1i2.185>
- Noer, R. M., Silalahi, A. D., Mulyasari, D., Sari, N., Ermawaty, E., Triharyadi, F., Tampubolon, D., & Bevoor, B. (2023). Improving the Degree of Health in the Elderly Through Health Checks and Education. *Pengabdian: Jurnal Abdimas*, 1(1), 8–13. <https://doi.org/10.55849/abdimas.v1i1.139>
- Oh, J. H. J., Basma, B., Bertone, A., & Luk, G. (2023). Assessments of English Reading and Language Comprehension in Bilingual Children: A Systematic Review 2010 to 2021. *Canadian Journal of School Psychology*, 08295735231183608. <https://doi.org/10.1177/08295735231183608>
- Pamuji, S., & Limei, S. (2023). The Managerial Competence Of The Madrasa Head In Improving Teacher Professionalism And Performance At Mi Al-Maarif Bojongsari, Cilacap District. *Pengabdian: Jurnal Abdimas*, 1(2), 66–74. <https://doi.org/10.55849/abdimas.v1i2.158>
- Räisänen, T. (2020). The Use of Multimodal Resources by Technical Managers and Their Peers in Meetings Using English as the Business Lingua Franca. *IEEE Transactions on Professional Communication*, 63(2), 172–187. <https://doi.org/10.1109/TPC.2020.2988759>



Ranal, A., Husniyah, H., Fienti, Y., Putri, S. A., Lenin, F., Musrika, M., Diana, D., & Xin, D. (2023). Physical Activity Training Education for the Elderly at Nursing Homes. *Pengabdian: Jurnal Abdimas*, 1(1), 14–19. <https://doi.org/10.55849/abdimas.v1i1.143>

Sohn, J., Park, I., Lee, G., & Choi, S. (2023). Exploring police legitimacy and other factors in predicting cooperation with police in the Atlanta Korean American community. *Policing: An International Journal*. <https://doi.org/10.1108/PIJPSM-02-2023-0032>

Stauss, K., Koh, E., Johnson-Carter, C., & Gonzales-Worthen, D. (2021). OneCommunity Reads: A Model for Latino Parent-Community Engagement and Its Effect on Grade-Level Reading Proficiency. *Education and Urban Society*, 53(4), 402–424. <https://doi.org/10.1177/0013124520928612>

Tarp, G. (2021). Building Dialogue Between Cultures: Expats' Way of Coping in a Foreign Country and Their Willingness to Communicate in a Foreign Language. Dalam N. Zarrinabadi & M. Pawlak (Ed.), *New Perspectives on Willingness to Communicate in a Second Language* (hlm. 55–84). Springer International Publishing. [https://doi.org/10.1007/978-3-030-67634-6\\_4](https://doi.org/10.1007/978-3-030-67634-6_4)

Utami, L. D., Amin, M., Mustafiyanti, M., & Alon, F. (2023). Masjid Friendly: Mosque Based Economic Empowerment. *Pengabdian: Jurnal Abdimas*, 1(2), 97–106. <https://doi.org/10.55849/abdimas.v1i2.186>

Video Component of Media Education in Direct and Reverse Acculturation at North Carolina State University and Texas Christian University. (2021). *International Journal of Media and Information Literacy*, 6(2). <https://doi.org/10.13187/ijmil.2021.2.426>

Wanti, L. P., Romadloni, A., Somantri, O., Sari, L., Prasetya, N. W. A., & Johanna, A. (2023). English Learning Assistance Using Interactive Media for Children with Special Needs to Improve Growth and Development. *Pengabdian: Jurnal Abdimas*, 1(2), 46–58. <https://doi.org/10.55849/abdimas.v1i2.155>

---

**Copyright Holder :**

© Herman Herman et al . (2023).

**First Publication Right :**

© Pengabdian: Jurnal Abdimas

**This article is under:**







# Solar light and ultrasound-assisted rapid Fenton's oxidation of 2,4,6-trichlorophenol: comparison, optimisation, and mineralisation

Shivani Yadav<sup>1</sup> · Sunil Kumar<sup>2</sup> · Anil Kumar Haritash<sup>1</sup>

Received: 28 June 2023 / Accepted: 2 September 2023

© The Author(s), under exclusive licence to Accademia Nazionale dei Lincei 2023

## Abstract

Chlorophenols are the persistent organic contaminants released in the aquatic bodies by industrial manufacturing units. Treatment of phenolic wastewater is an arduous process for conventional treatment methods because of their high stability and complexity. The present study deals with degradation of 2,4,6-Trichlorophenol, using Photo-Fenton's process. The study commences the optimisation and validation of different regulating parameters like pH, oxidant ( $\text{H}_2\text{O}_2$ ), and  $\text{Fe}^{2+}$  ions at variable concentration in batch mode. At pH 3.0,  $\text{Fe}^{2+}$  0.5 mM, and  $\text{H}_2\text{O}_2$  (10.0 mM), complete degradation of trichlorophenol was observed within 6 min of reaction time. The mineralisation of the model pollutant was studied over TOC analyser and HPLC. Response Surface Plots were drawn to define the interactive relationship between process governing Fenton's process. Under optimized conditions, different Fenton-integrated processes, such as Solar-Fenton, Sono-Fenton, and Sono-Photo-Fenton ( $\text{UV}_{365}$ ), were compared for degradation of TCP. Among all the processes, Solar-Fenton resulted in rapid and complete degradation of TCP within 5 min. The mineralisation efficiency of Fenton's process, Solar-Fenton, Sono-Fenton, Sono-Photo-Fenton, and Photo-Fenton's processes were 50%, 98%, 90%, 75%, and 50%, respectively. The results indicated Solar-assisted Fenton as potential and efficient approach toward degradation of Trichlorophenol.

**Keywords** Wastewater · Trichlorophenol · Photo-Fenton · Fenton-integrated processes · AOPs

## Abbreviations

AOPs	Advanced oxidation processes
BBD	Box–Behnken design
CPs	Chlorophenols
HPLC	High-performance liquid chromatography
POP	Persistent organic pollutants
ROS	Reactive oxygen species
RSM	Response surface methodology
TCP	2,4,6-Trichlorophenol
TOC	Total organic carbon
USEPA	United States Environmental Protection Agency

## 1 Introduction

The prevalence of Chlorophenols (CPs) is continuously detected in various environmental matrices like water, soil, sediments, air, food commodities, and biological tissues, as well. These refractory organic contaminants are released in the natural environment via industrial manufacturing units related to pharmaceuticals, agricultural products, wood preservatives, paper, etc. (Guo et al. 2017). CPs are formed as reaction by-product during disinfection process of drinking water, incomplete combustion or incineration of organic waste or burning of chlorophenol-treated wood as well (Vlastos et al. 2016). The pollutants are likely to enter the aquatic ecosystem via soil erosion, leaching, agricultural run-off, atmospheric deposition, and volatilization (Czaplicka 2004). High solubility of CPs in water results in entry into the food chain and their lipophilic property allows passage through the lipid bilayer membrane of organisms, thus leading to accumulation (Yadav et al. 2023). Owing to the bioaccumulation, the organochlorines bear the potential of acute and chronic toxicity to diverse class of organisms, thus casting a signal of health concern for the livings (Michałowicz 2010). CPs belongs to the class of persistent

Technical Paper Submitted to: Rendiconti Lincei. Scienze Fisiche e Naturali.

✉ Shivani Yadav  
shivaniyadav\_phd2k19@dtu.ac.in

- <sup>1</sup> Environmental Microbiology and Bioremediation Laboratory, Department of Environmental Engineering, Delhi Technological University, Delhi 110042, India
- <sup>2</sup> Solaris Chemtech Industries, Bhuj, Gujarat 370001, India

organic pollutants (POPs), possess low biodegradability, high toxicity, carcinogenic, mutagenic, and teratogenic characteristics, and are categorized as priority toxic compounds by USEPA (ATSDR 2007). The International Agency for Research on Cancer has also classified them as Group 2B probable carcinogens (IARC 2004). Literature has reported the worldwide production of CPs exceeding 200,000 tons per year (Ge et al. 2017). Although the occurrence of these emerging pollutants is in trace amounts in the natural environment, usually in ng/l to µg/l, they are reportedly found in the concentration up to 190 mg/l in polluted surface and ground water (Olaniran and Igbinsola 2011). Considering the harmful effect over the aquatic environment, emphasis is given on managing the water quality and subsequently removing the CPs from aquatic environment. The removal of these chloroderivatives of phenols depends on the position and number of chlorine atoms on the benzene ring (Yadav et al. 2023). The presence of chlorine atoms imposes steric hindrance, inductive effect, and resonance effect (Xu and Wang 2013). The degree and extent of ionization decreases with increase in the count of chlorine atoms. This is in proposition with the high pKa values that multichlorophenols (dichlorophenols, trichlorophenols, tetrachlorophenols, and pentachlorophenols) strongly interact with oxide surfaces than monochlorophenols through chemical bonding resulting in accumulation in soils and sediments (Czaplicka 2004). Because of the chemical and structural stability, traditional methods cannot break the C–Cl bonds associated with aromatic structures; biological treatment is inefficient in treating these harmful organics when present in lower concentration, whereas high levels of CPs inhibit the growth of microbial community (Czaplicka 2004). The metabolites generated from their partial degradation like chlorocatechols or chlorine substituted fission products suppress the development of halo aromatic-utilizing microorganisms (Annachhatre and Gheewala 1996). On the other hand, the advancement of Advanced Oxidation Processes (AOPs) has indeed received the attention of scientific world as promising wastewater treatment methodologies. AOPs have high removal efficiency, high rate of reaction, less generation of waste (sludge), and application to diverse spectrum of pollutants like pharmaceuticals, textile dyes, pesticides, landfill leachate, etc. (Sharma et al. 2016; Pipil et al. 2022; Bilal et al. 2018). These processes rely on the *in-situ* generation of highly reactive and oxidizing radical species (ROS). The AOPs are classified on the basis of formation of different oxidation radicals, such as HO·, HO<sub>2</sub>·, O·, sulfate radicals, etc. (Yadav et al. 2023). These powerful ROS interact with the organic molecules and induce progressive oxidation, transforming the parent compounds into more innocuous and biodegradable intermediate product (Yadav et al. 2022). The treatment process is non-selective in nature, thus can easily react with broad category of contaminants. AOPs

has been found effective toward treating various classes of contaminants, including micropollutants, endocrine-disrupting chemicals (EDCs), persistent organic pollutants (POPs), and antibiotics like amoxicillin (Verma and Haritash 2020; Yadav et al. 2023). The treated effluent can be efficiently reused in various industrial and domestic processes. The process reduces the toxicity levels, simultaneously increasing the biodegradability of pollutants, further bringing down to permissible limits (Yadav et al. 2022). The current study deals with treatment of 2,4,6-Trichlorophenol (TCP) using Fenton's process coupled with solar energy and ultrasound; comparison of efficiency toward degradation, and optimisation of process with maximum degradation and mineralisation.

## 2 Materials and methods

### 2.1 Chemicals

Analytical grade 2,4,6-TCP (purity 98%) was procured from Thermo Fisher Scientific (USA), FeSO<sub>4</sub>·7H<sub>2</sub>O as source of Fe (II), Hydrogen peroxide H<sub>2</sub>O<sub>2</sub> (30% w/v) which was used as an oxidant was purchased from Central Drug House (CDH), India; and HPLC grade methanol was obtained from Merck & Co (India). For pH adjustment, NaOH and H<sub>2</sub>SO<sub>4</sub> were procured from Central Drug House (CDH), India. Sodium acetate, hydroxylammonium chloride, and 1,10-phenanthroline were purchased from SRL Chemicals, India.

## 3 Experimental procedure

Experimentation work was performed based on Fenton, Photo (UV<sub>365</sub>) Fenton, Solar-Fenton, Sono-Fenton, and Sono-Photo-Fenton. The optimization of key operating parameters like pH, oxidant dose, and iron concentration against degradation of 2,4,6-TCP was studied. Synthetically prepared stock solution of 2,4,6-TCP of strength 100 mg/l was prepared using ultrapure Type-I water and was stored in dark. The Photo-Fenton's treatment for 2,4,6-TCP was performed by varying the concentration of Fe(II) and H<sub>2</sub>O<sub>2</sub> at different pH according to the runs provided by Response Surface Methodology (RSM)-based Box–Behnken Design. The concentration of 2,4,6-TCP was kept constant (100 mg/l) and optimized conditions were noted. TCP degradation was examined using varying concentrations of Fe(II) (0.1 mM to 0.5 mM) and the oxidant H<sub>2</sub>O<sub>2</sub> (10.0 mM to 60.0 mM). The research was carried out in a glass beaker with 200 ml of a synthetically prepared stock solution of 2,4,6-TCP (100 mg/l). The reaction solution was subjected to continuous magnetic shaking at 200 rpm under continuous

irradiation of UV light ( $\lambda$ -365 nm). The Photo-Fenton treatments were executed in a fabricated UV chamber integrated with eight UV tubes (Phillips 36 W each) of wavelength 365 nm having aerial arrangement, positioned parallel to each other (Fig. 1). The combined source intensity of UV chamber as calculated by Verma and Haritash (2019) was  $672 \text{ W/m}^2$ . A distance of 10.0 cm was maintained between UV source and sample for effective UV penetration. An aliquot of 5.0 ml was extracted using pre-rinsed syringe for analysis at regular intervals of 1 min. The above procedure was repeated until maximum/complete degradation was achieved or the residual concentration of TCP got stabilized. All the experiments were performed in triplicates to assure the reliability of analysis.

### 3.1 Analysis of 2,4,6-TCP

A double-beam UV–Vis spectrophotometer (Lab India manufacture UV 3092 model) was used to determine degradation of TCP. The absorption spectra of 2,4,6-TCP were detected in the wavelength range of 190 nm–800 nm and the wavelength of maximum absorption ( $\lambda_{\text{max}}$ ) was observed at 293 nm, which was also reported in the studies (Pandit et al. 2001; Gaya et al. 2010; Zhu et al. 2021). The calibration graph for the model pollutant was plotted at regular intervals of 10 mg/l in a concentration range from 0.0 mg/l to 100.0 mg/l. The following equation (Eq. 1) was used to calculate the removal efficiency of the processes:

$$\text{Efficiency (\%)} = \left[ \frac{C_i - C_t}{C_i} \right] \times 100, \quad (1)$$

where  $C_i$  is the initial concentration of TCP; and  $C_t$  is the concentration of TCP at time ' $t$ '.

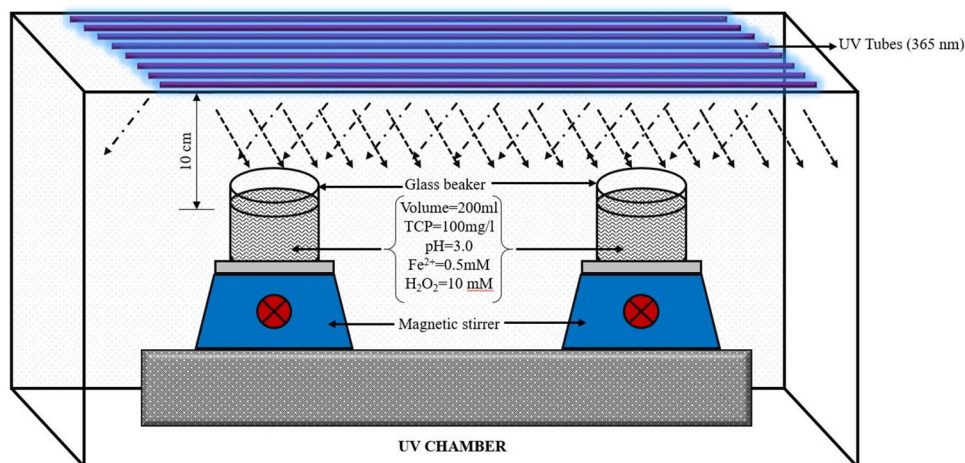
The residual reaction concentration during the oxidation of TCP was also ascertained using the chromatographic

analysis. The HPLC system (Shimadzu, LC-20AD) equipped with UV/Vis detector (SPD-20A) and C-18 column (Inertsil® ODS-3V,  $5 \mu\text{m}$   $4.6 \times 250 \text{ mm}$ ) was used for the analysis. A standard curve of model chemical (TCP) was plotted at 10 mg/l intervals in the concentration range of 0.0–100.0 mg/l over the HPLC system. The mobile phase of methanol and water (70:30 v/v) was used in binary mode at flow rate of 1.5 ml/min for separation by HPLC, and the response was recorded at 290 nm. Mineralisation of TCP was confirmed over TOC Analyser (TOC-L Shimadzu Make, Japan) as well to validate the observations over UV–Vis spectrophotometer and HPLC. The residual concentration of ferrous ion and total dissolved iron was determined using spectrophotometer at 510 nm using 1,10-phenanthroline method as mentioned in APHA (APHA, Method 3500-Fe 1997). Residual concentration of  $\text{H}_2\text{O}_2$  in the solution was determined through iodimetric titration.

### 3.2 Design of experiment and statistical analysis

RSM, a multivariate statistical tool widely used for designing of experiment and optimization of process parameters, was used for process optimisation in the present study. RSM based on Box–Behnken Design (BBD) using JSM Design of Expert ver. 16.2.0 software was used to obtain optimized conditions toward degradation of TCP. The BBD is a three-level fractional design with a central point. Three independent operating variable factors, (A) pH (1.0–5.0), (B) ferrous ion concentration (0.1 mM–0.5 mM), and (C) oxidant dose (10.0 mM–60.0 mM), led to a total of 15 experiments (Table 1). A second-order polynomial equation as shown in

**Fig. 1** Schematic presentation of Photo-Fenton setup



**Table 1** Box–Behnken design for three independent variables, and predicted and observed responses for TCP degradation

Run	pH	Fe <sup>2+</sup>	H <sub>2</sub> O <sub>2</sub>	Removal efficiency (%)	
				Observed	Predicted
1	3	0.3	35	75	75
2	3	0.1	10	25	25
3	3	0.3	35	75	75
4	5	0.5	35	71	78
5	1	0.1	35	18	11
6	3	0.3	35	75	75
7	5	0.3	60	64	57
8	5	0.1	35	11	10
9	3	0.5	60	55	55
10	3	0.5	10	100	92
11	1	0.3	60	20	19
12	1	0.3	10	24	31
13	3	0.1	60	45	53
14	1	0.5	35	12	13
15	5	0.3	10	54	55

Eq. 2 was used to determine the mathematical relationship between independent variables (A, B, C) and response (Y)

$$Y = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_{11} A^2 + \beta_{22} B^2 + \beta_{33} C^2 + \beta_{12} AB + \beta_{13} AC + \beta_{23} BC. \quad (2)$$

To test the adequacy of the model, analysis of variance (ANOVA) was calculated with predictability at 95% confidence level. The interactive and individual response of independent variables against TCP degradation is illustrated using respective surface plots.

## 4 Results and discussion

### 4.1 Degradation of 2,4,6-TCP

The effect of different parameters was studied by varying the concentration of regulating factors based on the runs given by JMP Design of Experiment (version 16.2) software. Keeping the TCP concentration constant (100 mg/l), the effects of factors like pH, Fe<sup>2+</sup>, and H<sub>2</sub>O<sub>2</sub> were examined in the range of 1.0–5.0, 0.1–0.5 mM, and 10.0–60.0 mM respectively (Table 1) for degradation of TCP. Addition of oxidant, H<sub>2</sub>O<sub>2</sub> imparted dark brown color to the reaction mixture having iron and TCP and the solution became turbid within few seconds. As the reaction proceeded, the dark color starts fading changing to light yellow and colorless at the end indicating completion of the reaction. Under optimum doses of pH 3.0 ± 0.2, Fe<sup>2+</sup> 0.5 mM, and H<sub>2</sub>O<sub>2</sub>

10.0 mM, TCP disappeared to non-detectable levels within 6 min of the reaction time during the process of Fenton's oxidation.

### 4.2 Effect of pH

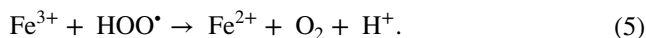
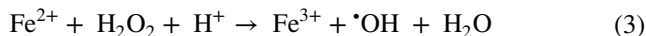
The pH of solution determines the removal efficiency of organic pollutants in the Photo-Fenton system. It is evident from literature that Fenton's process works efficiently under acidic condition (pH = 3 optimum) (Verma and Haritash 2019; Sharma et al. 2016). The present study was carried out at pH 1.0, 3.0, and 5.0 in which at pH 3.0, complete (100%) degradation of TCP was observed within 6–7 min of the reaction time. This is attributed because of the stability of Fe<sup>2+</sup> and H<sub>2</sub>O<sub>2</sub> which is more at pH 3.0. The results are in strong agreement with observation in the studies carried out by Karci et al. (2012) and Kavitha and Palanivelu (2016). However, at pH 1.0, only 10%–25% of the removal was achieved which took twice the time as taken by the degradation process at pH 3.0. This is due to the formation of complex iron species and oxonium (H<sub>3</sub>O<sub>2</sub><sup>+</sup>), while H<sup>+</sup> ions scavenge the hydroxyl radicals (Pipil et al. 2022). The solubility of Fe (II) is low under acidic pH. Moreover, at pH 5.0, 60%–70% degradation of TCP was observed and the time taken for degradation was twofold as compared to degradation at pH 3.0. Under alkaline condition, the solubility of Fe(III)/Fe(II) decreases, and iron precipitates as Fe(OH)<sub>3</sub>/Fe(OH)<sub>2</sub> sludge, while H<sub>2</sub>O<sub>2</sub> dissociates to form water and oxygen. This results in inactivation and inhibition of further generation of ·OH radicals (Pipil et al. 2022). Also, the electrostatic force of repulsion between CP and catalyst surface (Fe acts as catalyst in Fenton's process) dominates under alkaline conditions reducing the dissociation of CPs (Kantar et al. 2019). Furthermore, the dissociation constant (pK<sub>a</sub>) for CPs is in the range from 4.7 for PCP to 9.41 for 4-CP and 6.23 for 2,4,6-TCP, suggest the ionization of polychlorophenols to begin at lower pH. Thus, in the present study, complete degradation of TCP was observed at optimum pH of 3.0.

### 4.3 Effect of Fe<sup>2+</sup>

Fe<sup>2+</sup> is a significant factor determining the removal rate of the recalcitrant pollutants in the Fenton's treatment (Pipil et al. 2022). In the present study, experiments were performed with iron concentrations of 0.1 mM, 0.3 mM, and 0.5 mM. It was observed that at iron concentration 0.5 mM and acidic conditions (pH = 3.0) and oxidant concentration 10.0 mM, rapid and complete degradation (100%) of TCP took place in 6 min. However, at lower concentration 0.1 mM and 0.3 mM, the maximum degradation observed was 45% (oxidant dose = 60.0 mM) and 75% (oxidant



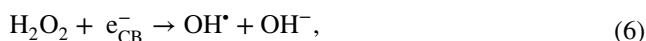
dose = 35.0 mM), respectively in 6 min. Redox recycling process of  $\text{Fe}^{2+}/\text{Fe}^{3+}$  takes place in the bulk solution generating reactive oxidative species (Eq. 3), while some ferrous ion species are generated through reaction (4) and (5)



Similar results were observed by Kavitha and Palanivelu (2004, 2016) with iron concentration of 0.4 mM and 0.8 mM toward degradation of phenol and TCP, respectively. The OH radicals generated from the reaction between iron and  $\text{H}_2\text{O}_2$  attack the  $\pi$  system of the aromatic phenolic ring accelerating the oxidation of TCP (Kavitha and Palanivelu 2004). The residual concentration of total dissolved iron and ferrous ions determined at the end of the experiment reported complete consumption of ferrous ions, whereas total dissolved iron concentration measured was 0.03 mM.

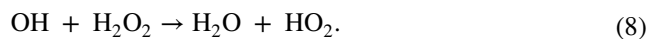
#### 4.4 Effect of $\text{H}_2\text{O}_2$

The effect of  $\text{H}_2\text{O}_2$  was studied at concentration of 10.0 mM, 35.0 mM, and 60.0 mM. Maximum removal efficiency (complete degradation, 100%) of TCP was observed with  $\text{H}_2\text{O}_2$  dose of 10.0 mM (pH = 3.0;  $\text{Fe}^{2+}$  = 0.5 mM) within 6 min of treatment time. On increasing the  $\text{H}_2\text{O}_2$  concentration, no significant improvement toward the degradation of TCP was observed. At maximum concentration of oxidant, i.e., 60.0 mM, only 55% degradation of TCP was achieved in 7–8 min of reaction time. Peroxide acts as electron acceptor and inhibits the recombination of electron-hole thus facilitates the generation of OH radicals (Eq. 6)



where  $e_{\text{CB}}^-$  refers to the presence of electrons in the conduction band.

$\text{H}_2\text{O}_2$  when present in excess amount acts as scavenger and self-destruction of peroxide occurs (Eq. 7). The hydroperoxyl radicals are produced from reaction between OH radicals and excess peroxide (Eq. 8). The reduction potential of these hydroperoxyl radicals (1.0 V) is less than that of hydroxyl radicals (2.8 V) thus is the most probable reason for decrease in the degradation rate with increase in  $\text{H}_2\text{O}_2$  concentration above the optimized value

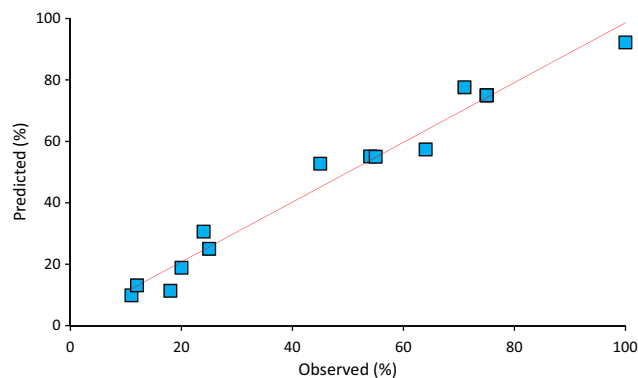


The results of the present study are in agreement with Karci et al. (2012) where, under the same optimized dose of  $\text{H}_2\text{O}_2$ , the degradation of 2,4-DCP took 8 min for complete degradation. The residual concentration of  $\text{H}_2\text{O}_2$  calculated at the end of the experiment was 2.5 mM, indicating that only 7.5 mM of  $\text{H}_2\text{O}_2$  was consumed during the complete oxidation of TCP. Literature has also reported that  $\text{H}_2\text{O}_2 > 15.0$  mM results in reduction in degradation of chlorophenolic compound in parallel with the scavenging effect of  $\text{H}_2\text{O}_2$  on OH radical (Gan et al. 2018).

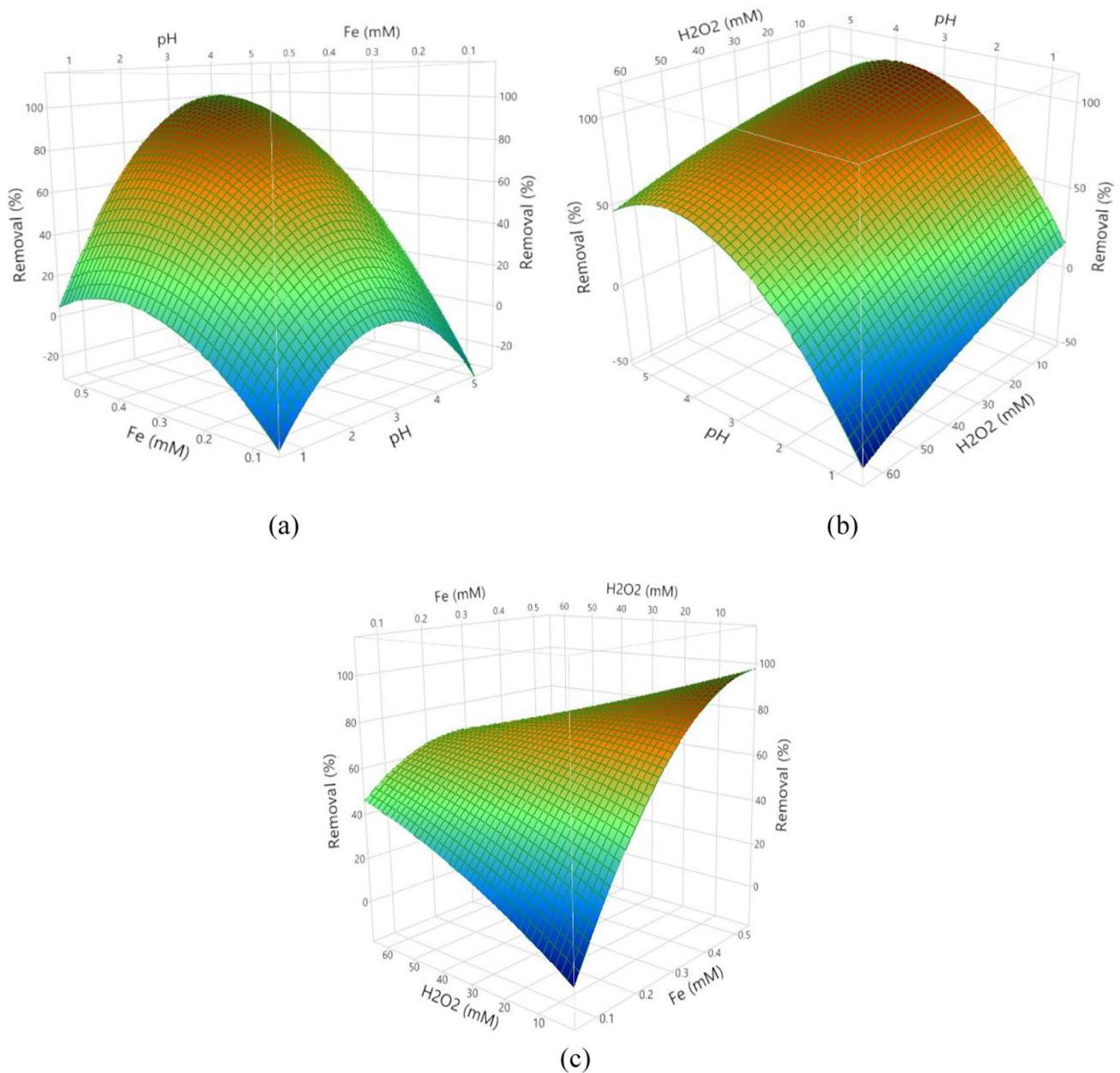
#### 4.5 Optimization of TCP degradation

An interactive relationship between process variables and studied response for degradation of TCP was assessed and response surface plots were drawn. The pH had a pronounced effect in governing the degradation of TCP. It played a determining role in controlling the dominant iron species, the activity of oxidant, and its stability. As seen in Fig. 3a and b, with increase in pH values, the removal efficiency also improved and was maximum between 3.0 and 4.0, whereas at pH 5.0, the removal efficiency started decreasing. At higher pH levels, iron precipitates as hydroxide and  $\text{H}_2\text{O}_2$  undergoes self-decomposition (Verma and Haritash 2019). The degradation of TCP was enhanced with increase in the dose of iron and was maximum at 0.5 mM with pH 3.0 and  $\text{H}_2\text{O}_2$  dose of 10.0 mM (Fig. 3a). Positive impact of addition of oxidant on degradation of TCP can be seen in Fig. 2b. The effect of variables, Fe and  $\text{H}_2\text{O}_2$  at pH 3.0 revealed that complete degradation of TCP took place with iron dose of 0.5 mM and  $\text{H}_2\text{O}_2$  concentration of 10.0 mM (Fig. 3c).

On the basis of model prediction and interaction of variable parameters, the optimized values for variable factors were deduced (Fig. 4). Based on the BBD, the optimized values were found to be 3.8 of pH, 0.53 mM of  $\text{Fe}^{2+}$ , and 10.14 mM



**Fig. 2** Normal plot of observed degradation (%) of 2,4,6-TCP against the Predicted value



**Fig. 3** Response surface plot of **a**  $\text{Fe}^{2+}$  and pH, **b** pH and  $\text{H}_2\text{O}_2$ , and **c**  $\text{Fe}^{2+}$  and  $\text{H}_2\text{O}_2$  as regulating parameters for degradation of 2,4,6-TCP

of  $\text{H}_2\text{O}_2$ . The results represent an insight of the effects of parameters toward degradation of TCP.

#### 4.6 Regression model and analysis of variance

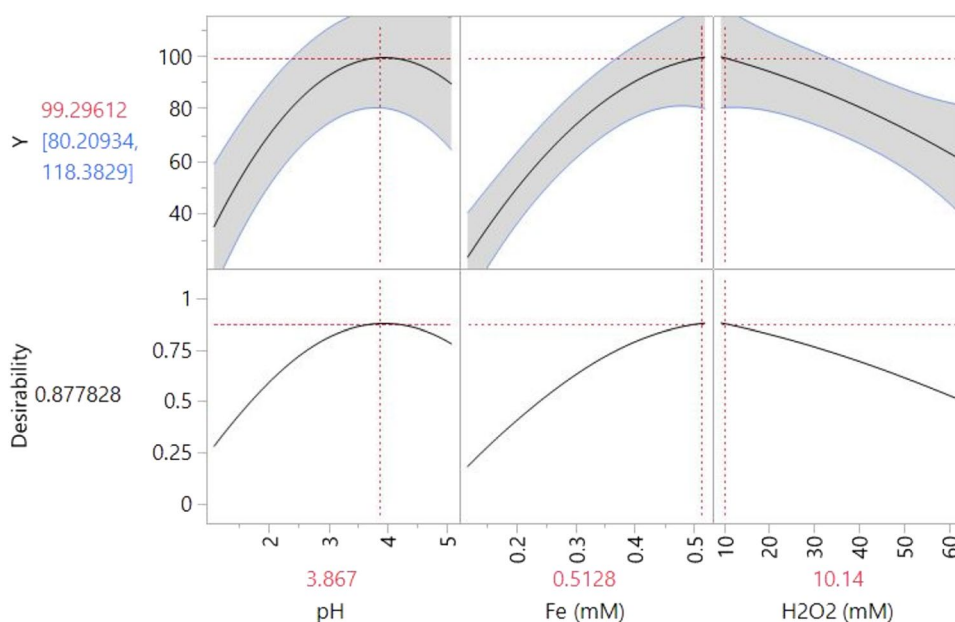
Analysis of the observed response (degradation of TCP) as given in Table 2 is represented by quadratic regression model using Eq. 2 shown below in the coded factors

$$Y = 75 + 15.75A + 17.375B - 2.375C \\ - 31.375A^2 - 15.625B^2 - 3.125C^2 \\ + 16.50AB - 3.5AC + 3.50BC.$$

To determine the goodness of the fit of the above equation, analysis of variance (ANOVA) of the experimental data was done, and the results are summarized in Table 2. *F*-value for the model suggests that it is significant for the value 20.18 with corresponding *p* value of regression being 0.002. The correlation coefficient ( $R^2$ ) between observed response and predicted response was 0.9732,



**Fig. 4** Optimization of regulating parameters (pH,  $\text{Fe}^{2+}$ , and  $\text{H}_2\text{O}_2$ ) for degradation of 2,4,6-TCP by Photo-Fenton ( $\text{UV}_{365}$ ) treatment



**Table 2** Analysis of variance (ANOVA) for percentage degradation of 2,4,6-TCP by Photo-Fenton treatment

Source	D.O.F	Sum of squares	Mean square	F-ratio	P
Regression	9	10,922.2	1213.58	20.18	0.002
pH	1	1984.5	1780.37	29.6	0.002
Fe	1	2415.1	1071.94	17.82	0.001
H <sub>2</sub> O <sub>2</sub>	1	45.1	192.76	3.2	0.426
pH*pH	1	3634.67	3634.67	60.43	0.0006
Fe*Fe	1	901.44	901.44	14.99	0.012
H <sub>2</sub> O <sub>2</sub> *H <sub>2</sub> O <sub>2</sub>	1	36.1	36.1	0.6	0.476
pH*Fe	1	2194.2	2194.2	12.16	0.008
pH*H <sub>2</sub> O <sub>2</sub>	1	1089	1089	18.1	0.408
Fe*H <sub>2</sub> O <sub>2</sub>	1	49	49	0.81	0.008
Residual error	1	1056.3	1056.25	17.56	0.009
Lack-of-fit	3	300.7	100.25		
Pure error	2	00	00	*	*
Cor total	14	11,222.9		0.002	

$R$ -squared = 0.9732; Adj.  $R$ -squared = 0.925

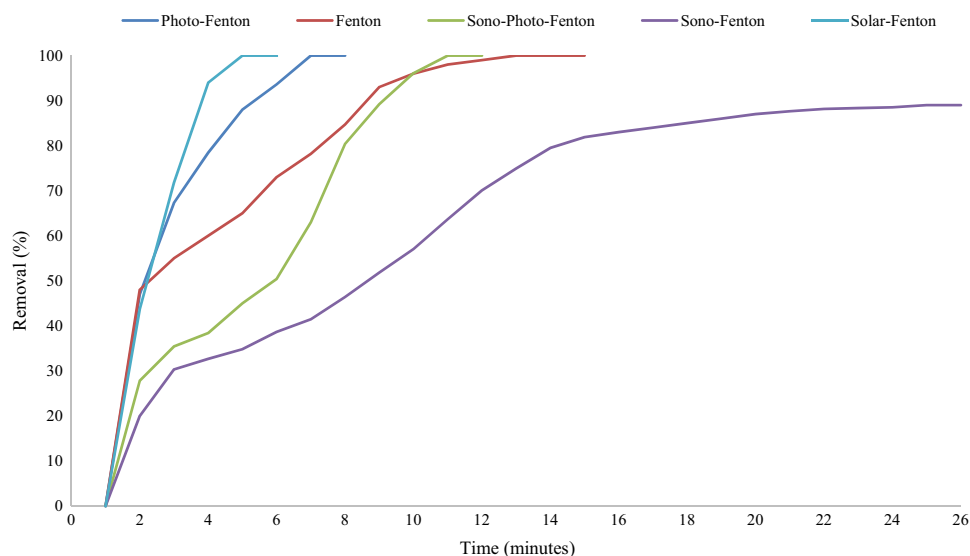
whereas the corresponding adjusted  $R^2$  value obtained was 0.9250. The values of  $R^2$  and adjusted  $R^2$  ensured the quadratic model a good fit with respect to the experimental response. The comparative plots of the observed value and predicted value for removal of TCP indicated significant agreement between the actual response against the response of model (Fig. 2). The significance of each independent variable was examined as per the  $p$  value ( $p$  value < 0.05 implies the term significant and vice versa). In the midst of test variables, the pH value (A) and Fe(II)

(B) had significant effect on the degradation of TCP. The quadratic effect coefficients of pH value ( $A^2$ ) were highly significant, whereas coefficient of quadratic effect of  $\text{Fe}^{2+}$  ( $B^2$ ), and interaction effect of pH and  $\text{Fe}^{2+}$  (AB), as well as  $\text{Fe}^{2+}$  and  $\text{H}_2\text{O}_2$  (BC) were slight significant model terms, while the rest other factors were statistically insignificant.

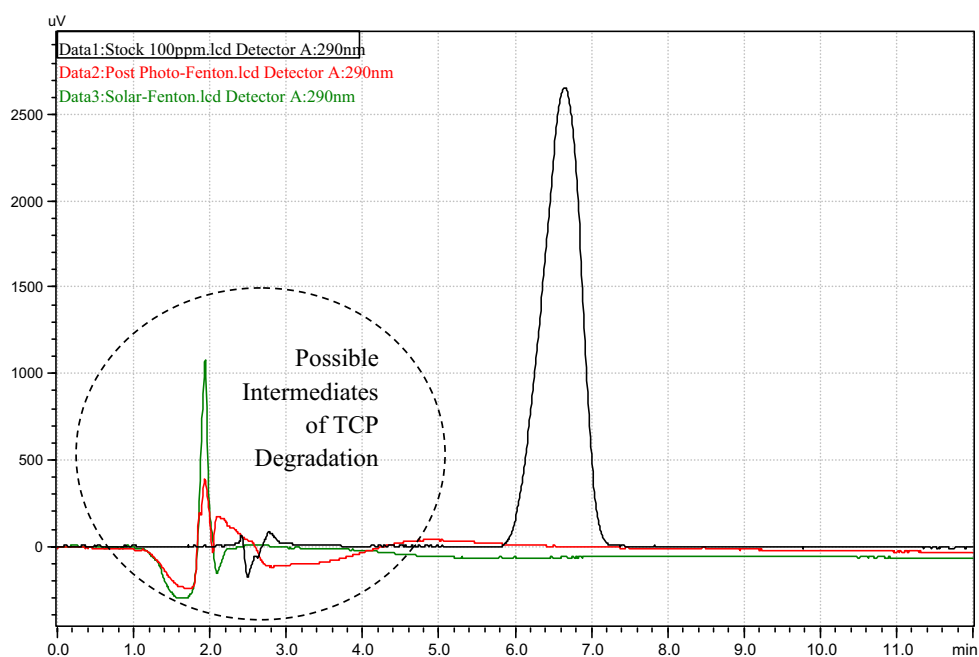
#### 4.7 Solar light and UV-assisted Fenton's treatment

A comparative study using different AOPs like Fenton's process, Solar-Fenton, Sonication, Sonolysis coupled with Fenton, and Sono-Photo-Fenton was conducted (Fig. 5). The experiments were performed at optimized conditions obtained in Photo-Fenton process, i.e.,  $\text{Fe}^{2+}$  0.5 mM, pH 3.0, and  $\text{H}_2\text{O}_2$  10.0 mM. The experiments involving sonication-integrated processes were conducted at operating frequency of 40 kHz. Among all the processes, Solar-Fenton was found to be most efficient, resulting in complete degradation of TCP within 5 min of the reaction time. Availability of sufficient high-intensity solar light resulted in a faster and complete degradation with 98% mineralisation of the model pollutant. It could also be reasoned from the fact that the average annual sunshine hours of India are 2780 h with land area receiving 4–7 kWh/m<sup>2</sup> of the solar energy (Solar Energy 2022). The Solar-Fenton experiments were conducted in presence of sunlight on a hot sunny day with temperature ranged between 43 °C and 46 °C. Early, quick, and complete degradation of TCP was observed in presence of solar light (Solar-Fenton) in comparison to  $\text{UV}_{365}$  Photo-Fenton. In case of Photo-Fenton ( $\text{UV}_{365}$ ), complete removal efficiency was attained within 6 min of the reaction duration because of the direct nucleophilic attack of the OH radicals. These

**Fig. 5** Comparison of Fenton's treatment with ultrasound and light (UV and solar) integrated Fenton for degradation of 2,4,6-TCP



**Fig. 6** The HPLC spectra of 2,4,6-TCO (100 mg/L untreated), Photo-Fenton (UV<sub>365</sub>) and Solar-Fenton-treated TCP-containing waste-water confirming mineralisation



reactive oxidation species attack the phenolic ring at the carbon not occupied with chlorine and abstract the  $\pi$  electron cloud over benzene ring causing mineralisation. However, the process resulted in only 50% mineralisation indicating the formation of intermediates. Presence of chlorine atoms at ortho and para position results in steric hindrance and lower mineralisation (Saritha et al. 2009). The HPLC analysis also confirmed the conversion of phenolic compound to intermediates (Fig. 6). When Fenton's process was used, though complete degradation was there, the system took 15 min,

twice the time taken by Photo-Fenton process. The process reported 50% mineralisation, while similar observation has also been reported by Saritha et al. (2009). Fenton's process when coupled with sonication resulted in 90% of degradation attained in 27 min of the reaction time. The slower rate of degradation is because both the processes compete for  $H_2O_2$  for generating OH radicals. Sonolysis coupled with Photo-Fenton demonstrated complete degradation within 15 min with higher mineralisation (75%) as compared to Photo-Fenton alone. The transmission of ultrasound waves

**Table 3** Comparison of different processes for maximum time taken for degradation

Processes	Conditions	Time for complete degradation (minutes)	Rate constant	Mineralisation efficiency (%)
Solar-Fenton	Fe <sup>2+</sup> : 0.5 mM H <sub>2</sub> O <sub>2</sub> : 10.0 mM pH: 3.0	5.0	0.92103	98
Photo-Fenton	Fe <sup>2+</sup> : 0.5 mM H <sub>2</sub> O <sub>2</sub> : 10.0 mM pH: 3.0 Light Source: UV tubes (365 nm) Intensity: 672W/m <sup>2</sup>	6.0	0.76753	50
Sono-Photo-Fenton	Fe <sup>2+</sup> : 0.5 mM H <sub>2</sub> O <sub>2</sub> : 10.0 mM pH: 3.0 Light Source: UV tubes (365 nm) Intensity: 672 W/m <sup>2</sup> Ultrasound: 40 kHz	12.0	0.38376	75
Fenton	Fe <sup>2+</sup> : 0.5 mM H <sub>2</sub> O <sub>2</sub> : 10.0 mM pH: 3.0	15.0	0.30701	50
Sono-Fenton	Fe <sup>2+</sup> : 0.5 mM H <sub>2</sub> O <sub>2</sub> : 10.0 mM pH: 3.0 Ultrasound: 40 kHz	27.0	0.00466	90

through the aqueous solution generates acoustic cavitation with production of highly reactive radicals (Anandan et al. 2020). The ultrasound irradiation induced cavitation effect generates heat, facilitates the proper mixing or mass transfer, and promotes the contact between materials and dispersion of contaminated layers of chemicals (Othmer and Overberger 1983). The physical impact of ultrasound accelerated the reaction by proper mixing of reagents and enhancing the surface area of the catalyst (Anandan et al. 2020). Thus, the synergistic effect of sonication and UV light enhanced the degradation and mineralisation of TCP. Table 3 reflects the summary of all the processes employed, conditions, and time taken toward maximum degradation.

#### 4.8 Kinetics of TCP degradation

The degradation kinetics for TCP during its degradation by Fenton's process was modeled to pseudo-first order kinetics using Eq. 9

$$\ln C_t/C_0 = -kt, \quad (9)$$

where  $C_0$  represents the initial concentration of TCP at time ' $t$ ' = 0,  $C_t$  is the final concentration of TCP at any time, ' $t$ ', and ' $k$ ' is the reaction rate constant. The pseudo-rate constant for all processes is represented in Table 3. It can be observed that the value of rate constant is maximum in case

of Solar-Fenton indicating higher degradation rate, whereas minimum in case of Sono-Fenton indicating slower degradation. Sunlight and UV-induced Photo-Fenton process provide photons, stimulating the regenerations of ferrous ions for the cyclic production of OH $\cdot$ . The rate of degradation followed the order:

Solar - Fenton > Photo (UV<sub>365</sub>) - Fenton  
> Sono - Photo - Fenton > Fenton  
> Sono - Fenton.

## 5 Conclusion

The study revealed that Fenton's process is an effective tool for degradation of 2,4,6-TCP in wastewater originating from pulp & paper and textile industries. Coupling Fenton's process with light (UV<sub>365</sub>) can significantly enhance the rate of TCP degradation. Under optimized conditions (Fe<sup>2+</sup> 0.5 mM; H<sub>2</sub>O<sub>2</sub> 10.0 mM and pH 3.0), complete degradation of TCP was attained, however only 50% mineralisation efficiency was observed. Integration of ultrasound with Fenton's treatment compromises with the efficiency of the process, and should be avoided. Since solar Fenton could rapidly degrade TCP, it may be treated as a sustainable and cost-effective solution to treatment

of TCP. Solar-Fenton was observed with nearly complete mineralisation of the chlorophenolic compound as compared to Photo-Fenton in similar stipulated time period. Solar-Fenton treatment for degradation of TCP represents a rapid degradation rate which may be owing to chemical oxidation by Fenton's process as well as photolysis under sunlight. The quick degradation results in the treatment of larger volumes of high strength wastewater in the stipulated time and ensures mineralisation, thereby removing toxicity associated with residual reaction intermediates. The study recommends chemical/Fenton's oxidation of secondarily treated wastewater of textile or pharmaceutical effluent to minimize the environmental effects of chlorophenols.

**Acknowledgements** The authors are grateful to Mr. Harsh Pipil, Research Scholar, Department of Environmental Engineering at DTU for the assistance during this study.

**Author contributions** SY: conceptualization, execution, data compilation, and draft writing. SK: conceptualization, supervision, reviewing, and editing. AKH: conceptualization, supervision, reviewing, and editing.

**Funding** The authors declare that no funds, grants, or other supports were received during the preparation of this manuscript.

**Availability of data and materials** Necessary data have been provided along with this manuscript.

## Declarations

**Conflict of interest** The authors have no relevant financial or non-financial interests to disclose.

**Ethical approval** Not applicable.

**Consent to participate** Authors provide the consent to participate.

**Consent to publish** Authors provide the consent to publish the manuscript.

## References

- Anandan S, Ponnusamy VK, Ashokkumar M (2020) A review on hybrid techniques for the degradation of organic pollutants in aqueous environment. *Ultrason Sonochem* 67:105130. <https://doi.org/10.1016/j.ultsonch.2020.105130>
- Annachhatre AP, Gheewala SH (1996) Biodegradation of chlorinated phenolic compounds. *Biotechnol Adv* 14:35–56. [https://doi.org/10.1016/0734-9750\(96\)00002-X](https://doi.org/10.1016/0734-9750(96)00002-X)
- APHA, Method 3500-Fe (1997) Standard methods for the examination of water and wastewater, 22nd edn. American Public Health Association/American Water Works Association/Water Environment Federation, Washington, DC
- ATSDR (2007) Comprehensive environmental response, compensation, and liability act (CERCLA) priority list of hazardous substances. ATSDR
- Bilal M, Rasheed T, Iqbal HMN, Hu H, Wang W, Zhang X (2018) Toxicological assessment and UV/TiO<sub>2</sub>-based induced degradation profile of reactive black 5 dye. *Environ Manag* 61(1):171–180. <https://doi.org/10.1007/s00267-017-0948-7>
- Czaplicka M (2004) Sources and transformations of chlorophenols in the natural environment. *Sci Total Environ* 322(1–3):21–39. <https://doi.org/10.1016/j.scitotenv.2003.09.015>
- Gan L, Li B, Guo M et al (2018) Mechanism for removing 2,4-dichlorophenol via adsorption and Fenton-like oxidation using iron-based nanoparticles. *Chemosphere* 206:168–174. <https://doi.org/10.1016/j.chemosphere.2018.04.162>
- Gaya UI, Abdullah AH, Hussein MZ, Zainal Z (2010) Photocatalytic removal of 2, 4, 6-trichlorophenol from water exploiting commercial ZnO powder. *Desalination* 263(1–3):176–182. <https://doi.org/10.1016/j.desal.2010.06.055>
- Ge T, Han J, Qi Y et al (2017) The toxic effects of chlorophenols and associated mechanisms in fish. *Aquat Toxicol* 184:78–93. <https://doi.org/10.1016/j.aquatox.2017.01.005>
- Guo M, Weng X, Wang T, Chen Z (2017) Biosynthesized iron-based nanoparticles used as a heterogeneous catalyst for the removal of 2,4-dichlorophenol. *Sep Purif Technol* 175:222–228. <https://doi.org/10.1016/j.seppur.2016.11.042>
- IARC (International Agency for Research on Cancer) (2004) Overall evaluations of carcinogenicity to humans. IARC Monographs, pp 1–82
- Kantar C, Oral O, Urken O et al (2019) Oxidative degradation of chlorophenolic compounds with pyrite-Fenton process. *Environ Pollut* 247:349–361. <https://doi.org/10.1016/j.envpol.2019.01.017>
- Karci A, Arslan-Alaton I, Olmez-Hanci T, Bekbölet M (2012) Transformation of 2, 4-dichlorophenol by H<sub>2</sub>O<sub>2</sub>/UV-C, Fenton and photo-Fenton processes: oxidation products and toxicity evolution. *J Photochem Photobiol A* 230(1):65–73. <https://doi.org/10.1016/j.jphotochem.2012.01.003>
- Kavitha V, Palanivelu K (2004) The role of ferrous ion in Fenton and photo-Fenton processes for the degradation of phenol. *Chemosphere* 55(9):1235–1243. <https://doi.org/10.1016/j.chemosphere.2003.12.022>
- Kavitha V, Palanivelu K (2016) Degradation of phenol and trichlorophenol by heterogeneous photo-Fenton process using Granular Ferric Hydroxide®: comparison with homogeneous system. *Int J Environ Sci Technol* 13:927–936. <https://doi.org/10.1007/s13762-015-0922-y>
- Michałowicz J (2010) 2,4,5-trichlorophenol and its derivatives induce biochemical and morphological changes in human peripheral blood lymphocytes in vitro. *Arch Environ Contam Toxicol* 59:670–678. <https://doi.org/10.1007/s00244-010-9508-3>
- Olaniran AO, Igbinsola EO (2011) Chlorophenols and other related derivatives of environmental concern: properties, distribution and microbial degradation processes. *Chemosphere* 83(10):1297–1306. <https://doi.org/10.1016/j.chemosphere.2011.04.009>
- Othmer CG, Overberger GT (1983) Ultrasonics. In: Seaborg C (ed) Kirk-Othmer encyclopedia of chemical technology, 23rd edn. Wiley, Cham, p 462. <https://doi.org/10.1021/j150469a016>
- Pandit AB, Gogate PR, Mujumdar S (2001) Ultrasonic degradation of 2, 4: 6 trichlorophenol in presence of TiO<sub>2</sub> catalyst. *Ultrason Sonochem* 8(3):227–231. [https://doi.org/10.1016/S1350-4177\(01\)00081-5](https://doi.org/10.1016/S1350-4177(01)00081-5)
- Pipil H, Yadav S, Chawla H et al (2022) Comparison of TiO<sub>2</sub> catalysis and Fenton's treatment for rapid degradation of Remazol Red Dye in textile industry effluent. *Rend Fis Acc Lincei* 33:105–114. <https://doi.org/10.1007/s12210-021-01040-x>
- Saritha P, Raj DSS, Aparna C et al (2009) Degradative oxidation of 2,4,6 trichlorophenol using advanced oxidation processes: a comparative study. *Water Air Soil Pollut* 200:169–179. <https://doi.org/10.1007/s11270-008-9901-y>

- Sharma A, Verma M, Haritash AK (2016) Degradation of toxic azo dye (AO7) using Fenton's process. *Adv Environ Res* 5(3):189–200. <https://doi.org/10.12989/aer.2016.5.3.189>
- Solar Energy (2022) Ministry of New and Renewable Energy. <https://mnre.gov.in/solar/current-status/>. Accessed 28 May 2022
- Verma M, Haritash AK (2019) Degradation of amoxicillin by Fenton and Fenton-integrated hybrid oxidation processes. *J Environ Chem Eng* 7(1):102886. <https://doi.org/10.1016/j.jece.2019.102886>
- Verma M, Haritash AK (2020) Photocatalytic degradation of Amoxicillin in pharmaceutical wastewater: a potential tool to manage residual antibiotics. *Environ Technol Innov* 20:101072. <https://doi.org/10.1016/j.eti.2020.101072>
- Vlastos D, Antonopoulou M, Konstantinou I (2016) Evaluation of toxicity and genotoxicity of 2-chlorophenol on bacteria, fish and human cells. *Sci Total Environ* 551:649–655. <https://doi.org/10.1016/j.scitotenv.2016.02.043>
- Xu LJ, Wang JL (2013) Degradation of chlorophenols using a novel Fe<sub>0</sub>/CeO<sub>2</sub> composite. *Appl Catal B* 142–143:396–405. <https://doi.org/10.1016/j.apcatb.2013.05.065>
- Yadav S, Pipil H, Chawla H, Taneja S, Kumar S, Haritash AK (2022) Textile industry wastewater treatment using eco-friendly techniques. In: Kanwar VS, Sharma SK, Prakasam C (eds) *Proceedings of international conference on innovative technologies for clean and sustainable development (ICITCSD-2021)*. Springer, Cham. [https://doi.org/10.1007/978-3-030-93936-6\\_6](https://doi.org/10.1007/978-3-030-93936-6_6)
- Yadav S, Kumar S, Haritash AK (2023) A comprehensive review of chlorophenols: Fate, toxicology and its treatment. *J Environ Manag* 342:118254. <https://doi.org/10.1016/j.jenvman.2023.118254>
- Zhu M, Lu J, Zhao Y, Guo Z, Hu Y, Liu Y, Zhu C (2021) Photochemical reactions between superoxide ions and 2, 4, 6-trichlorophenol in atmospheric aqueous environments. *Chemosphere* 279:130537. <https://doi.org/10.1016/j.chemosphere.2021.130537>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

# Speech Dereverberation with Frequency Domain Autoregressive Modeling

Anurenjan Purushothaman, *Member, IEEE*, Debottam Dutta, Rohit Kumar, *Student Member, IEEE* and Sriram Ganapathy, *Senior Member, IEEE*

**Abstract**—Speech applications in far-field real world settings often deal with signals that are corrupted by reverberation. The task of dereverberation constitutes an important step to improve the audible quality and to reduce the error rates in applications like automatic speech recognition (ASR). We propose a unified framework of speech dereverberation for improving the speech quality and the ASR performance using the approach of envelope-carrier decomposition provided by an autoregressive (AR) model. The AR model is applied in the frequency domain of the sub-band speech signals to separate the envelope and carrier parts. A novel neural architecture based on dual path long short term memory (DPLSTM) model is proposed, which jointly enhances the sub-band envelope and carrier components. The dereverberated envelope-carrier signals are modulated and the sub-band signals are synthesized to reconstruct the audio signal back. The DPLSTM model for dereverberation of envelope and carrier components also allows the joint learning of the network weights for the down stream ASR task. In the ASR tasks on the REVERB challenge dataset as well as on the VOICES dataset, we illustrate that the joint learning of speech dereverberation network and the E2E ASR model yields significant performance improvements over the baseline ASR system trained on log-mel spectrogram as well as other benchmarks for dereverberation (average relative improvements of 10-24% over the baseline system). The speech quality improvements, evaluated using subjective listening tests, further highlight the improved quality of the reconstructed audio.

**Index Terms**—Frequency domain auto-regressive modeling, Dereverberation, end-to-end ASR, Joint modeling.

## I. INTRODUCTION

THE wide spread adoption of voice technologies like meeting assistants, smart speakers, in-car entertainment systems, and virtual assistants imply that the audio signal at the input of these system is impacted by reverberation and noise artifacts [1]. The performance of the downstream applications like, automatic speech recognition, speaker/language recognition, emotion recognition or voice activity detection, is shown to degrade significantly in reverberant conditions [2]–[6]. The performance deterioration is primarily attributed to the smearing of the temporal envelopes caused by reverberation [7]. The temporal smearing is caused by the emplacement of the direct path signal on reflected signals, resulting in a weighted summation of delayed components [8].

This paper was partly funded by grants from the Samsung Research India, Bangalore.

A. Purushothaman, and S. Ganapathy are with the Learning and Extraction of Acoustic Patterns (LEAP) lab, Department of Electrical Engineering, Indian Institute of Science, Bangalore, India, 560012. D. dutta is with University of Illinois Urbana-Champaign. R. Kumar is associated with Johns Hopkins University. e-mail: {anurenjanp, srirang}@iisc.ac.in

One of the approaches to deal with the adverse far-field conditions is to develop a front-end which performs signal enhancement. Several techniques for dereverberation like signal processing based (for example, weighted prediction error (WPE) [9]), mask estimation based (for example, time-frequency mask estimation [10]) and multi-channel beamforming based (for example, time-delay estimation [11], generalized eigen-value [12], [13]) have been explored to improve the signal quality. On the other hand, another effective approach for system development in reverberant conditions is that of multi-condition training [14]. However, even with these pre-processing and multi-condition training methods, the beamformed signal contains significant amount of temporal smoothing which adversely impacts the ASR performance [15].

In the traditional setting, the first step in the analysis of a signal is the short-term Fourier transform (STFT). The key assumptions about the convolution model of reverberation artifacts, is applicable for a long-analysis window in the time domain, or using convolutional transfer function with cross-band filters in the STFT domain [16], [17]. In our case, we use the former approach of long analysis window and explore dereverberation in the sub-band envelope domain. As the reverberation is a long-term convolution effect, we highlight that room impulse response (typically with a  $T_{60} > 400ms$ ) can be absorbed as a multiplication in the frequency domain, as well as a convolution in the sub-band envelope domain.

In this paper, we investigate the effect of reverberation on the long-term sub-band signals of speech using an envelope-carrier decomposition. The extraction of the sub-band envelope is achieved using the autoregressive (AR) modeling approach in the spectral domain, termed as frequency domain linear prediction (FDLP). Our previous work showed that a feature level enhancement with the FDLP envelope improves speech recognition performance [18], [19]. However, the prior works did not allow the reconstruction of the audio signal for quality improvement. Further, the enhancement of the carrier signal was not addressed in the previous work due to the challenges in the handling the impulsive nature of the carrier signal.

In this paper, we propose a novel approach to the joint dereverberation of the envelope and carrier signals using a neural modeling framework. While using the sub-band signals directly, the sample level de-convolution with a suitable loss function can be a difficult design choice to learn using neural models. Hence, we propose using an envelope-carrier decomposition of the sub-band signals. Our rationale for the envelope-carrier decomposition based setup is the fact the envelope information is alone used in the ASR experiments.



Thus, the ASR loss has to impact only the envelope dereverberation branch. However, the carrier and the envelope components are part of the signal reconstruction branch.

We develop a dual path long short term memory (DPLSTM) architecture for the dereverberation of the temporal envelope and carrier signals. In our case, the goal of the neural model is to perform a dereverberation of the envelope and the carrier components of the sub-band signal. These signals have a time profile, with varying dynamic range and properties. Further, merging all the sub-band signals in the decomposition also brings in a frequency profile. Thus, the design choice of the neural model, for enhancing the sub-band envelope-carrier signals, has to learn the sequence level patterns in both the time and frequency domains. The DPLSTM [20] is a suitable choice, as the model is able to integrate information effectively in both the time and frequency domains.

Following the dereverberation step, the sub-band modulation and synthesis step generates the reconstructed audio signal. The neural enhancement and sub-band synthesis can also be implemented as a part of the larger neural pipeline for downstream tasks like ASR, thereby enabling the joint learning of the ASR and dereverberation model parameters. We refer to the proposed approach as Dual path dereverberation using Frequency domain Auto-Regressive modeling (DFAR) and the joint end-to-end model as E2E-DFAR.

Various ASR experiments are performed on the REVERB challenge dataset [21] as well as the VOICES dataset [22], [23]. The key contributions from this work, over the prior work [18], can be summarized as follows,

- Proposing an analysis for dereverberation with a sub-band decomposition and envelope-carrier demodulation.
- Proposing a dual-path long short time memory model named, DPLSTM for the dereverberation of sub-band envelope and carrier signals. This approach is termed as DFAR.
- Developing a joint learning scheme, where the ASR model and the DFAR model are optimized in a single end-to-end framework. This model is referred to as the E2E-DFAR.
- Evaluating the proposed approaches on speech quality improvement tasks as well as on ASR tasks on two benchmark datasets - REVERB challenge dataset and the VOICES dataset.

## II. RELATED PRIOR WORK

### A. Enhancement and dereverberation

For speech enhancement, Xu et. al. [24] devised a mapping from noisy speech to clean speech using a supervised neural network. In a similar manner, ideal ratio mask based neural mappings [25] have been explored for speech separation tasks. On the dereverberation front, Zhao et. al. proposed an LSTM model for late reflection prediction in the spectrogram domain [26]. Han et. al [27] developed a spectral mapping approach using the log-magnitude inputs and Williamson et. al [10] proposed a mask-based approach for dereverberation on the complex short-term Fourier transform. In a different line of

work, speech enhancement in the time domain was pursued by Pandey et. al [28].

The application of speech dereverberation as a pre-processing step for downstream applications like ASR have been explored in several works (for example, [29]–[31]). The recent years have seen the use of recurrent neural network architectures for dereverberation. For example, Maas et. al [32], utilized a recurrent neural network (RNN) to establish mapping between noise-corrupted input features and their corresponding clean targets. Also, the use of a context-aware recurrent neural network-based convolutional encoder-decoder architecture was investigated by Santos et. al. [33].

### B. Robust multi-channel ASR

In the design of robust ASR, Generalized sidelobe canceller (GSC) [34], [35] is a common approach. It was introduced by Li et. al in [36], where the authors proposed a neural network-based generalized side-lobe canceller. To combine spectral and spatial information from multiple channels using attention layers, an end-to-end multi-channel transformer was investigated in [37]. In another attention modelling approach, the streaming ASR model based on monotonic chunk-wise attention was proposed by Kim et. al in [38]. Ganapathy et. al. [4] proposed a 3-D CNN model for far-field ASR.

### C. Joint modeling of enhancement and ASR

The attempt proposed by Wang et. al. [39] incorporates a DNN based speech separation model coupled with a DNN based acoustic model. The work reported by Wu et. al. [40] explored a unification of separately trained speech enhancement neural model and the acoustic model, where the joint model is fine-tuned to improve the ASR performance. Here, the DNN based dereverberation front end leverages the knowledge about reverberation time. While traditional GSC is optimized for signal level criteria, the neural network-based GSC, proposed by Li et. al [36], was optimized for ASR cost function.

## III. PROPOSED DFAR APPROACH

### A. Quadrature Mirror Filter (QMF)

For the sub-band decomposition, we had the following design considerations

- The decomposition approach should allow the long-term artifacts of reverberation to be captured in the sub-band domain as a convolution,
- The analysis method should allow a perfect reconstruction back to the audio using the synthesis part, and
- The sub-band components should be critically sampled for efficient computation of the dereverberated components in a deep neural model.

The quadrature mirror filter (QMF) met all the above requirements and hence, this work has used the QMF analysis and synthesis for speech dereverberation task.

A quadrature mirror filter (QMF) is a filter whose magnitude response is a mirror reflection at quadrature frequency ( $\frac{\pi}{2}$ ) of another filter [41]. In signal processing, the QMF filter-pairs are used for the design of perfect reconstruction filter

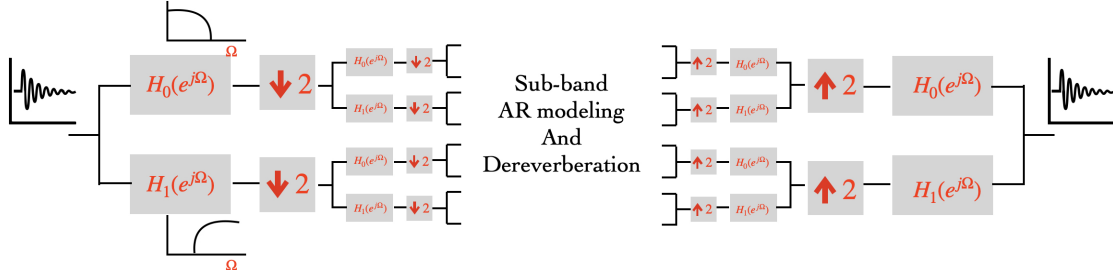


Fig. 1. Illustration of a 4-channel uniform QMF decomposition using a 2-stage binary QMF tree. In our work, we use 64-channel decomposition, using a 6-way binary tree.

banks. Let  $H_0(e^{j\Omega})$  and  $H_1(e^{j\Omega})$  denote low-pass and high-pass filter's frequency domain function, where  $\Omega$  is the digital frequency. In addition to the quadrature property ( $H_1(e^{j\Omega}) = H_0(e^{j(\Omega-\pi)})$ ), the filters used in QMF filter-banks also satisfy the complimentary property,

$$|H_0(e^{j\Omega})|^2 + |H_1(e^{j\Omega})|^2 = 1. \quad (1)$$

The design of sub-band decomposition scheme with QMF involves a series of filtering and down-sampling operations for the analysis [42]. The synthesis is achieved by up-sampling and filtering operations. A tree-like structure can be formed using a recursive decomposition operation. The down-sampling process enables a critical rate of processing, where the sum of the number of samples in each sub-band equals the number of the samples in the full-band signal.

In this work, we use an uniform 64-band Quadrature Mirror Filter bank (QMF) for decomposing the input signal into 64 uniformly spaced frequency bands. Inspired by the audio decomposition scheme outlined in Motlicek et. al. [43], we use a 6-level binary tree structure. The schematic of the sub-band decomposition is shown in Fig. 1. For the implementation in a neural pipeline, the down-sampling operation is equivalent to a stride, while the up-sampling operation is that of un-pooling.

### B. Autoregressive modeling of temporal envelopes

The application of linear prediction model in the frequency domain, an approach called frequency domain linear prediction (FDLP), enables the modeling of the temporal envelopes of a signal with an autoregressive (AR) model [8], [44]. The sub-band signal is transformed to the spectral domain using a discrete cosine transform (DCT) [8], where a linear prediction model is applied.

Let the sub-band signal be denoted as  $x_q[n]$ , where  $q = 1, \dots, Q$  denotes the sub-band index. The analytic signal, in signal processing theory, is a complex valued function, whose real value is the original signal while the imaginary value is the Hilbert transform of the signal. It finds application in single side-band amplitude modulation and quadrature filtering. Let the analytic version of sub-band signal,  $x_q^a[n]$  be denoted as,  $x_q^a[n]$ . The corresponding analytic signal in the frequency domain,  $X_q^a[k]$  can be shown to be the one-sided discrete Fourier transform (DFT) [8] of the even symmetric version of  $x_q[n]$ .

We apply linear prediction (LP) on the frequency domain signal,  $X_q^a[k]$ . The corresponding LP coefficients are denoted

by  $\{b_p\}_{p=0}^m$ , where  $m$  is the order of the LP. The temporal envelope estimate of  $x_q^a[n]$ , is given by,

$$e_q[n] = \frac{\alpha}{|\sum_{p=0}^m b_p e^{-2\pi i p n}|^2} \quad (2)$$

where  $\alpha$  denotes the LP gain. The envelope represents the autoregressive model of the Hilbert envelope. In this paper, we use the Burg method [45] for estimating the AR envelope.

The corresponding carrier (remaining residual signal),  $c_q[n]$  is found as,

$$c_q[n] = \frac{x_q[n]}{\sqrt{e_q[n]}} \quad (3)$$

The division operation in the expression above is well defined as the envelope given in Eq. (2) is always positive. Further, the modeling of the temporal envelopes using the AR model ensures that the peaks of the sub-band signal in the time-domain are well represented [46], [47].

### C. Effect of reverberation on envelope and carrier signals

The effect of reverberation on the time-domain speech signal can be expressed in the form of a convolution operation,

$$y[n] = x[n] * r[n], \quad (4)$$

where  $x[n]$  denotes the clean speech signal,  $r[n]$  is the impulse response of the room and  $y[n]$ , is the reverberant speech signal. The room response function can be further split into two parts,  $r[n] = r_e[n] + r_l[n]$ , where  $r_e[n]$  and  $r_l[n]$  are the early and late reflection components, respectively.

Let  $x_q[n]$ ,  $r_q[n]$  and  $y_q[n]$  denote the sub-band versions of the clean speech, room-response function and the reverberant speech signal respectively. Assuming an ideal band-pass filtering, it can be shown that the analytic signal,  $x_q^a[n]$ , is given by [8], [48],

$$y_q^a[n] = \frac{1}{2}[x_q^a[n] * r_q^a[n]], \quad (5)$$

For band-pass filters with narrow band-width, the envelopes of the reverberant speech can be approximated as [18],

$$e_{yq}[n] \simeq \frac{1}{2}e_{xq}[n] * e_{rq}[n], \quad (6)$$

where  $e_{yq}[n]$ ,  $e_{xq}[n]$ ,  $e_{rq}[n]$  denote the sub-band envelopes of reverberant speech, clean speech and room response respectively. Prior efforts in envelope normalization focus on suppressing the linear effects of reverberation by setting the

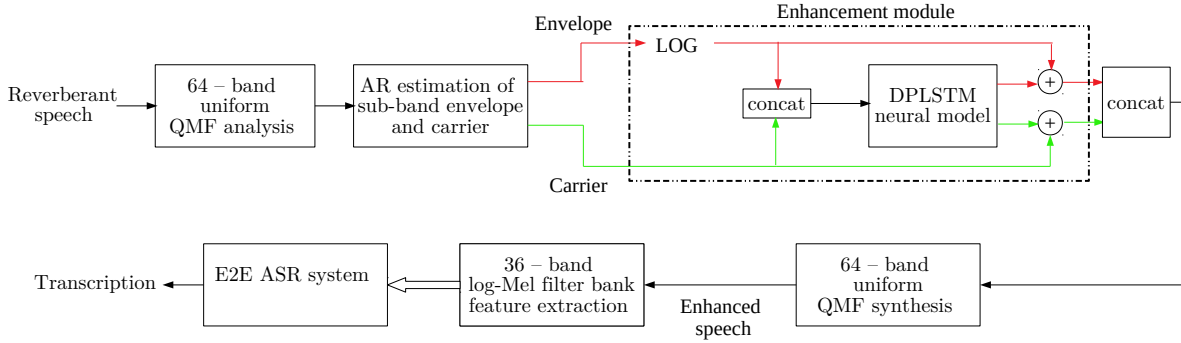


Fig. 2. Block schematic of speech dereverberation model, the feature extraction module and the E2E ASR model. The red arrows denote the envelopes,  $e[n]$ , and the green arrows represent the carrier,  $c[n]$ . The entire model can be constructed as an end-to-end neural framework.

gain of the reconstructed envelopes to unity [49]. However, in this work, we develop neural models that can remove the non-linear effects of reverberation. The reverberant sub-band envelope can also be viewed as an additive model [18], [50].

$$e_{yq}[n] = e_{yqe}[n] + e_{yql}[n], \quad (7)$$

where,  $e_{yqe}[n]$  is the early reflection component (which includes the direct path and the early reflections), while  $e_{yql}[n]$  is the late reflection part of the sub-band envelope  $e_{yq}[n]$ .

The key assumptions about the reverberation model of Eq. (4-6), is a long-analysis window in the time domain. As the reverberation is a long-term convolution effect, we highlight that the room impulse response (typically with a  $T60 > 400ms$ ) can be absorbed as a multiplication in the frequency domain, as well as a convolution in the sub-band envelope domain, only in the case of a long analysis window. The widely used short-time Fourier transform (STFT) does not capture the room impulse response function directly, and hence does not allow a convolutive modeling of the artifacts. Further, the phase effects in STFT domain are somewhat cumbersome to model. The above mentioned issues of STFT are also verified experimentally in Sec. V.

**Envelope enhancement:** A neural model can be used to learn late reflection component  $e_{xql}[n]$  from the sub-band temporal envelope  $e_{xq}[n]$ . The predicted late reflection component can be subtracted from the sub-band envelope to suppress the artifacts of reverberation.

We pose the problem in the log domain to reduce the dynamic range of the envelope magnitude. The neural model is trained with reverberant sub-band envelopes ( $\log(e_{xq}[n])$ ) as input. The model outputs the gain (in the log domain, i.e.,  $\log(\frac{e_{sq}[n]}{e_{xq}[n]})$ ). This gain is added in the log-domain to generate dereverberated signal envelope ( $\log(\hat{e}_{sq}[n])$ ).

**Envelope-carrier dereverberation model:** In a similar manner, the non-linear mapping between the reverberant carrier,  $c_{xq}[n]$  and clean carrier,  $c_{sq}[n]$ , can be learned using a neural network. A neural model is trained with reverberant sub-band carrier ( $c_{xq}[n]$ ) as input and model outputs the residual (an estimate of the late reflection component,  $c_{xql}[n]$ ), which when added with the reverberant carrier generates the estimate of source signal carrier ( $\hat{c}_{sq}[n]$ ). Instead of independent operations of dereverberation of the envelope and the carrier, we propose to learn the mapping between clean and

reverberant versions of both the envelope and the carrier in a joint model. The input to the neural model is the sub-band reverberant envelope spliced with the corresponding carrier signal. The network is trained to output the late reflection components of both the envelope and carrier. With this approach, the model also learns the non-linear relationships between the envelope and carrier signals for the dereverberation task. From the model output, the estimate of the clean sub-band signal  $\hat{s}_q[n]$  is generated. In our implementation, the audio signal is divided into non overlapping segments of 1 sec. length and passed through the envelope-carrier dereverberation model. The model is outlined in Fig. 2.

#### D. DFAR model architecture using DPLSTM

We propose the dual path long short term model (DPLSTM) for the dereverberation of the envelope-carrier components of the sub-band signal. Our proposed model is inspired by dual path RNN proposed by Luo et. al [20]. The block schematic of the DPLSTM model architecture is shown in Fig. 3. For 1 sec. of audio sampled at 16 kHz, the envelope ( $E^y$ ) and carrier ( $C^y$ ) components of the critically sampled sub-band signals (64 channel QMF decomposition) are of length 250. The envelope/carrier signals of all the sub-bands, for the reverberant signal ( $Y$ ), is of size  $64 \times 250$ . The combined envelope-carrier input is therefore of size  $128 \times 250$ , which forms the input to the DPLSTM model. The DPLSTM model outputs are also of the same size of the input, and the model is trained using the mean squared error (MSE) loss.

The proposed DPLSTM has two paths, one LSTM path models the recurrence along the time dimension, while the other models the recurrence along the frequency dimension. We use two separate 3-layer LSTM architectures for these paths. The output dimensions are kept the same as the input dimension for each of these paths. The frequency recurrence LSTM output is transposed and these are concatenated in the frequency dimension. This combined output is fed to a multi layer bi-directional LSTM, which performs recurrence over time. The final output is split into sub-band specific envelope and carrier components. The modulation of the envelope with the respective carrier components generates the sub-band signals, which are passed through the QMF synthesis to generate the full-band dereverberated signal.

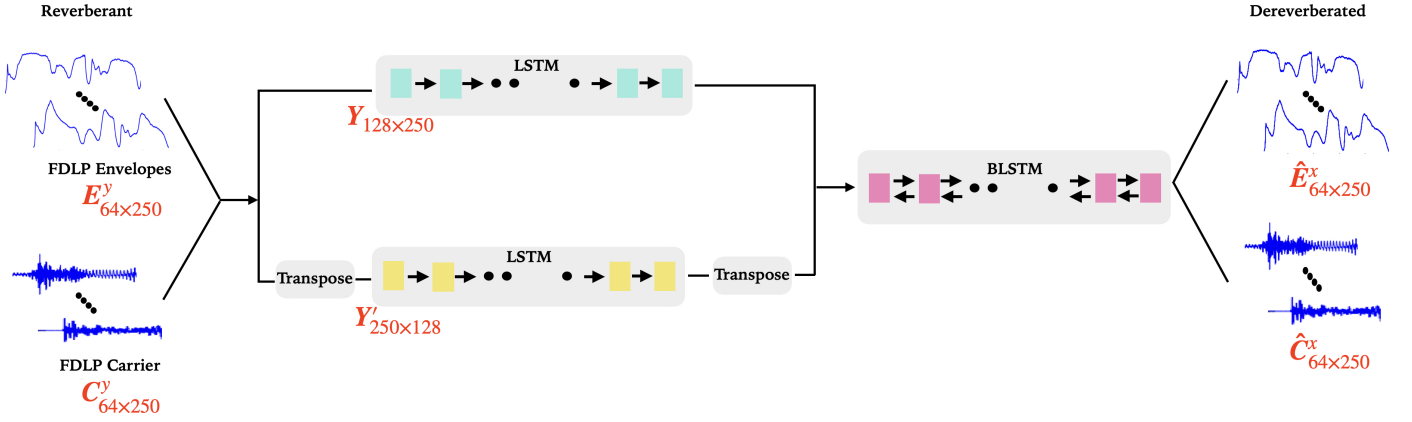


Fig. 3. The dual path LSTM model architecture for envelope-carrier dereverberation. The top LSTM path models the recurrence along the time dimension while the one on the bottom models the recurrence along the frequency dimension.

#### E. Joint learning of dereverberation model for ASR

The joint learning of the envelope-carrier dereverberation module with the E2E ASR architecture is achieved by combining the two separate models to train a single joint neural model. This is shown in Fig. 2. We initialize the modules with weights obtained from the independent training of each component. Specifically, the envelope-carrier dereverberation model is trained using MSE loss, which is followed by a sub-band synthesis (right side half of Fig. 1). The QMF synthesis is implemented using a 1-D CNN layer to generate the dereverberated speech signal. Further, the E2E ASR architecture is separately trained on the log-mel filter bank features, obtained from the dereverberated speech. The mel-filter bank feature generation can also be implemented using a neural framework. Thus, the final model, composed of neural components from the envelope-carrier dereverberation, sub-band synthesis, feature extraction and ASR, can now be jointly optimized using the E2E ASR loss function. This model is referred to as E2E-DFAR model<sup>1</sup>. The trainable components are the DPLSTM model and the ASR model parameters, while the sub-band synthesis and feature extraction parameters are not learnable.

### IV. EXPERIMENTAL SETUP

#### A. Datasets

1) *REVERB Challenge ASR*: The audio samples in REVERB challenge dataset [51] are 8 channel recordings with both real and simulated reverberant conditions. The real samples are utterances from MC-WSJ-AV corpus [52], spoken by human speakers in a noisy reverberant room. The simulated samples of the dataset are generated by convolving six different room impulse responses with the clean WSJCAM0 recordings followed by the addition of noise at the signal-to-noise ratio (SNR) of 20 dB. The training data consists of 7861 ( ~ 17.5 hours) utterances which are obtained by convolving WSJCAM0 train data with 24 measured RIRs. The reverberation time of the measured impulse responses range from 0.2 to 0.8 sec. The training, development and evaluation

data sets consist of 92, 15 and 38 speakers respectively. The development data consists of 1663 (3.3 hours) utterances and the evaluation data consists of 2548 (5.4 hours) utterances.

2) *VOiCES Dataset*: The VOiCES training set is a subset (80 hours) of the LibriSpeech dataset. This set has utterances from 427 speakers recorded in clean environments with close-talking microphones. The development and evaluation sets are far-field microphone recordings from diverse room dimensions, environments and noise conditions containing 19 and 20 hours of speech, respectively. The three sets namely training, development and evaluation, do not have any overlap in terms of the speakers. The robustness of the developed models is challenged by the mismatch that exists between the training and development/evaluation sets. We artificially added reverberation and noise on the 80 hours training set, which served as the training set for all the E2E ASR experiments on the VOiCES dataset. The development set contains 20 hours of distant recordings from the 200 speakers. The evaluation data of 19 hours consists of recordings 100 speakers. The training set has 22741 utterances, development set has 4318 utterances and evaluation set has 4600 utterances.

#### B. E2E ASR baseline system

For all the ASR experiments, we use the weighted prediction error based pre-processing [9] and unsupervised generalized eigenvalue (GEV) beamforming [13]. The baseline features are 36-dimensional log-mel filter bank features with frequency range from 200 Hz to 6500 Hz. The ESPnet toolkit [57] is used to perform all the end-to-end ASR experiments, with a Pytorch backend [58]. The model architecture uses 12 conformer encoder layers with 2048 units in the projection layer. The 6-layer transformer architecture with 2048 units in the projection layer serves as the decoder. Both connectionist temporal cost (CTC) loss and attention based cross entropy (CE) loss are used in the training, with CTC-weight set at 0.3 [59]. A single layer of 1000 LSTM cell recurrent neural network is used for language modeling (RNN-LM). For training the model, we use stochastic gradient descent (SGD) optimizer with a batch size of 32. For language model training, data is augmented from Wall Street Journal (WSJ) corpus.

<sup>1</sup>The implementation of the work can be found in <https://github.com/anurenjan/DFAR>

TABLE I

WER (%) ON THE REVERB DATASET FOR ENVELOPE/CARRIER, ENVELOPE-CARRIER DEREVERBERATION (DFAR) AND THE JOINT E2E-DFAR MODEL. THE RELATIVE IMPROVEMENTS (%) PERTAIN TO THE COMPARISON OF THE VARIOUS CONFIGURATIONS W.R.T. THE BF-FBANK BASELINE SYSTEM.

Model Config.	Dev			Eval		
	Real	Sim.	Avg. [Rel. Imp.]	Real	Sim	Avg. [Rel. Imp.]
BF-FBANK (baseline)	12.8	8.7	10.8 [- - -]	11.9	7.9	9.9 [- - -]
DCCRN [53] + BF-FBANK	17.4	10.4	13.9 [-28.7]	15.3	8.8	12.1 [-22.2]
Fullsubnet + BF-FBANK [54]	11.8	7.9	9.9 [+8.3]	10.5	7.4	9.0 [+9.1]
Deep non-linear filter [55] + BF-FBANK	12.4	8.1	10.3 [+4.6]	10.5	7.2	8.9 [+10.1]
Reverb. time shortening [56] + BF-FBANK	11.5	7.6	9.6 [+11.1]	10.1	7.6	8.9 [+10.1]
STFT Deverb. + BF-FBANK	12.0	7.8	9.9 [+8.3]	10.8	7.3	9.1 [+8.1]
Sub-band sig. Dereverb. + BF-FBANK	13.3	9.8	11.6 [-7.4]	12.8	8.6	10.7 [-8.1]
FDLP Env. Derevb. + BF-FBANK	12.7	8.5	10.6 [+1.9]	10.1	7.8	9.0 [+9.1]
FDLP Carr. Dereverb. + BF-FBANK	11.2	8.3	9.8 [+9.3]	10.8	7.6	9.2 [+7.1]
DFAR + BF-FBANK	10.6	7.6	9.1 [+15.7]	9.1	6.9	8.0 [+19.2]
E2E-DFAR	<b>9.4</b>	<b>6.4</b>	<b>7.9 [+26.9]</b>	<b>7.3</b>	<b>5.7</b>	<b>6.5 [+34.3]</b>

### C. Performance metrics

#### 1) ASR performance metrics:

- **WER/CER** (Word/Character Error Rate): The word/character error rate is given by the ratio of number of word/character insertions, deletions and substitutions in the system output to the total number of words/characters in the reference.

#### 2) Speech quality metrics:

- **SRMR**: Speech to reverberation modulation ratio (SRMR) is a non intrusive measure. Here, a representation is obtained using an auditory-inspired filter bank analysis of critical band temporal envelopes of the signal. The modulation spectral information is used to get an adaptive measure termed as speech to reverberation modulation energy ratio [60], [61]. A higher value indicates an improved quality of the given speech signal.
- **MOS** (Mean Opinion Score): To evaluate the performance of dereverberation algorithms, subjective quality and intelligibility measurement methods are needed. The most widely used subjective method is the ITU-T standard [62], where a panel of listeners are asked to rate the quality/intelligibility of the audio.

## V. EXPERIMENTS AND RESULTS

The baseline features are the beamformed log-mel filter-bank energy features (denoted as BF-FBANK).

### A. REVERB Challenge ASR

The word error rates (WER) for the dereverberation experiments are shown in Table I. Note that, all the experiments use the same input features (log-mel filter bank features) along with the same E2E ASR architecture (conformer encoder and transformer decoder). The only difference between the various rows, reported in Table I, is the dereverberation pre-processing applied on the raw audio waveform. All the dereverberation experiments use the DPLSTM architecture described in Sec. III.

TABLE II

COMPARISON OF THE RESULTS WITH OTHER WORKS REPORTED ON THE REVERB CHALLENGE DATASET.

System	Eval-sim.	Eval-real	Avg.
Subramanian et. al. [63]	6.6	10.6	8.6
Heymann et. al. [64]	-	10.8	-
Fujita et. al. [65]	<b>4.9</b>	9.8	7.4
Purushothaman et. al. [18]	7.1	12.1	9.6
Zhang et. al. [66]	-	10.0	-
This work	5.7	<b>7.3</b>	<b>6.5</b>

TABLE III

WER (%) IN REVERB DATASET FOR DIFFERENT ARCHITECTURES FOR THE DEREVERBERATION MODEL.

Model Config.	Dev			Eval		
	Real	Sim	Avg	Real	Sim	Avg
Baseline	12.8	8.7	10.8	11.9	7.9	9.9
CLSTM	14.5	9.7	12.1	12.4	9.1	10.8
4-layer LSTM	12.5	8.0	10.3	10.1	7.1	8.9
DPLSTM	<b>10.6</b>	<b>7.6</b>	<b>9.1</b>	<b>9.1</b>	<b>6.9</b>	<b>8.0</b>

1) *Various dereverberation configurations*: In Table I, the first row is the baseline result with the beamformed audio (unsupervised GEV beamforming [13] and weighted prediction error (WPE) processing [9].

The next set of rows compare several prior works.

- Fullsubnet - A full-band and sub-band fusion model for speech enhancement [54].
- DCCRN - Deep complex convolution recurrent neural network model for speech enhancement [53].
- Deep non-linear filter for multi-channel audio [55]
- Reverberation time shortening [56]

The prior works are trained on the same data settings as used in the DFAR framework. All the prior works, except DCCRN (which is not designed for ASR), improve the baseline system in range of 8-11% in terms of relative WER. However, the proposed DFAR/E2E-DFAR approach is observed to provide

TABLE IV  
WER (%) IN REVERB DATASET FOR HYPER PARAMETER  $\lambda$ , IN  
 $MSE\ loss = \lambda \times env.\ loss + (1 - \lambda) \times carr.\ loss$ .

Parameter $\lambda$	Dev			Eval		
	Real	Simu	Avg	Real	Simu	Avg
0	12	8.2	10.1	10.4	7.5	9.0
0.2	11.9	8.6	10.3	10.7	7.7	9.2
0.4	11.6	8.2	9.9	10.1	7.2	8.7
0.5	11.3	<b>7.2</b>	9.3	9.7	<b>6.5</b>	8.1
0.6	<b>10.6</b>	7.6	<b>9.1</b>	<b>9.1</b>	6.9	<b>8.0</b>
0.8	13.1	8.7	10.9	10.9	7.9	9.4
1	13.5	8.0	10.8	10.4	6.9	8.7

the best WER, with relative improvement in WER of 19/34% on the evaluation data.

In Table I, we have also performed two ASR experiments - i) using STFT inputs (log magnitude), and ii) using the sub-band signal directly without the envelope-carrier decomposition. Both these experiments, use the DPLSTM dereverberation model proposed in this work. As seen in Table I, the dereverberation on the STFT magnitude component improves the ASR systems significantly over the baseline, while the dereverberation on the sub-band signal directly is not effective. However, the STFT approach is also seen to be inferior to the DFAR approach where the envelope-carrier dereverberation is performed.

The fourth set of rows corresponds to the WER results with envelope/carrier based dereverberation alone. The relative improvements of 2 – 9% are seen here compared to the baseline BF-FBANK. Separately, with dereverberation based on the carrier signal alone, a similar improvement is achieved. Further, the dereverberation of the temporal envelope and carrier components in a combined fashion using the DPLSTM model improves the ASR results over the separate dereverberation of envelope/carrier components. Here, average relative improvements of 16% and 19% are seen in the development set and evaluation set respectively, over the BF-FBANK baseline system for the DFAR approach.

The final row in Table I reports the results using the joint learning of the dereverberation network and the E2E ASR model. The E2E-DFAR is initialized using the dereverberation model and the E2E model trained separately. The proposed E2E-DFAR model yields average relative improvements of 27% and 34% on the development set and evaluation set respectively over the baseline system. The joint training is also shown to improve over the set up of having separate networks for dereverberation and E2E ASR. While the DFAR model is trained only on simulated reverberation conditions, the WER improvement in real condition is seen to be more pronounced than those observed in the simulated data. This indicates that the model can generalize well to unseen reverberation conditions in the real-world.

2) *Comparison with prior works*: The comparison of the results from prior works reported on the REVERB challenge dataset is given in Table II. The Table includes results from end-to-end ASR systems [63], [65], [66] as well as the joint enhancement and ASR modeling work reported in [64]. We

TABLE V  
PERFORMANCE (WER %) ON THE VOICES DATASET.

Model Config.	Dev	Eval
FBANK (baseline)	40.3	50.8
+ Env. derevb.	38.4	48.6
+ Env.-carr. derevb. (DFAR)	37.1	45.4
+ E2E-DFAR	<b>36.4</b>	<b>44.7</b>

also compare with our prior work reported in [18]. Specifically, many of the prior works compared in Table II are based on STFT based enhancement. The work reported in Subramanian et. al. [63], used a neural beamforming approach in the STFT domain, while the efforts described in Heymann et. al. [64], used a long-short term memory network for mask estimation in power spectral domain (PSD). The dynamic convolution method proposed in Fujita et. al. [65] used deconvolution of log-mel spectrogram features. Similar to the proposed work, all these efforts have also used the E2E ASR model training. As seen in Table II, the proposed work improves over these prior works considered here, further highlighting the benefits of the dereverberation in the sub-band time domain using long-term envelope-carrier based DPLSTM models.

3) *Dereverberation model architecture*: The ASR experiments on the REVERB challenge dataset, pertaining to the choice of different model architectures used in the dereverberation model, are listed in Table III. We have experimented with convolutional LSTM (CLSTM) [50] and time-domain LSTM (4-layer LSTM) architecture [67] in addition to the DPLSTM approach. As seen here, the Dual-path recurrence based DPLSTM gives the best word error rate in comparison with the other LSTM neural architectures considered. This may be attributed to the joint time-frequency recurrence performed to the other approaches which perform only time domain recurrence.

4) *Dereverberation loss function*: The MSE loss function used in the DPLSTM model training consists of a combination of loss values from the envelope and the carrier components. We experimented with the hyper parameter,  $\lambda$ , which controls the proportion of envelope based loss and carrier based loss in the total loss ( $Total\ loss = \lambda \times env.\ loss + (1 - \lambda) \times carr.\ loss$ ). The ASR results for the various choices of the hyper parameter  $\lambda$  are shown in Table IV. Empirically, the value of  $\lambda = 0.6$  gives the best WER on the REVERB challenge dataset. Further, the choice of  $\lambda = 1$  or  $\lambda = 0$ , corresponding to envelope/carrier only dereverberation, are inferior to other choices of  $\lambda$ , indicating that the joint dereverberation of the envelope and carrier components is beneficial.

## B. VOICES ASR

The ASR setup used in the VOICES dataset followed the ESPnet recipe with the conformer encoder and a transformer decoder. The rest of the model parameters and hyperparameters are kept similar to the ones in the REVERB challenge dataset. The WER results on the VOICES dataset are given in Table V. The dereverberation of the envelope alone provides an absolute improvement of 1.9% and 2.2% on the development and evaluation data respectively, compared



TABLE VI  
SRMR VALUES ON THE REVERB DATASET FOR VARIOUS SIGNAL ENHANCEMENT STRATEGIES.

Signal	SRMR				
	Dev. (Real)	Dev. (Sim.)	Eval. (Real)	Eval. (Sim.)	REVB. (Train)
Unsupervised GEV beamforming [13]	5.18	4.1	4.58	4.67	4.23
+ WPE [9]	5.35	4.2	4.61	4.75	4.48
+ DCCRN [53]	5.43	4.37	4.63	4.94	4.67
+ Fullsubnet [54]	5.36	4.32	4.64	4.97	4.63
+ Deep Non-Linear Filters [55]	5.51	4.22	4.64	5.02	4.61
+ Reverberation Time Shortening Target [56]	5.49	<b>4.57</b>	4.62	5.2	4.58
+ STFT Mag. + DPLSTM	5.44	4.33	4.64	4.94	4.6
+ Sub-band signal + DPLSTM	5.45	4.28	4.61	4.87	4.63
+ env. derevb. (this work)	4.62	3.83	4.12	4.25	4.11
+ crr. derevb. (this work)	5.52	4.46	4.69	5.27	4.77
+ env. & crr. derevb. [DFAR] (this work)	<b>5.52</b>	4.47	<b>4.69</b>	<b>5.27</b>	<b>4.77</b>

TABLE VII  
MOS VALUES IN REVERB DATASET FOR ENVELOPE AND CARRIER BASED ENHANCEMENTS.

	ET Real - near	ET Real - far	ET Simu - near	ET Simu - far
Baseline - GEV [13] + WPE [9]	3.78	3.65	3.74	4.12
+ env.-carr. derevb. [DFAR] (this work)	<b>3.98</b>	<b>3.67</b>	<b>4.01</b>	<b>4.40</b>

to the FBANK baseline system. The dereverberation based on envelope-carrier modeling further improves the results. An absolute improvement of 3.3%/5.4% on the development/evaluation data is achieved, compared to the FBANK baseline. Further, the joint training on envelope-carrier dereverberation network with the ASR model improves the WER results. We observe relative improvements of 10% and 12% on the development and evaluation data respectively .

### C. Speech quality evaluation

A comparison of the SRMR values for different dereverberation approaches is reported in Table VI. Here, we compare the baseline unsupervised GEV beamforming [13] and weighted prediction error (WPE) [9] with various strategies for beamforming. The deep complex convolutional recurrent network (DCCRN) based speech enhancement [53] is also implemented on the REVERB dataset, and these results are reported in Table VI. While the envelope based dereverberation did not improve the SRMR values, the carrier based dereverberation is shown to improve the SRMR results. Further, the DFAR model also achieves similar improvements in SRMR for all the conditions over the baseline approach (GEV+WPE) and the DCCRN approach.

We conducted a subjective evaluation to further assess the performance of the dereverberation method. The subjects were asked to rate the quality of the audio on a scale of 1 to 5, 1 being poor and 5 being excellent. The subjects listened to the audio in a relatively quiet room with a high quality Sennheiser headset. We perform the A-B listening test, where the two versions of the same audio file were played, the first one with GEV + WPE dereverberation and the second one with the proposed dereverberation approach. We chose 20 audio samples, from four different conditions (real and simulated data and from near and far rooms) for this evaluation and recruited 20 subjects.

The subjective results are shown in Table VII. As seen, the proposed speech dereverberation scheme shows improvement in subjective MOS scores for all the conditions considered. The subjective results validate the signal quality improvements observed in the SRMR values (Table VI).

## VI. CONCLUSION

In this paper, we propose a speech dereverberation model using frequency domain linear prediction based sub-band envelope-carrier decomposition. The sub-band envelope and carrier components are processed through a dereverberation network. A novel neural architecture, based on dual path recurrence, is proposed for dereverberation. Using the joint learning of the neural speech dereverberation module and the E2E ASR model, we perform several speech recognition experiments on the REVERB challenge dataset as well as on the VOICES dataset. These results show that the proposed approach improves over the state of art E2E ASR systems based on mel filterbank features.

The dereverberation approach proposed in this paper also reconstructs the audio signal, which makes it useful for audio quality improvement applications as well as other speech processing systems in addition to the ASR system. We have further evaluated the reconstruction quality subjectively and objectively on the REVERB challenge dataset. The quality measurements show that the proposed speech dereverberation method improves speech quality over the baseline framework of weighted prediction error. The ablation studies on various architecture choices provides justification for the choice of the DPLSTM network architecture. Given that the proposed model allows the reconstruction of the audio signal, it can be used in conjunction with self-supervised neural approaches for representation learning of speech as well. This will form part of our future investigation.

## REFERENCES

- [1] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, "Far-field automatic speech recognition," *Proceedings of the IEEE*, vol. 109, no. 2, pp. 124–148, 2020.
- [2] T. Hain, L. Burget, J. Dines, P. N. Garner, F. Grézl, A. El Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, "Transcribing meetings with the AMIDA systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 486–498, 2012.
- [3] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low latency acoustic modeling using temporal convolution and LSTMs," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 373–377, 2018.
- [4] S. Ganapathy and V. Peddinti, "3-d CNN models for far-field multi-channel speech recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5499–5503.
- [5] A. Gusev, V. Volokhov, T. Andzhukhaev, S. Novoselov, G. Lavrentyeva, M. Volkova, A. Gazizullina, A. Shulipa, A. Gorlanov, A. Avdeeva et al., "Deep speaker embeddings for far-field speaker recognition on short utterances," *arXiv preprint arXiv:2002.06033*, 2020.
- [6] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2023–2032, 2007.
- [7] T. Yoshioka et al., "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.
- [8] S. Ganapathy, "Signal analysis using autoregressive models of amplitude modulation," Ph.D. dissertation, Johns Hopkins University, 2012.
- [9] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE TASLP*, vol. 18, no. 7, pp. 1717–1731, 2010.
- [10] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM transactions on Audio, Speech, and Language processing*, vol. 25, no. 7, pp. 1492–1501, 2017.
- [11] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.
- [12] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 5, pp. 1529–1539, 2007.
- [13] R. Kumar, A. Sreeram, A. Purushothaman, and S. Ganapathy, "Unsupervised neural mask estimator for generalized eigen-value beamforming based ASR," in *IEEE ICASSP*, 2020, pp. 7494–7498.
- [14] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7398–7402.
- [15] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low latency acoustic modeling using temporal convolution and LSTMs," *IEEE Signal Processing Letters*, vol. 25, issue 3, pp. 373–377, 2017.
- [16] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 546–555, 2009.
- [17] Y. Avargel and I. Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE transactions on Audio, Speech, and Language processing*, vol. 15, no. 4, pp. 1305–1319, 2007.
- [18] A. Purushothaman, A. Sreeram, R. Kumar, and S. Ganapathy, "Dereverberation of autoregressive envelopes for far-field speech recognition," *Computer Speech & Language*, vol. 72, p. 101277, 2022.
- [19] A. Purushothaman, A. Sreeram, and S. Ganapathy, "3-D acoustic modeling for far-field multi-channel speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6964–6968.
- [20] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 46–50.
- [21] K. Kinoshita et al., "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *IEEE WASPAA*, 2013, pp. 1–4.
- [22] C. Richey, M. Barrios et al., "VOICES obscured in complex environmental settings (voices) corpus," *arXiv preprint arXiv:1804.05053*, 2018.
- [23] M. Nandwana, J. Van Hout, M. McLaren, C. Richey, A. Lawson, and M. Barrios, "The voices from a distance challenge 2019 evaluation plan," *arXiv preprint arXiv:1902.10828*, 2019.
- [24] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.
- [25] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.
- [26] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Late reverberation suppression using recurrent neural networks with long short-term memory," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5434–5438.
- [27] K. Han, Y. Wang, and D. Wang, "Learning spectral mapping for speech dereverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4628–4632.
- [28] A. Pandey and D. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.
- [29] M. Wöllmer, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, "Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6822–6826.
- [30] Z. Chen, S. Watanabe, H. Erdogan, and J. R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [31] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.
- [32] A. L. Maas, T. M. O'Neil, A. Y. Hannun, and A. Y. Ng, "Recurrent neural network feature enhancement: The 2nd chime challenge," in *Proceedings The 2nd CHiME Workshop on Machine Listening in Multisource Environments held in conjunction with ICASSP*, 2013, pp. 79–80.
- [33] J. F. Santos and T. H. Falk, "Speech dereverberation with context-aware recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1236–1246, 2018.
- [34] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on antennas and propagation*, vol. 30, no. 1, pp. 27–34, 1982.
- [35] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.
- [36] G. Li, S. Liang, S. Nie, W. Liu, and Z. Yang, "Deep neural network-based generalized sidelobe canceller for dual-channel far-field speech recognition," *Neural Networks*, vol. 141, pp. 225–237, 2021.
- [37] F.-J. Chang, M. Radfar, A. Mouchtaris, B. King, and S. Kunzmann, "End-to-end multi-channel transformer for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5884–5888.
- [38] C. Kim, A. Garg, D. Gowda, S. Mun, and C. Han, "Streaming end-to-end speech recognition with jointly trained neural feature enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6773–6777.
- [39] Z.-Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 796–806, 2016.
- [40] B. Wu, K. Li, F. Ge, Z. Huang, M. Yang, S. M. Siniscalchi, and C. Lee, "An end-to-end deep learning approach to simultaneous speech dereverberation and acoustic modeling for robust speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1289–1300, 2017.
- [41] P. Vaidyanathan, "Quadrature mirror filter banks, m-band extensions and perfect-reconstruction techniques," *IEEE Assp Magazine*, vol. 4, no. 3, pp. 4–20, 1987.
- [42] P. P. Vaidyanathan, *Multirate systems and filter banks*. Pearson Education India, 2006.
- [43] P. Motlicek, S. Ganapathy, H. Hermansky, and H. Garudadri, "Scalable wide-band audio codec based on frequency domain linear prediction," IDIAP, Tech. Rep., 2007.
- [44] M. Athineos and D. P. Ellis, "Frequency-domain linear prediction for temporal features," 2003.

- [45] J. P. Burg, *Maximum entropy spectral analysis*. Stanford University, 1975.
- [46] S. Ganapathy, P. Motlicek, and H. Hermansky, "Autoregressive models of amplitude modulations in audio compression," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 6, pp. 1624–1631, 2009.
- [47] S. Ganapathy, S. H. Mallidi, and H. Hermansky, "Robust feature extraction using modulation filtering of autoregressive models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 8, pp. 1285–1295, 2014.
- [48] S. Thomas, S. Ganapathy, and H. Hermansky, "Recognition of reverberant speech using frequency domain linear prediction," *IEEE Signal Processing Letters*, vol. 15, pp. 681–684, 2008.
- [49] S. Ganapathy, J. Pelecanos, and M. K. Omar, "Feature normalization for speaker verification in room reverberation," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4836–4839.
- [50] R. Kumar, A. Purushothaman, A. Sreeram, and S. Ganapathy, "End-to-end speech recognition with joint dereverberation of sub-band autoregressive envelopes," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6057–6061.
- [51] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.
- [52] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005, pp. 357–362.
- [53] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," *Proceedings of Interspeech*, 2020.
- [54] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6633–6637.
- [55] K. Tesch and T. Gerkmann, "Insights into deep non-linear filters for improved multi-channel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 563–575, 2022.
- [56] R. Zhou, W. Zhu, and X. Li, "Speech dereverberation with a reverberation time shortening target," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [57] S. Watanabe, T. Hori *et al.*, "ESPNNet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.
- [58] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS-W*, 2017.
- [59] S. Karita, N. Chen *et al.*, "A comparative study on transformer vs RNN in speech applications," in *IEEE ASRU*, 2019, pp. 449–456.
- [60] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.
- [61] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 55–59.
- [62] P. RECOMMENDATION, "Itu-tp. 808," 2018.
- [63] A. Subramanian, X. Wang, S. Watanabe, T. Taniguchi, D. Tran, and Y. Fujita, "An investigation of end-to-end multichannel speech recognition for reverberant and mismatch conditions," *arXiv preprint arXiv:1904.09049*, 2019.
- [64] J. Heymann, L. Drude, R. Haeb-Umbach, K. Kinoshita, and T. Nakatani, "Joint optimization of neural network-based wpe dereverberation and acoustic model for robust online ASR." *IEEE ICASSP*, 2019, pp. 6655–6659.
- [65] Y. Fujita, A. Subramanian, M. Omachi, and S. Watanabe, "Attention-based asr with lightweight and dynamic convolutions," in *ICASSP*. IEEE, 2020, pp. 7034–7038.
- [66] W. Zhang, A. Subramanian, X. Chang, S. Watanabe, and Y. Qian, "End-to-end far-field speech recognition with unified dereverberation and beamforming," *arXiv preprint arXiv:2005.10479*, 2020.
- [67] M. Mimura, S. Sakai, and T. Kawahara, "Speech dereverberation using long short-term memory," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.



from Govt. Engineering College, Barton Hill, Trivandrum in 2006, the Master's degree in Signal Processing from College of Engineering, Trivandrum in 2008. In March 2008 he joined the Department of ECE, College of Engineering, Trivandrum as Assistant Professor.



Technology, Silchar in 2018.



doctoral studies, he worked as a research fellow in the Learning and Extraction of Acoustic Patterns (LEAP) lab at the Indian Institute of Science, Bangalore, India.



Technology from College of Engineering, Trivandrum, India and Master of Engineering from the Indian Institute of Science, Bangalore. He has also worked as a Research Assistant in Idiap Research Institute, Switzerland from 2006 to 2008. At the LEAP lab, his research interests include signal processing, machine learning methodologies for speech and speaker recognition and auditory neuro-science. He is a subject editor for the Speech Communications journal and a senior member of the IEEE.

**Anurenjan Purushothaman** is associated with the Learning and Extraction of Acoustic Patterns (LEAP) lab, Department of Electrical Engineering, Indian Institute of Science, Bangalore, India, 56001, where he is pursuing his PhD under the guidance of Prof. Sriram Ganapathy. He is also associated with College of Engineering, Trivandrum and Government Engineering College, Idukki, where he is an Assistant Professor in the Department of Electronics & Communication. He received the Bachelor's degree in Electronics & Communication Engineering from Govt. Engineering College, Barton Hill, Trivandrum in 2006, the Master's degree in Signal Processing from College of Engineering, Trivandrum in 2008. In March 2008 he joined the Department of ECE, College of Engineering, Trivandrum as Assistant Professor.

**Debottam Dutta** is currently a PhD student at the Signals & Inference Research Group (SiNRG) at University of Illinois at Urbana-Champaign (UIUC). Prior to joining UIUC, he worked as a Senior Research Fellow for a year at the Learning and Extraction of Acoustic Patterns (LEAP) lab, Indian Institute of Science, Bangalore. He obtained his Master of Technology degree in Signal Processing from Indian Institute of Science, Bangalore in 2021 and Bachelor's degree in Electronics and Communication Engineering from National Institute of

**Rohit Kumar** is currently associated with the Laboratory for Computational Audio Perception (LCAP) at Johns Hopkins University, Baltimore, Maryland. He is pursuing his PhD under the guidance of Prof. Mounya Elhilali. Rohit earned his Bachelor's degree in Electronics & Communication Engineering from Delhi Technological University (formerly Delhi College of Engineering) in 2017 and his Master's degree in Signal Processing from the Indian Institute of Science, Bangalore, India, in 2020, under the supervision of Prof. Sriram Ganapathy. Prior to his

**Sriram Ganapathy** is an Associate Professor at the Electrical Engineering, Indian Institute of Science, Bangalore, where he heads the activities of the learning and extraction of acoustic patterns (LEAP) lab. He is also associated with the Google Research India, Bangalore. Prior to joining the Indian Institute of Science, he was a research staff member at the IBM Watson Research Center, Yorktown Heights. He received his Doctor of Philosophy from the Center for Language and Speech Processing, Johns Hopkins University. He obtained his Bachelor of

ACCEPTED MANUSCRIPT

# Strategies in design of self-propelling hybrid micro/nanobots for bioengineering applications

To cite this article before publication: Saurabh Shivalkar *et al* 2023 *Biomed. Mater.* in press <https://doi.org/10.1088/1748-605X/acf975>

## Manuscript version: Accepted Manuscript

Accepted Manuscript is “the version of the article accepted for publication including all changes made as a result of the peer review process, and which may also include the addition to the article by IOP Publishing of a header, an article ID, a cover sheet and/or an ‘Accepted Manuscript’ watermark, but excluding any other editing, typesetting or other changes made by IOP Publishing and/or its licensors”

This Accepted Manuscript is © 2023 IOP Publishing Ltd.



During the embargo period (the 12 month period from the publication of the Version of Record of this article), the Accepted Manuscript is fully protected by copyright and cannot be reused or reposted elsewhere.

As the Version of Record of this article is going to be / has been published on a subscription basis, this Accepted Manuscript will be available for reuse under a CC BY-NC-ND 3.0 licence after the 12 month embargo period.

After the embargo period, everyone is permitted to use copy and redistribute this article for non-commercial purposes only, provided that they adhere to all the terms of the licence <https://creativecommons.org/licences/by-nc-nd/3.0>

Although reasonable endeavours have been taken to obtain all necessary permissions from third parties to include their copyrighted content within this article, their full citation and copyright line may not be present in this Accepted Manuscript version. Before using any content from this article, please refer to the Version of Record on IOPscience once published for full citation and copyright details, as permissions may be required. All third party content is fully copyright protected, unless specifically stated otherwise in the figure caption in the Version of Record.

View the [article online](#) for updates and enhancements.

# Strategies in Design of Self-Propelling Hybrid Micro/Nanobots for Bioengineering Applications

Saurabh Shivalkar<sup>1#\*</sup>, Anwesha Roy<sup>2#</sup>, Shrutika Chaudhary<sup>3#</sup>, Sintu Kumar Samanta<sup>1</sup>, Pallabi Chowdhary<sup>4</sup>, Amaresh Kumar Sahoo<sup>1\*</sup>

<sup>1</sup>Department of Applied Sciences, Indian Institute of Information Technology, Allahabad, UP India.  
<sup>2</sup>Department of Biotechnology, Heritage Institute of Technology, Kolkata, West Bengal, India.  
<sup>3</sup>Department of Biotechnology, Delhi Technological University, Delhi, India.  
<sup>4</sup>Department of Biotechnology, M.S. Ramaiah University of Applied Sciences, Bengaluru, Karnataka, India.  
#Equal Contribution

E-mail: [asahoo@iiita.ac.in](mailto:asahoo@iiita.ac.in)  
E-mail: [rss2019001@iiita.ac.in](mailto:rss2019001@iiita.ac.in)

Received xxxxxx  
Accepted for publication xxxxxx  
Published xxxxxx

## Abstract

Micro/nanobots are integrated devices developed from engineered nanomaterials that have evolved significantly over the past decades. They can potentially be pre-programmed to operate robustly at numerous hard-to-reach organ/tissues/cellular sites for multiple bioengineering applications such as early disease diagnosis, precision surgeries, targeted drug delivery, cancer therapeutics, bio-imaging, biomolecules isolation, detoxification, bio-sensing, and clearing up clogged arteries with high soaring effectiveness and minimal exhaustion of power. Several techniques have been introduced in recent years to develop programmable, biocompatible, and energy-efficient micro/nanobots. Therefore, the primary focus of most of these techniques is to develop hybrid micro/nanobots that are an optimized combination of purely synthetic or biodegradable bots suitable for the execution of user-defined tasks more precisely and efficiently. Recent progress has been illustrated here as an overview of a few of the achievable construction principles to be used to make biomedical micro/nanobots and explores the pivotal ventures of nanotechnology-moderated development of catalytic autonomous bots. Furthermore, it is also foregrounding their advancement offering an insight into the recent trends and subsequent prospects, opportunities, and challenges involved in the accomplishments of the effective multifarious bioengineering applications.

**Keywords:** Nanomotors; Micromotors; Bioengineering; Drug Delivery; Biosensing; Biomedical engineering

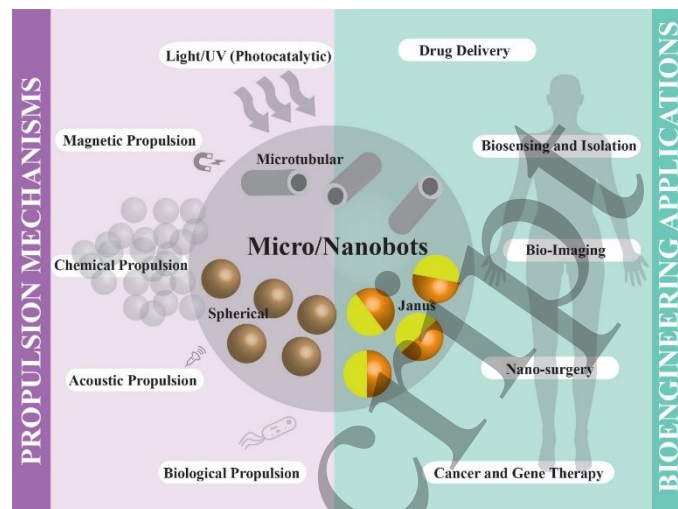
## 1. Introduction

As we are getting a deeper understanding of biological systems, interest in the behaviour of nanoscale materials in the biological environment has grown exponentially. Nanoscale materials are engineered to design better and biocompatible integrated systems for the diagnosis and cure of several diseases. Micro/nanobots are one remarkable example of engineered nanoscale materials that have huge prospects in the field of biomedical sciences. Their assembly is based on molecular nanotechnology and mechano-synthetic chemistry that enabled them to move autonomously within various complex mediums, including biological environments [1,2,3]. These miniature devices (ranging from micro to nanoscale sizes) are intriguing designs of biological systems and are potentially being self-propelled to overcome low Reynolds number, viscous drag, and Brownian motion in liquid medium under physical or chemical stimulation. The extent of research in this field is dedicatedly objectified to mimic natural mechanisms to get better efficiency and performance. Interestingly, these could be able to perform several challenging tasks much more accurately and with high precision such as cell penetration, molecular delivery, bioremediation, and removal of environmental contaminants. Micro/nanobots can be characterized mainly according to the



sources of energy input (fuel) required for their propulsion [4]. The sources may be chemical stimuli like pH, catalysis, or diffusiophoresis, and physical stimuli such as magnetic field, light, ultrasonic wave, electrical field, etc. The performance of the micro/nanobots greatly relies on material properties, uniqueness in design, and engineering at a molecular level that involves the amalgamation of several disciplines such as nanotechnology, biochemistry, fluid mechanics, and material sciences [5,6]. Over the few past decades, emphasis on this research field is strongly made to obtain proof of concept and development of understanding of parameters associated with the physicochemical properties of the bots. Opportunities for surface functionalization and modifications provide a huge scope for constructing micro/nanobots having better targeting ability [4,5,18]. Moreover, advancements in material design and processing allowed researchers to construct bots that offer less toxicity by using biocompatible energy or fuel sources. Thus, it is to be deemed that recently developed integrated systems like micro/nanobots might offer an unprecedented scope of use for various biomedical applications.

There have been several interesting and potential advancements occurred in the last few years in various aspects of the micro/nanobots with different motion mechanism and their potential applications in the field of bioengineering are discussed herein in Scheme 1. For example, the nano/microstructure of these bots allows easy administration inside the body, making it the most convenient cohesive system for the delivery of the therapeutics at the programmed site and employing it for accurate diagnosis purposes [7]. The movement at the nano or microscale comes with its challenges, however, that has been addressed more precisely by several recent studies. This scaling down leads to significant changes in the factors such as the effect of viscosity and Brownian diffusion. Thus, making the motion of these nanobots more challenging particularly within complex biological mediums. Nevertheless, smart design strategies of these micro/nanobots are expeditiously increasing in the last few years that offer a better performance index. Furthermore, significant attention is given to the key challenges in the field of biomedical such as targeted delivery of drugs/genes to the tumour cells, high precision surgeries and biopsies of cardiac vascular clogging, and sensing the intrinsic physiological and physiochemical changes near the diseased sites. Thus, recent progress in the area of design of micro/nanobots in a broad range of bioengineering applications like imaging, efficient diagnosis, drug delivery, nano-surgery, etc., are elaborately discussed here. Additionally, available nanoscale materials for 'smart' hybrid micro/nanobots, which are indeed an optimized combination of synthetic or biopolymers are described systematically to explore the standardization features of the design of biocompatible micro/nanobots for next-generation bioengineering applications [8].



**Scheme 1:** Schematics of propulsion mechanisms and potential bioengineering applications of micro/nanobots.

## 2. Design strategies and propulsion mechanism of micro/nanobots

To design realistic micro/nanobots, they must have the potential to transform energy into motion. Therefore, various synthesis methods were explored to obtain desired propulsion in the bots. Different strategies in the development of the bots resulted in micro/nanobots with different characteristics and features such as shape size and propulsion mechanisms. Therefore, different names are given to these micro/nanobots by different research groups. Frequently came across names of these micro/nanobots are microjets, nanorockets, micropropellers, microrobots, nanotrain, microhelices, micro or nano-swimmers, self-propelling microsphere, micro or nano engine etc [9-14]. These design strategies paved the pathways for developing application-specific micro/nanobots. One of the key features among all types of micro/nanobots is their ability to move autonomously and perform tasks simultaneously [4,5,18].

Being extremely small-scale objects, the autonomous movement of Micro/nanobots is generally influenced by various physical forces that are not as significant at larger scales. Changes in physical factors like viscosity, drag force, and Brownian motion possess a huge impact on the motion of micro/nanobots. Moreover, a change in Reynolds Number ( $Re$ ) which is a ratio of the inertial and viscous force of fluids also creates a determinant effect on the propulsion of the micro/nanobots. The equation of the  $Re$  of fluids is given below.

$$Re = \frac{\text{Inertial Force}}{\text{Viscous Drag Force}} = \frac{\rho v l}{\eta} \quad (2.1)$$

Where  $\rho$  is the density of the fluid,  $v$  is the velocity of the nanobots;  $l$  is the length of the nanobot and  $\eta$  is the dynamic viscosity of the fluid. Generally,  $Re$  of any system is



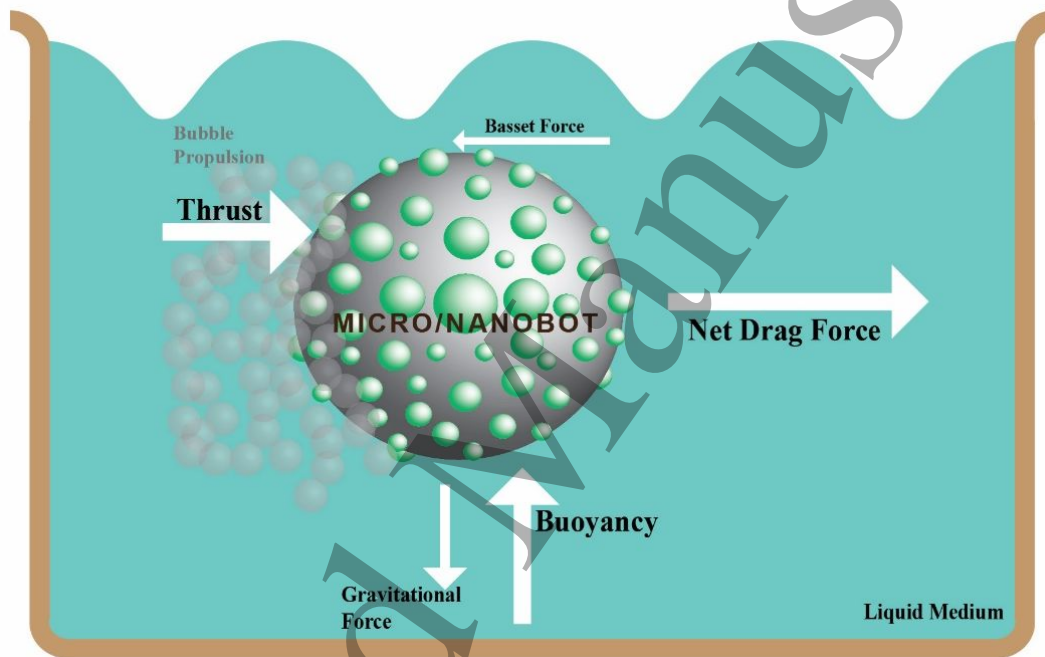
designed in such a way where inertial forces dominate over viscous drag forces but at the micro or nanoscale,  $Re$  becomes very low due to dominating effect of viscous drag forces over the inertial forces. Edward Millis Purcell has explained this concept in the 'Scallop Theorem' that stated life at low Reynold Number [15].

On the other hand, at the micro or nanoscale Brownian diffusion also interferes with the direction of the movement of micro/nanobots. This is the reason why most of the micro/nanobots developed to this date produce random motion and frequent reorientation if not physically guided [4,5,8,17]. Brownian diffusion is characterized by the diffusion coefficient ( $D$ ) that is associated with the thermal energy and the size of the particles present in the micro or nanoscale environment. Therefore, for controlling the direction and the

motion of the micro/nanobots, Brownian diffusion is very crucial in the design and development of these bots [16]. Brownian diffusion is defined by Stoke- Einstein relationship given below.

$$D = \frac{K_b T}{6\pi\eta r} \quad (2.2)$$

Where,  $K_b$ ,  $T$ ,  $\eta$ , and  $r$  are Boltzmann constant, absolute temperature, viscosity, a radius of particles respectively. The movement of the micro/nanobots is generally due to the combination of forces or resultant forces of various forces gravitational, viscous, buoyancy, and hydrodynamic forces (Fig. 1). The resultant force referred to is known as Effective Drag Force ( $F_v$ ) [17].



**Fig. 1:** Illustration of forces acting on the micro/nanobot moving in a liquid medium using a bubble propulsion mechanism. The arrow and their thickness represent the direction in which force is applied and the strength of the force respectively.

The drag force acting on the micro/nanobots is effective because of the combination of three forces. These three forces or the total drag force act oppositely to the micro/nanobots' direction under steady-state conditions ( $F_d$ ), the second force is the mass component where the resistance exerted by the accelerated sphere is equivalent to the magnitude in irrotational motion ( $F_2$ ), and last one is the basset force acting on the surface of the micro/nanobots during their motion in the solution ( $F_3$ ). By using the total drag force,  $F_v$  has been calculated and given by equation 2.4 [17].

$$F_v = F_1 + F_2 + F_3 \quad (2.3)$$

$$F_v = 6\pi\eta Rv + \frac{2}{3}\pi R^3\rho a + 6R^2\sqrt{\pi\eta\rho} a \int_0^t \frac{dt'}{\sqrt{t-t'}} \quad (2.4)$$

To derive such autonomous movement, micro/nanobots were exposed to chemical fuel or physical techniques (Table 1).

Various propulsion mechanisms are described here:

### 2.1. Chemical fuel based

This mechanism of propulsion is also known as the onboard method; where micro/nanobots are independent of any external energy or power sources for the movement. Therefore, they are dependent on chemical fuel for their autonomous movement. Chemically driven micro/nanobots are the most common and extensively explored miniaturized devices that move autonomously due to the catalytic reactions in the fluids. For example, surface catalysis of hydrogen peroxide ( $\text{H}_2\text{O}_2$ ) releasing oxygen ( $\text{O}_2$ ) in a liquid phase is the key principle of propulsion (Fig. 2a) [18,19,20]. Bots developed from nanoscale materials either in the nano or micro range essentially possessed a high potential to drift along with biological fluids thus, making them autonomous and capable of performing complex tasks. Several experiments and studies have been carried out to obtain remarkable speed from thrust produced by oxygen bubbles. However, controlling the speed is another important aspect for applications of catalytic micro/nanobots in biomedical sciences [21]. The speed of chemically driven bots was controlled by altering the concentration, pH, and temperature of chemical fuels [21].  $\text{H}_2\text{O}_2$  is frequently used as fuel. However, when the objective is kept within the boundaries of biocompatibility and low toxicity to biological systems several other types of fuels have also been explored. Biologically available glucose or urea and sometimes a combination of acid or base have also been reported as biocompatible fuels [22]. Other than catalysis, micro/nanobots were driven using chemical gradients or electrochemical gradients, but these mechanisms were not as prevalent as catalysis due to their application-based challenges. For example, the self-electrophoresis (electrochemical gradient) mechanism is incompatible with biological fluids of high ionic strength. For the directional movement, chemically powered micro/nanobots are based on two methods. In the first approach, ceaseless bubble development in the constricted chamber of microbots following the jet-like ejection through a nozzle drove the long-range directional movement [19]. In a second way, local chemical gradients were created surrounding the microbots causing a self-phoretic thrust force [19]. Efforts have consistently been made in this field to obtain the best architectural design for its self-propulsion, controlling motion accurately, and understanding its motion mechanism.

**Bubble propulsion through catalysis:** Bubble propulsion through catalysis of micro/nanobots is the most commonly explored mechanism. Several research are focusing on different designs and shapes such as nanorods, tubular, and spiral functions through bubble propulsion [4,23,24]. Spontaneous decompositions of  $\text{H}_2\text{O}_2$  into  $\text{H}_2\text{O}$  and  $\text{O}_2$  in the solution-born catalytic micro/nanobots surface lead to the formation of gas bubbles. This in turn drives the bots and provides motion in some random direction mostly away from

the catalyst. The most extensively investigated catalyst for the decomposition of  $\text{H}_2\text{O}_2$  in the bubble propulsion mechanism is platinum (Pt). However, platinum was found to be toxic to the biological system, thus its use to catalyze these bots for biomedical applications is restricted [25]. Attempts have been made consistently in the past few years to design less toxic or biocompatible micro/nanobots. In this line of interest, the development of Pt-free micro/nanobots, several hybrid composite materials have been proposed such as silver (Ag)/manganese dioxide ( $\text{MnO}_2$ ) and magnesium (Mg)/aluminium (Al) has also been introduced. The Ag and  $\text{MnO}_2$  nanoparticles-based bots have excellent catalytic properties thus efficiently decomposing  $\text{H}_2\text{O}_2$  into oxygen bubbles leading to fast propulsion of the bots. Speed could easily be regulated by changing the concentration of hydrogen peroxide in the solution. Even at the lowest concentration of 0.1% of  $\text{H}_2\text{O}_2$ , efficient propulsion can be achieved. In comparison to Mg/Al, the Ag/ $\text{MnO}_2$  bots have high-efficiency low manufacturing cost and excellent stability, and prolonged bubble propulsion [25]. This was a better alternative to Pt-based micro/nanobots for diverse practical applications. Although, a few additional functionalization and modification of Ag/ $\text{MnO}_2$  nanobots were still needed to be done, to inculcate their biocompatibility and incorporate them into bioengineering applications.

The catalytic hydrogel-based soft vehicle is another example of a bubble propulsion bot developed by injection loading method having very low consumption of energy. Key factors such as propulsion ability and reusability necessary for practicability were also explored at their best. Rapid bubble formation from the decomposition of  $\text{H}_2\text{O}_2$  in the aqueous solution endowed the efficient propulsion of the bot. An astounding speed of 3.84 mm/s in 10% (w/w) of  $\text{H}_2\text{O}_2$  solution was reported [26]. Since it was a hydrogel-based bot, it had the remarkable ability of loading, which can be practically six times before it degrades/diminishes. This prototype was further modulated by making a template of oil/water emulsion into hydrogel soft bot to obtain an improved speed of 4.33 mm/s while keeping other parameters the same. This was due to the low boiling oil phase of emulsion hydrogels conjugate liberating high catalytic reaction heat causing higher bubble formation and propulsion. These bots possess high stability, better loading capability, reusability, and good speed which are essential for efficient drug delivery. As the  $\text{H}_2\text{O}_2$  was concentration increased, the speed of the bot also increased due to the higher catalysis rate. However, it became challenging to regulate the directionality of the bot's movement with increased speed. To overcome these challenges researchers started focusing on the synthesis of magnetic micro/nanobots. Catalytic micro/nanobots developed from magnetic nanoscale materials were convenient in controlling and guiding the direction of the bot's movements. For example, magnetic microbots were designed

from ferromagnetic cobalt ferrite ( $\text{CoFe}_2\text{O}_4$ ) and doped with palladium nanoparticles to retain their catalytic properties [17]. Although the size of cobalt ferrite nanoparticles was 40 nm for better visualization of its functionality under an optical microscope it was agglomerated to 150  $\mu\text{m}$ . The decomposition of  $\text{H}_2\text{O}_2$  in an aqueous solution using this Pd-doped magnetic nanobot showed better chemical stability and higher oxidation [17].

Furthermore, bubble-propelled tubular microswimmers were developed with a catalytic inner layer. The tubular structure enabled easy collection along with enhanced bubbles emission in constricted reactors [27]. As the gas had assembled and the internal gas pressure went up, the bubble accelerated towards one end getting rejected from an opening to the external surrounding. This autonomous strategy in the form of chemotaxis for the movement of bubble-propelled tubular microswimmers involved a chemotherapeutic drug-loaded swimmer that could move along the  $\text{H}_2\text{O}_2$  concentration slope towards higher fuel concentration.  $\text{H}_2\text{O}_2$  is the main fuel type for both bubble-propelled and self-phoretic swimmers and is highly corrosive which can be a major problem while applying it in biomedical applications. So, enzymes have been considered a biocompatible alternative to inorganic catalysts [19].

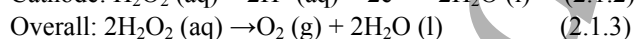
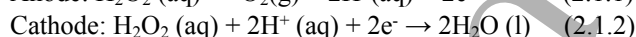
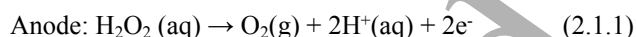
Different micro/nanobots have different shapes, their mechanism will not get affected by their shapes and sizes. An autonomously moving micro/nanobot follows bubble propulsion and propels via the bursting of oxygen bubbles in the presence of  $\text{H}_2\text{O}_2$  as fuel. Hu et al 2020 demonstrated the synthesis of a micro-vehicle made up of three metals (Au-Ni-Pt) and possessed a circular steering propulsion mechanism in the presence of  $\text{H}_2\text{O}_2$  as fuel. It was found that the decomposition of  $\text{H}_2\text{O}_2$  via a surface of the off-centred Pt nano-engine enabled the production of oxygen bubbles and the bursting of oxygen bubbles produced thrust which helped the Au-Ni-Pt micro-vehicle to move forward clock-wise or anti-clockwise direction circularly [28]. Similarly, Reddy et al in 2014 demonstrated the synthesis of Pt/Pd alloy & Au Janus micro-disk. The disk was pinned upright at the liquid/air interface and could show autonomous movement in the presence of  $\text{H}_2\text{O}_2$  fuel by decomposing it and producing  $\text{O}_2$  bubbles. These oxygen bubbles lifted the microdisk at Pd-Pt coated upper rim and aided in smooth displacement. But when the bubbles burst, the disk was propelled ballistically and proving that during bubble formation, thus propulsion was observed three times lower than the magnitude when the bubbles burst [29]. Apart from the disc-like micro/nanobots, Soler et al, mentioned the synthesis of a tubular microjet made up of photoresists on which Fe/Pt nanomembranes were deposited. Previously, the challenge was to move the microjet in the blood sample having considerably high viscosity. However, when the blood samples were diluted by ten times

and the temperature was raised from 25  $^\circ\text{C}$  to 37  $^\circ\text{C}$ , the speed of the microjet increased from 0 to 60  $\mu\text{m s}^{-1}$ . Similarly, the microjets were also tested for serum samples. At 50% concentration of serum, microjets showed a speed of 30  $\mu\text{m s}^{-1}$  when the temperature was raised from 25 to 37  $^\circ\text{C}$  proving that there is an effect of temperature and concentration of the fuel/fluid in the locomotion of microjets [30]. A Pt-coated silica microbead model was formed and tested for its velocity in the presence of hydrogen peroxide fuel. The Pt layer decomposed the  $\text{H}_2\text{O}_2$  into water and oxygen bubbles. The driving force behind the movement of the microsphere was the detachment of the oxygen bubbles upon bursting which created a thrust and helped the microsphere to move [31]. Apart from the disc-like micro/nanomotors and tubular microjets, a rod-shaped nanomotor made up of gold and platinum showed autonomous movement in the presence of 2-3% hydrogen peroxide. The nanorods moved along their axis with a speed of ten body lengths per second in the direction of platinum [32]. So, irrespective of the different shapes/sizes, a micro/nanomotor will follow the same bubble propulsion mechanism when hydrogen peroxide has used as fuel for locomotion. Furthermore, the concentration of hydrogen peroxide is directly proportional to the speed of the gold (Au)-nickel (Ni)-platinum (Pt) micro/nanobots [137-139,141]. The bursting of oxygen bubbles powers micro/nanobots that can freely be driven in a circular motion in hydrogen peroxide solution [139,141]. The presence of an external magnetic field guides the micro/nanobots in any specific directions that assist in improving the accuracy bot's functionality which can very well be reversed with the withdrawal of the external magnetic field. Along the magnetic field lines, the nanobots can move ahead [138,139]. Additionally, modulation with magnetic rotation at the tail could change the orientation of the gold (Au)-nickel (Ni)-platinum (Pt) nanorods. [142]. Moreover, the increase in temperature of hydrogen peroxide solution is directly proportional to the speed of the gold (Au)-nickel (Ni)-platinum (Pt) micro/nanobots [138,140]. Therefore, when the temperature was below 27  $^\circ\text{C}$ , larger bots move rapidly in contrast to the smaller ones that propel quickly at above 27  $^\circ\text{C}$  [140]. Chemical reaction, magnetic field-dependent and temperature – all these can be employed to have accurate control over the direction and speed of micro/nanobots [139].

*Diffusiophoresis propulsion:* Diffusiophoresis is one of the common mechanisms of gradient-based propulsion of micro/nanobots. It is most widely associated with developing micro/nanobots with asymmetric distribution of catalase/dopant. In this integrated system, the dopant is carefully positioned preferentially on one side of the nanobots for the accumulation of  $\text{H}_2\text{O}_2$  oxidation products or bubble formation. This creates a concentration gradient of oxidation products on the surface of micro/nanobots. Eventually, a critical point comes when there is an accumulation of oxidation product on the side of the bot causing a locally

higher concentration and the product starts to diffuse away. This diffusion leads to the propulsion of the micro/nanobots.

**Self-electrophoresis propulsion:** Self-electrophoresis propulsion in micro/nanobots requires a specific synthesis method of membrane template-assisted electrodeposition with sequential deposition of varied metals that could chemically generate an electrical gradient in the presence of fuel. It works on the principle of electrochemical cells where one end/side of bots acts as a cathode while the other as an anode. A proton gradient is developed along the axis of the micro/nanobots [33]. This assists in the movement of these bots with the gradient. For example, the synthesis of Janus micro/nanobots using a fabrication technique based on asymmetric bipolar electrodeposition of metallic thin films on microbeads is suitable for propulsion using this mechanism [34,35]. Similarly, this propulsion mechanism is applied to bimetallic nanorods that could generate electric gradients within bimetallic nanorods. Most widely coupled Au/Pt nanorods-based bots were developed that could in the presence of  $\text{H}_2\text{O}_2$  in the aqueous solution. During Au/Pt coupling, negatively charged electrons were generated at Pt (anode) and positively charged protons at Au (cathode). Both were utilized later at the cathode for the reduction of  $\text{H}_2\text{O}_2$  to  $\text{H}_2\text{O}$  and  $\text{O}_2$ . Oxygen produced was further reduced into the water by a four-electron reduction mechanism, making it a bubble-free mechanism of propulsion [24]. This mechanism of catalytic reduction was due to the net migration of electrons from an anode (Pt) to the cathode (Au), generating a proton gradient along the axis of the nanorod. These bimetallic rods functioned as tubular nanobots found to be effective for various applications. The electrokinetic disproportion of hydrogen peroxide asymmetrically at the two metal surfaces was shown as [24]:



**pH-induced movement:** Dey et al. demonstrated the synthesis of an autonomously moving microsphere of palladium (Pd) nanoparticles coated over the surface of a copolymer made up of polystyrene divinyl benzene. The polymer microsphere was of 2  $\mu\text{m}$  diameter, while Pd nanoparticles were of size 70 nm [36]. In the presence of 5%  $\text{H}_2\text{O}_2$ , the microsphere was observed to have an autonomous movement. Chemical and pH gradients were introduced in the aqueous solution to control the directionality of the propulsion mechanism of the microsphere. For this purpose, 0.3N sodium hydroxide (NaOH) was added dropwise to the medium continuously, to obtain good propulsion in the places where the concentration of NaOH was high. The microsphere was moving from a low to a high NaOH concentration, the movement was increased due to the increase in the rate of

decomposition of  $\text{H}_2\text{O}_2$ . Moreover, the average speed of these microspheres was observed as  $1.5 \times 10^{-3}$  m/s i.e., around 15 body lengths/second in the pH varying between 3.1 to 7.1. At a pH range between 7.1 to 10.1, the average speed was  $2 \times 10^{-3}$  m/s which was around 20 body lengths/second.

The Marangoni effect is responsible for the movement of nano/micro-scale liquid droplets. The change in the interfacial tension due to the absorption and desorption of solutes over the liquid surface was due to the mass transfer on the surface of the liquid. An oil droplet of size 500  $\mu\text{m}$  was developed on the surface of di(2-ethyl hexyl) phosphoric acid (DEHPA) that possess a pH-induced movement inside a liquid solution [37]. On increasing the pH, deprotonation of DEHPA takes place leading to an increase in the mobility of oil droplets. As the pH was increased up to 8-13.5 using NaOH, more deprotonation was observed due to a decrease in their interfacial tension. The speed of the oil droplet was 6 mm/s at a pH of 11.2 but did not show any motion beyond 13.3 pH.

## 2.2. Fuel free

**Electric field driven:** The idea of using an alternating electric field for creating autonomous movement in semiconductor diodes of millimetre range was further explored to produce nanobots by miniaturizing semiconductor diodes. The directional movement of these diodes was due to the generation of electro-osmotic flow around them [38]. The principle was investigated and applied on developed poly(pyrrole)-cadmium (PPy-Cd) and CdSe-Au-CdSe nanowires. The spatially uniform alternating electric field was applied to induce electro-osmotic flow around nanowires. This induction produced a uniform directional motion having a speed of 17 mm/s.

**Magnetic field driven:** Magnetically propelling micro/nanobots has several advantages over other mechanism. The major advantage is it can be used non-invasively with easy control of motion direction for targeted applications. Therefore, they are free from the use of excess toxic chemical-stimulated propulsion. To stimulate different types of motion in the bots, different kinds of magnetic fields such as homogenous, rotation, oscillating, or inhomogeneous were applied. Nanoscale materials used for designing bots could be ferromagnetic, diamagnetic, or paramagnetic and can be moulded into several shapes such as helical, nanowire, nanoscrews, tubular, or polymeric nanobeads [39,40].

For example, magnetically propelled nanoscrews were developed from silica helices using a shadow growth technique on a nanolithographically patterned substrate (Fig. 2b) [18,41]. As nickel possesses some magnetic properties, nickel segments of thickness 40 nm were incorporated over the silica helices of diameter 70 nm, leading to a total approximate width of 120 nm. The spacing, length, and pitch of the nanoscrews were 130 nm, 400 nm, and 100 nm

respectively. The magnetization inside and outside of the plane was  $1.13 \times 10^{-6}$  emu/mm<sup>2</sup> and  $0.13 \times 10^{-6}$  emu/mm<sup>2</sup> respectively. The nano-screws were magnetized diametrically because of the lower magnetization outside the plane resulting in higher permeability over a wide range of magnetic field strength [18]. In another example, CoPt alloy nanowires-based surface walker was prepared using template-assisted galvanostatic electrodeposition and had semi-hard magnetic properties [42]. This property of nanowires allows easy modulation with the applied/external magnetic field and retains their magnetization direction thereafter. The size and morphology of the nanowires were obtained using scanning electron microscopy and energy-dispersive X-ray spectroscopy. Nanowires of diameter 200 nm and length 5  $\mu$ m were pre-magnetized under an applied magnetic field. The propulsion is seen to be rotating around the x-axis, the y-axis experiences drag force and the z-axis experiences drag force away from the wall. As the magnetic field strength gradually moved up to 15 mT the frequency of the rotation motion increased from 1 Hz to 5 Hz thus, exhibiting a tumbling motion [42]. The magnetic actuation method of mobile nanobots involved time-varying rotating magnetic fields and gradient pulling. Cells and tissues can be infiltrated by magnetic fields in quite a safe manner which made them a highly promising method for biomedical applications [43]. There are several advantages of using a magnetic actuation-based method that involves wireless powering, biocompatible energy sources, and guided controlled motion; hence it is suitable for both in-vitro and in-vivo applications as well as the development of lab-on-a-chip based sensors. Major limitations of using magnetic field as an operating method for nanobots involved trouble in selective agent addressability and greater cost demand for medical applications [44,45].

**Acoustic propulsion:** Ultrasound has a frequency value above the human hearing threshold range. It can be used as an external stimulus to guide and generate nanobots. An experiment was conducted on an electrodeposited metallic rod, using continuous or pulsed ultrasonic waves for its propulsion. The metallic rod prepared by template electrodeposition leads to the formation of convex or concave ends, causing asymmetry. The asymmetry within the rods plays an important role by creating a non-uniform distribution of ultrasound pressure. Thus, generating an ultrasonic gradient which becomes the driving force for the metallic rods and propels it in a specified direction. This driving mechanism for directional motion using acoustic waves is proposed as self-acoustophoresis.

The actuation method by the auditory field involved auditory transmission energy and audible rolling. These are the promising origin for the micro/nanobot motility in constrained pathways. The force of acoustic radiation caused due to the stagnant wave was generated while a back-and-forth

reflection of the sound wave via resonator was done in a biologically safe manner. In the acoustic operation method, the pathway of propulsion can be known beforehand and effectively changed by wave functions, which made this technique useful in handling the motility of numerous microbots thus accumulating these in the required specific locations. In acoustic actuation schemes, oscillating bubbles trapped within micro-swimmer bodies generated sufficient thrust. In the case of in vivo systematic motion, an acoustically activated flagellum was used to propel artificial micro-swimmers via a fluid by a low amplitude oscillation of the tail in the existence of moving auditory waves [46]. Acoustic radiation force can be created by focusing the ultrasonic beam to confine an individual or a collection of micro-objects by physical means. This method has the potential to be applied for drug delivery and detoxification (Fig. 2c) [18]. The acoustic actuation method involved biocompatible energy sources and reliability for lab-on-a-chip and in vitro uses. The major limitation was in-vivo usage which needed the development of proper instrumentation [45].

**Light-driven propulsion:** Several compounds such as spirogyra, titanium dioxide (TiO<sub>2</sub>), silver monochloride (AgCl), etc possess photo-catalytic and optical properties. Therefore, it can be easily sensitized by the light of some specific wavelength. These properties were harnessed to develop micro/nanobots driven by light stimulation. AgCl when sensitized under UV, it dissociated to produce protons and chloride ions. An electrolyte gradient was formed when protons started to diffuse away from AgCl which is opposite from the chloride ion. This resulted in an autonomous motion, which was as high as 10 cm s<sup>-1</sup>. On the other hand, TiO<sub>2</sub> possessed high photocatalytic properties which were exploited to develop light-stimulated micro/nanobots and devices (Fig. 2d) [47,48]. Self-diffusiophoresis of TiO<sub>2</sub> occurred when induced under UV light, causing the motion of bots having a speed range from 103 mms<sup>-1</sup>. When TiO<sub>2</sub>-based micro/nanobots were propelled under UV illumination, a sufficient amount of hydrogen bubbles was produced for the micro/nanobots propulsion as mentioned in the reaction given below [4]:

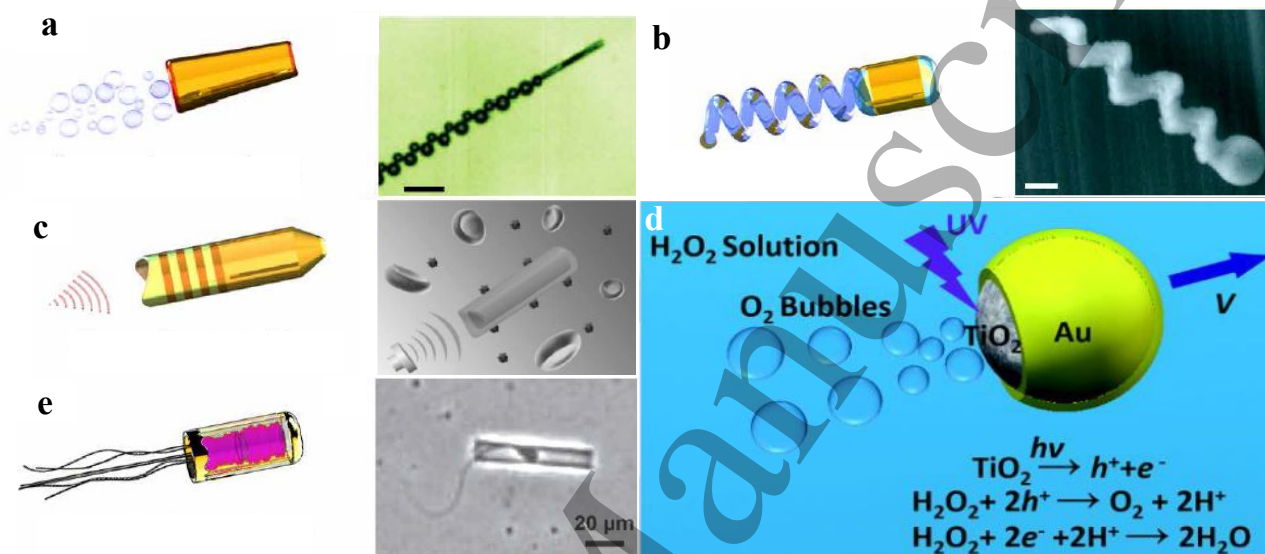


Spiropyran is an organic compound having tremendous photochromic properties. Micro/nanobots based on spiropyrans were not exactly driven by a light source but they acted as a molecular switch for controlling their motion and speed. Spiropyrans-based bots in an aqueous medium consist of different carbon chains length. These molecules were labelled as spiropyrans SP-C1, SP-C8, and SP-C18 having



methyl, octyl, and octadecyl side chains, respectively [49]. Fundamentally, the light propulsion mechanism involves the light-induced formation of thermal gradients surrounding the micro/nanobots that ultimately helps in bots' movement. A thermal slope over an air-liquid interface led to the fluid flow in the cooler site pathway due to the temperature dependence nature of the surface towards their application. A huge number of bubbles were created and controlled with the use of a liquid crystal device that can alter a laser wavefront into numerous outputs. Light can also move directly via untethered microdevices. Any laser source with optical trapping can be

used to hold the position of micro/nanobots in a three-dimensional pattern. The sub-micron resolution and power transmission in numerous directions are the major advantages of using the light-induced propulsion method for micro/nanobots. Other advantages of this mechanism involve producing travelling waves, simple selective agent addressability, and reliability for lab-on-a-chip and in vitro applications. However, this operating method is restricted to two-dimensional patterns and is unsuitable for in vivo conditions [49].



**Fig. 2:** Representation of propulsion mechanism of various types of micro/nanobots. a) Chemically powered bubble propulsion mechanism in micro-rocket. (Adapted with permission from reference. [18], Copyright 2017 AAAS); b) Magnetically propelled helical nano-swimmers. (Adapted with permission from reference [18], Copyright 2017 AAAS); c) Ultrasound-propelled micro/nanobots used for biodegradation. (Adapted with permission from reference [18] Copyright 2017 AAAS); d) Light-driven Au-TiO<sub>2</sub> microbots in the presence of hydrogen peroxide fuel. (Adapted with permission from reference [48], Copyright 2019 Wiley); e) Biologically actuated sperm-based hybrid microbots. (Adapted with permission from reference [18] Copyright 2017 AAAS).

### 2.3. Biohybrid micro/nanobots

Mostly, micro/nanobots were developed with the potential of being driven chemically. However, it cannot be suitable for most biological applications due to the toxicity of fuels such as hydrogen peroxide or hydrazine. Therefore, these are not suitable for most in-vivo biomedical applications. Biohybrid micro/nanobots are being explored to overcome such toxicity issues [50]. Biohybrid bots are the integration of functionality, efficacy, and merits of biological systems into artificial micro/nanobots. Precise control and biocompatibility of biohybrid micro/nanobots could provide new scope for the treatment and diagnosis of various chronic and incurable diseases. The current micro/nanobot's development mechanism is focusing on incorporating biological fluids, enzymes, biomolecules or biological cells. For example, the use of catalase bio-enzyme in nanobots for the efficient

decomposition of hydrogen peroxide to convert chemical energy to mechanical energy. Catalytic carbon microfibres driven by glucose and oxygen are another example of a biohybrid bot.

Biological cells based biohybrid microbots such as bacteria, spermatozoon, and muscle cells access the chemical energy from the environment into spontaneous work, modulating their energy by reacting with forces, mechanical strain, and chemicals in their surroundings. In this approach, single cells or tissue were physically integrated with synthetic components to harness the propulsion and sensing capabilities of operational microbots. The biohybrid micro/nanobots are mainly suitable to operate only in biological media. The main design strategy for the effective powering of these bots involves the physical attachment of blood cells, spermatozoon, muscle cells or any biological cells with synthetically engineered bodies [19,51,52,73]. Integrated



sensing and mobility, inherent compatibility with biological media, and high efficiency in energy output are some of the major advantages of using biohybrid micro/nanobots. Limitations of these micro/nanobots include live cells that can function only in delicate conditions (such as 37°C, 5% carbon dioxide, and nutrients) to survive. Similarly, biohybrids were also designed by tagging micro/nanobots with live organisms to assist in autonomous movement. The idea behind such micro/nanobots is harnessing the ability of natural swimmers such as bacteria, cells, or spermatozoon. These cells are motile within the biological system thus integrating with the micro/nanobots provides autonomous movement.

Biological systems are complex structures that function with utmost efficiency and sustainability. A mimicking biological system with artificial structures is extremely hard to achieve. Therefore, the integration of live cells and organisms with micro/nanobots can help improve the performance of the bot in biological systems. For example, spermatozoon was used to design bots that could move through the viscous media of biological systems (Fig. 2e) [18]. Thus, it is to be deemed that spermatozoa have huge potential to serve as a carrier within the biological system. This concept is made live when spermatozoon drag attached micro/nanotubes to perform biological functions (Fig. 3A) [33,34,43,]. Another example is using magnetotactic bacteria

possessing magnetic crystals within them to move along with the geometric fields. These bacteria are integrated with nanoscale materials and biological systems to perform nano-surgery in vascular systems (Fig. 3B) [53-57]. Wu et al in 2014 demonstrated the synthesis of RBC-based motors where iron oxide nanoparticles were incorporated in the RBC cells (Fig. 3C) [50]. The RBC-based motors were propelled under the influence of both acoustic and magnetic fields. When this motor was propelled in the bloodstream, it showed no biofouling effect as well. The RBC motor was not only biocompatible but also remained unaffected by the macrophages present in the blood. This property of the RBC motor can be used wisely in biomedical applications [58]. In another approach, a polymer-based nanorocket was designed using polysaccharides of opposite charges. Chitosan (CHI) and alginate (ALG) are positively and negatively charged polysaccharides respectively. Both were assembled layer-by-layer over a porous polycarbonate membrane having a thickness of 10 µm and the diameter of the pore was around 600nm. After 18 layer-by-layer depositions, ALG was found to be present at the inner layer. Then, deposition of poly (diallyldimethylammonium chloride) (PDADMAC)-stabilized Pt NPs was done inside the template pores. Later, the polycarbonate membrane was dissolved in CH<sub>2</sub>Cl<sub>2</sub> to obtain the Pt NPs modified (CHI/ALG) nanotubes [58].

**Table 1:** Design strategies and their respective advantages of various types of micro/nanobots.

S.No.	Micro/Nanobots	Material	Fabrication method	Propulsion mechanism	Advantages	Ref.
1.	Au/PEDOT/Pt micromotor	Au/PEDOT/Pt	Template electrodeposition	H <sub>2</sub> O <sub>2</sub> Driven	Cancer biomarker miRNA-21 detection;  Breast cancer treatment	[59]
2.	Hybrid stomatocyte nanomotors	Pt NPs/ PEG-b-PCL/ PEG-b-PS	Self-assembly	H <sub>2</sub> O <sub>2</sub> Driven	Anti-cancer drug delivery;  Controlled drug release	[60]
3.	Needle-type microrobot (MR)	Ni/TiO <sub>2</sub>	3D Laser Lithography;  Physical Vapor Deposition (PVD)	Magnetically Driven	Stable drug delivery	[61]
4.	Biomimetic micropropellers	SiO <sub>2</sub> beads/ Al <sub>2</sub> O <sub>3</sub>	Physical Vapor Deposition (PVD); Glancing Angle Deposition (GLAD)	Enzyme (Urease) Driven	Penetrates in biological gels	[62]

5.	Biohybrid magnetite microrobot	<i>S. platensis</i> / Fe <sub>3</sub> O <sub>4</sub> NPs	Dip-coating process	Magnetically Driven	In vivo fluorescence imaging; Remote diagnostic sensing	[63]
6.	Magnetic microhelices	Ni/Ti Bilayer	Direct Layer Writing	Magnetically Driven	Spermatozoon transportation; Cellular cargo delivery	[64]
7.	Nanoswimmer	Polypyrrole (PPy)/ Ni/Au	Multistep electrodeposition	Acoustically Driven	Linear propulsion; Biosensing; Bioimaging	[65]
8.	Supramolecular nanotrain	CNT/ Liposome	Soft Lithography and Photo Lithography	Light Driven/ Electrically Driven	lab-on-a-chip applications; medical diagnosis; biosensing	[66]

#### 2.4. Intelligent micro/nanobots

Micro/nanobots must self-adaptively go to the target places and complete the task in uncertain or constantly changing physical, chemical or physiological parameters. To avoid being controlled by fluctuating external signals, micro/nanobots must be intelligent or smart so that they could modulate themselves with different motion behaviours and functions.

*Intelligent taxis mechanism:* Numerous intelligent micro/nanobots with self-navigation/self-targeting have been elucidated based on their strategic movements toward or away from stimulus sources. Therefore, received a lot of attention over the past few years. In this regard, the mechanism of the intelligent response of such micro/nanobots was demonstrated using two models. First, for the asymmetric micro/nanobots the local vector fields, such as those created by gravity, flows, magnetic fields and concentration gradients cause overall alignment that further helps these bots to swim toward or away from the signal sources. Second, isotropic micro/nanobots generally create propulsion forces directed towards the local vector field, regardless of their Brownian motions [67].

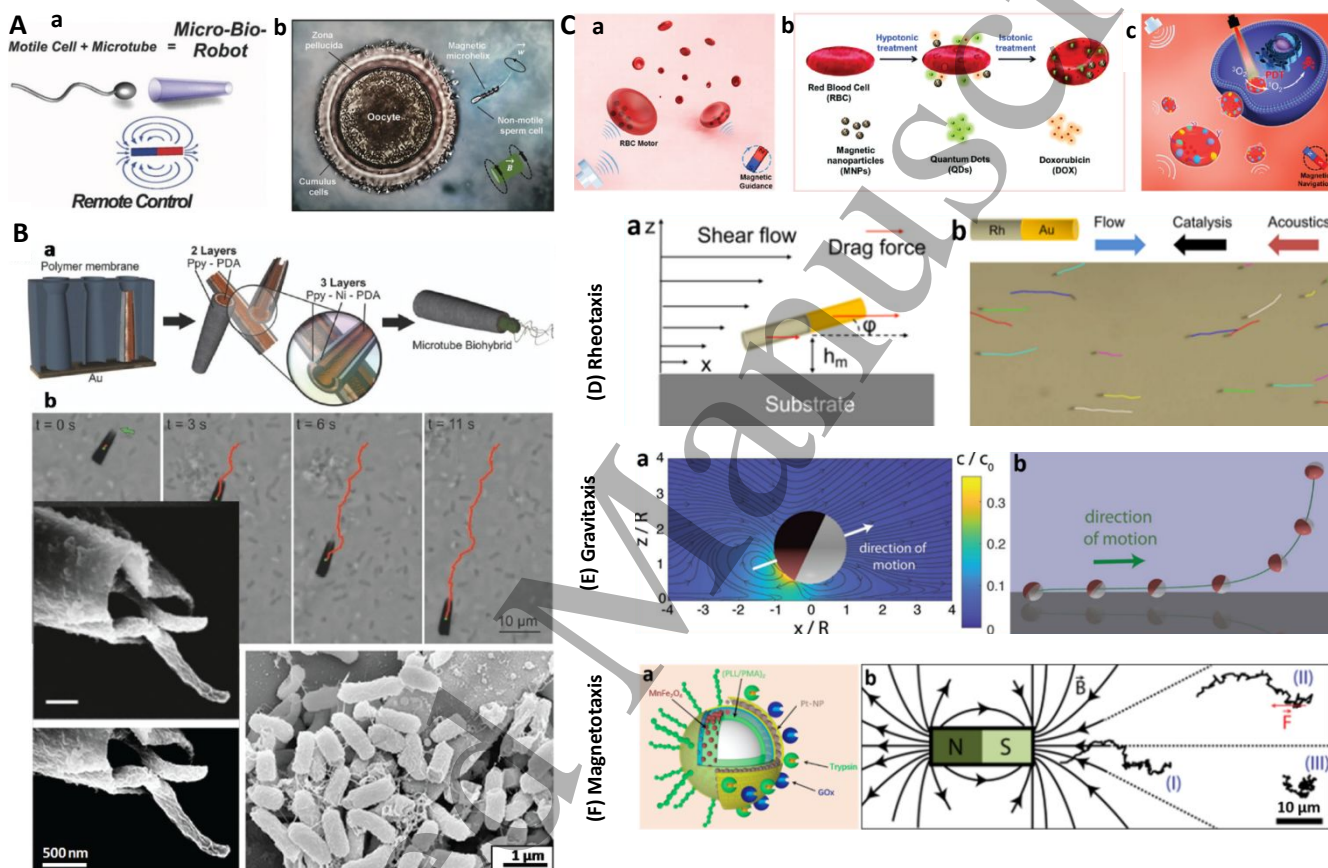
Asymmetric micro/nanobots with tailored responses were subjected to a force  $F$  when exposed to a vector field produced from a flow, gravity, or magnetic field. These bots can self-reorient along the local vector field if their symmetry axes were out of alignment due to an aligning torque ( $M$ ) produced by force  $F$ . Rheotaxis, chemotaxis, gravitaxis, and magnetotaxis in micro/nanobots was thus accomplished. In asymmetric micro/nanobots, overall tactic motions were primarily due to a combination of random motions brought on by Brownian rotation and directional motions along the vector field. Since aligning torque  $M$  was powerful enough to restrict their rotating Brownian diffusion, these bots could move

roughly in a straight path. If not, they will steer off in any direction from the signal source [67]. For example, chemotactic micro/nanobots always travel in the opposite direction of the chemical gradient and frequently engage in intelligent "deviating-rectifying" behaviour. In addition to an asymmetric distribution of mass, asymmetric physical or chemical fields that surrounds micro/nanobots along the direction of gravity can also cause them to reorient, creating negative gravitaxis in bots [67]. For instance, Singh and coworkers synthesized TiO<sub>2</sub>/SiO<sub>2</sub> Janus micro/nanobots that were light-induced negative-gravitactic swimmers under illumination from the bottom of the substrate (Fig.3E). In their demonstration, the bots were moved and reoriented to a cap-down configuration, where it exhibited negative gravitaxis, due to the photocatalytic breakdown of H<sub>2</sub>O<sub>2</sub> by TiO<sub>2</sub> that causes a local fluid flow in its proximity. These smart taxis were used to design "micro/nanoelevators" and several possible applications, including vertical freight transport and dynamic separation of active and passive molecules, based on the development of negative gravitational movements [68]. Similarly, rheotaxis (positive/negative) can be achieved when aligning torque  $M$  of the micro/nanobots overpowers the Brownian motion that keeps their orientation unchanged. Based on this concept, Rh/Au bimetallic microbots were developed in which both positive and negative rheotaxis can be achieved when a combination of acoustic field and chemical fuel was provided (Fig.3D) [69]. Furthermore, micro/nanobots with anisotropically incorporated magnetic components can produce intelligent responses, termed magnetotaxis. Microbots developed by incorporating Ni-Pt into the stomatocyte polymer matrix can potentially produce a response corresponding to the direction of the magnetic field. In another example, Janus micro/nanobots developed from Mn-Fe<sub>2</sub>O<sub>4</sub> nanoparticles can propel in the presence of fuel

with their orientation aligned by the presence of a sufficient magnetic field. In this magnetotactic movement, curved trajectories can also be observed as soon as the intensity of the magnetic field is reduced during the motion (Fig.3F) [70,71].

In another example, electrically powered micro/nanobots were used for intelligent cargo delivery. Herein, a 3D orthogonal microelectrode system was developed having platinum-gold (Pt-Au) nanobots that could move independently between a planar four-level microelectrode. When an alternating electric field was applied to these planar quadrupole microelectrodes, the catalytic nanobots could

smartly align with the electric field direction. These bots then moved autonomously along the direction of the alternating electric field with the Pt segment as the leading edge. As the bots approached the cargo (gold nanorods), they were put together by the interaction of electric dipoles. Thus, strongly anchoring together at the edge of the metal microdock in the presence of an alternating electric field produced due to the induction of a nearby electric field. As the alternating electric field is switched off, the nanobots could instantly release cargo and be ready for the next task [72].



**Fig. 3:** **A)** Spermatozoa hybrid micro/nanobots: (a) Structural schematic of sperm cells coupled with magnetic microtubes to remotely guide the bots (Adapted with permission from reference [73], Copyright 2013 Wiley); (b) Microscopic representation of transportation of immotile sperm to cumulus cells using microbots (Adapted with permission from reference [64], Copyright 2016 Americal Chemical Society); **B)** Bacteria hybrid micro/nanobots: (a) Schematic of *E. coli*-propelled tubular microbots (Adapted with permission from reference [56], Copyright 2017 Wiley), (b) Time-lapse images of the movement of *E. coli*-propelled tubular microbots. Copyright 2017, Small. (inset images) SEM of an MSR-1 *E. coli*-propelled tubular microbots. Copyright 2017, ACS Nano (Adapted with permission from reference [57], Copyright 2017 American Chemical Society); **(C)** Red Blood Cells hybrid micro/nanobots: (a) Schematic of acoustically propelled and magnetically guided motion of RBC microbots (Adapted with permission from reference [50], Copyright 2014 American Chemical Society), (b) Illustration of RBC hybrid microbots carrying multiple cargos (Adapted with permission from reference [51], Copyright 2015 Royal Society of Chemistry), (c) Illustration of photodynamic therapy using RBC hybrid microbots (Adapted with permission from reference [52], Copyright 2013 Wiley); **D)** (a) orientation mechanism of Rh/Au bimetallic micro/nanobots, (b) rheotactic movement of Rh/Au bimetallic bots in presence of fuel  $H_2O_2$  as well as acoustic field (Adapted with permission from reference [69], Copyright 2017 American Chemical Society); **E)** (a) Concentration field and

fluid flow of  $\text{TiO}_2/\text{SiO}_2$  Janus micro/nanobots, (b) its lift-off trajectory (Adapted with permission from reference [68], Copyright 2018 Wiley); **F)** Magnetotactic micro/nanobots: Magnetotactic trajectories of the double-fueled magnetic microbots (Adapted with permission from reference [70], Copyright 2017 American Chemical Society).

*Nature-inspired intelligent motion and function:* In the natural world, bacteria can use their flagella to swim in low Reynolds number liquid conditions. Nelson's group made the first demonstration of synthetic bacterial flagella employing helical propulsion, drawing inspiration from the motion characteristic of bacteria [74]. The helical micro/nanobots have soft-magnetic metal square "head" and a helical tail with a shape and size similar to a natural flagellum developed using a self-scrolling approach. A weak, rotating magnetic field that resembles bacterial flagella could be used to propel and guide the helical motor through the water Fig 4A. Zhang et al. built synthetic bacterial flagella-based bots with functional liposomes for in-vitro transport of calcein to mouse cells by modulating an external magnetic field based on the bacterial flagella investigations mentioned above (Fig 4B) [75]. Likewise, an artificial multilinked two-arm microbot had been shown to mimic natural swimming movements. The two-arm microbots exhibit an effective "freestyle" swimming style comparable to that of humans in a planar oscillating magnetic field. It was made up of gold, while two arms made of nickel, and two porous silver hinges connected the arms and body. The two-arm microbot's maximum speed was observed as  $59.6 \mu\text{m/s}$ , while both speed and direction can be altered as required (Fig 4C) [76].

Generally, fish movement is dependent on body swing and induced surrounding fluid flow, unlike the freestyle stroke of humans. In another example, inspired by the swimming behaviour that animals Li et al. developed magnetically powered artificial microbots that resemble fish. The deformable fish-shaped microbots body comprises one gold head, one gold caudal fin, two nickel bodies, and silver hinges connecting each part. This design was imitated by the nanofish microbots, which were created using a template electrodeposition and chemical etching approach. The nanofish microbots may produce periodic mechanical deformations and display travelling-wave motions in a planar oscillating magnetic field (Fig 4D) [77].

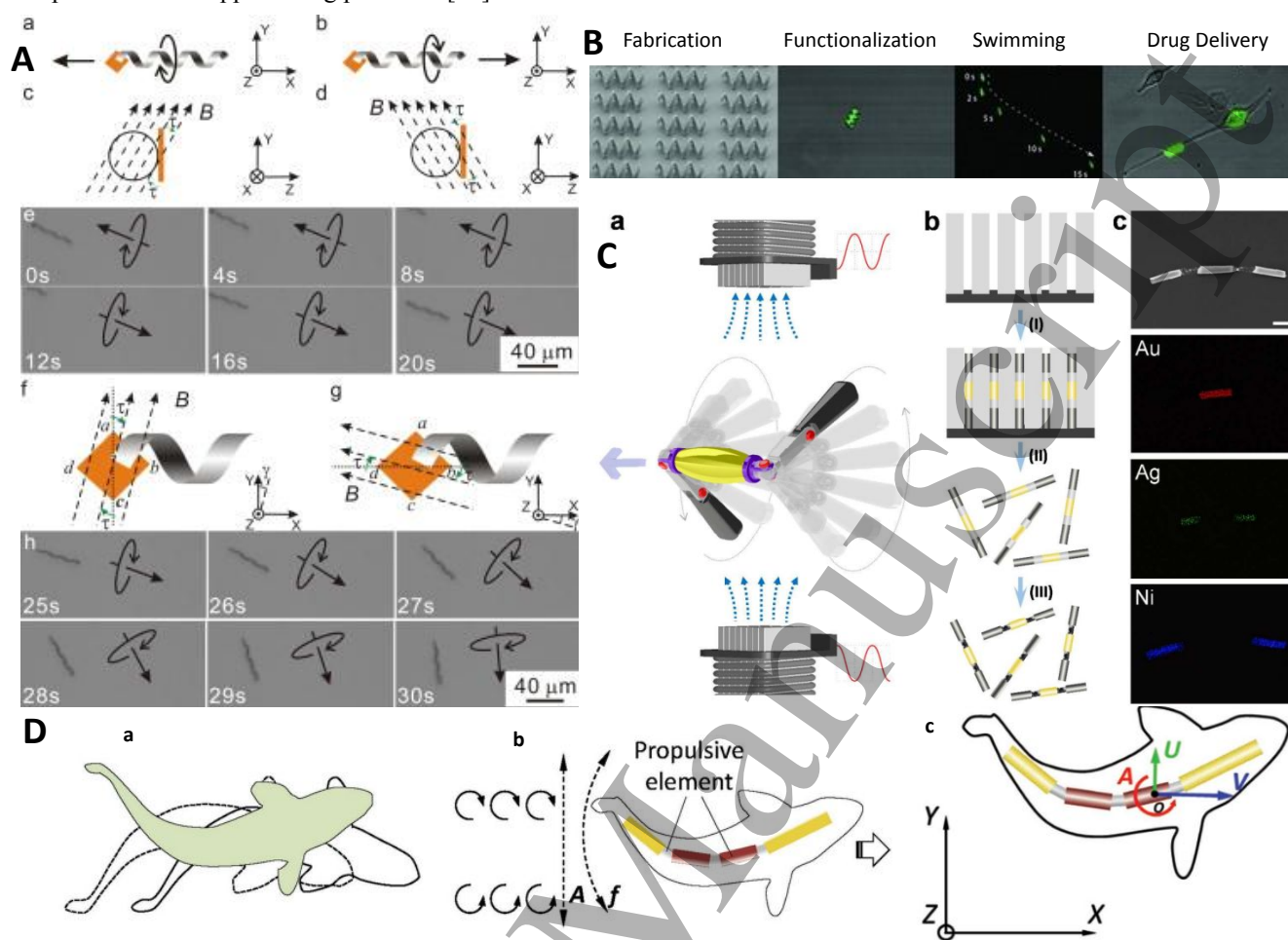
In another approach, Walker et al. demonstrated the development of a micro-propeller moving under a magnetic field. In this approach, a micro-propeller was developed which could mimic the functions of *H. pylori* and can penetrate through the gastric mucus. For the interaction with the mucus surface, modulation was done by bile salts, polyethylene glycol (PEG), some polymers, ligands and many more. But when the modulation was done by immobilizing urease on the surface of micro-propellers, two advantages were observed: (a) mobility of the micro-propellers increased, and (b) urease was capable of degrading the mucus lining. This approach was done to prove that the biomimetic micro-propellers not only

showed higher mobility on modulation via enzymes but were also capable of penetrating the mucus layers [62].

Medina-Sanchez et al. presented microbots made up of micro helices capable of transporting spermatozoon. This approach was made because of the problems associated with the movement of the spermatozoon. Thus, there was a requirement to develop a system where spermatozoon could be delivered so that they can perform their natural functions. The ferromagnetic microbots consisted of polymeric microhelix with a coating of NiTi soft-magnetic bilayer. When the spermatozoon was loaded on the microhelix, the Ni layer of the microhelix followed magnetic propulsion and attained a velocity of  $17.6 \pm 3.53 \mu\text{m/s}$ . When the spermatozoon was delivered to the oocyte, their fusion made this approach successful. Under the influence of the magnetic field, the transportation and fusion of the spermatozoon with the oocyte made this approach applicable in future as well [78].

Swarming is another example of a common phenomenon in the natural world. Natural swarm behaviours, as compared to individual activities, show exceptional advantages in protecting against confounding predators, foraging for prey, and structural flexibility as per the environment. Artificial micro/nanobot swarms have been constructed to achieve collective behaviours and functionalities, drawing inspiration from the swarm behaviour in nature. Functionalized  $\text{Fe}_3\text{O}_4$  nanoparticles were used to create a reconfigurable microswarm that resembled an ant bridge in an oscillating magnetic field [79]. By adjusting the magnetic field, one may control the structure and behaviours of microswarms. Similarly, drawing inspiration from the swarm behaviour and function of the ant bridge stretchable ribbon-like  $\text{Fe}_3\text{O}_4$  nanobots were created to provide a conductive channel for electrons between two isolated electrodes. As a result, the reconfigurable microswarm demonstrated an exceptional capacity to intelligently restore damaged microcircuits in biomedical devices [79]. Moreover, a tornado-like 3D microswarm was developed by fusing the aforementioned 2D swarm behaviour under a magnetic field and light-induced vertical hovering. This behaviour was motivated by the antigravity migration of zooplankton toward the light. The four steps of the swarm's reconfigurable collective behaviour from 2D to 3D are rising, hovering, oscillation, and landing. In the wild, prey animals frequently congregate into bigger groups to confuse or elude predators. Based on this behaviour, an intriguing predator-prey interaction system was developed in which diffusiophoretic repulsive and appealing microparticles acted as predators and prey, respectively. Before being combined with predator microbots, the  $\text{TiO}_2$  microbots formed a swarm of prey particles. As a result of the interactions between repulsive and attracting diffusiophore,

the prey swarms began to alter their group shape in response to the position of the approaching predators [80].



**Fig 4:** **A)** (a-f) Bacterial flagella inspired movement mechanism of helical microbots (Adapted with permission from reference [74], Copyright 2009 Applied Physics Letters); **B)** SEM images of development, functionalization, movement and drug delivery potential of the microbots having corkscrew propulsion like bacterial flagella (Adapted with permission from reference [75], Copyright 2014 Wiley); **C)** Two-arm nanobots inspired by freestyle swimming: (a) schematics, (b) development of multilinked two-arm magnetic nanobots, (c) SEM image and EDX mapping of two-arm nanobots (Adapted with permission from reference [76], Copyright 2017 American Chemical Society); **D)** Fish inspired nanobots: (a) Illustration of various movement forms of natural fish, (b) design and actuation of nanofish. (c) Schematic of the defined coordinate system for the simulation of the undulatory motion of nanofish (Adapted with permission from reference [77], Copyright 2013 Wiley).

### 3. Applications of micro/nanobots in biomedical sciences

Micro/nanobots are highly miniaturized devices and offer the scope of functionalization with various small and biomolecules suitable for 'programmed' applications. The micro/nanobots thus may prevail over low Reynolds number viscous drag and show Brownian motion by transforming applied or local fuel into mechanical energy-producing desired movement has become an important tool for biomedical applications. The power control, viability, and

usability of profuse nanoscale robotic prototypes have been enhanced by immense efforts from the micro/nanobots community. The growing interest towards micro/nanobots extends true prospects for various biomedical applications, to carry out localized diagnosis, remove biopsy samples, and take images, as these bots could easily transverse via multiplex biological media or thin capillaries. These micro/nanobots can independently liberate their payloads at predetermined destinations [81]. Biocompatible micro/nanobots can degenerate, dissipate and vanish after their mission is completed. Amplified tissue invasions and payload retention capacity have been revealed by in-vitro micro/nanobots



performance. Unbound micro/nanobots constitute a prospective substitute for invasive medical robots and passive drug carriers [19]. Some of the major and potential biomedical applications of micro/nanobots include targeted drug delivery, precision surgery, sensing of biological targets, and detoxification (Table 2) [18].

### 3.1. Integrated and targeted drug delivery

Most of the micro/nanobots reported are designed to act as transport cargoes. This includes the applications of targeted drug/gene delivery. In 2009 Martel et al explored polar magnetotactic bacteria to develop micro/ nanobots for drug cargo loading [82]. Later researchers demonstrated that drug-loaded liposomes, nanospheres, and/or free drugs may deliver by the micro/nanobots. An interesting example of this is the use of catalytic nanobots of Ni/Pt alloy for the transport of doxorubicin-loaded liposomes. The precise delivery of the drug is possible due to the interaction of ferromagnetic nickel of catalytic nanobots with iron oxide captured inside liposomes particles or nanospheres. Similarly, Au/Pt microbots are interacted electrostatically to associate with positively charged polystyrene microspheres while streptavidin functionalized polystyrene microspheres are linked through biotin-avidin interactions for their efficient delivery.

Micro/nanobots can potentially transfer as well as deliver therapeutic payloads directly to the target areas, thus boosting the therapeutic effects and lowering systemic side effects of highly toxic drugs. These drug-delivery micro/nanobots count on controlled motion and have a dearth of the force and navigation needed for targeted drug delivery and tissue invasion. Drug delivery bots should have properties such as propelling or hydrodynamic drag force, systematic navigation, cargo-towing/release, and tissue invasion to attain accurate delivery of therapeutic loads to targeted regions [33,83]. A multi-fold tubiform polymeric microbot, enfolding the anticancer drug doxorubicin through an absorbent membrane template-assisted layer-by-layer assembly was successful in delivering the drug load to the cancer cells [84]. Chemically driven Janus microbots were synthesized to develop an active cargo delivery system with a potential for full diffusion and amplification. In an approach, biocompatible drug-loaded Mg-based Janus microbots showed systemic self-propulsion in simulated body fluid or blood plasma. These microbots had embedded payload along with pH-sensitive polymer for gastric acid neutralization (Fig. 5a).[18] The magnetic microbot's machinery did direct and targeted drug delivery by carrying drug-loaded magnetic materials. Similarly, the calcium carbonate ( $\text{CaCO}_3$ ) microbot autonomously moves under the impact of  $\text{CO}_2$  bubbles generated due to the chemical reaction of tranexamic acid in an aqueous solution for delivering thrombin, a haemostatic drug for wound healing (Fig. 5c) [85]. In another example, ultrasound-driven nanobots

were developed which executed fast drug delivery towards cancer cells while light-driven micro/nanobots performed cargo delivery and release of payload at the targeted site under stimuli-responsive disruption (Fig. 5b) [86,87]. In intracellular delivery, nanobots can infiltrate via cellular membranes and directly deliver numerous therapeutic components into cells like quick internalization. The motion of ultrasound-powered gold nanowire-based bots inside living cells was utilized for expedited intracellular siRNA delivery [88]. For delivery of pDNA to human embryonic kidney cells, magnetic micro-swimmers of helical shape were used [89]. Synthetic nanobots that can be fuelled by physiological media like gastric acid and water were used in in-vitro applications. These bots can bear an enormous load of various cargos, release payloads in a responsive autonomous way and ultimately degrade themselves to non-toxic by-products [89]. As a model system, zinc-based microbots were used for the acid-driven motion in the stomach of a mouse, which significantly magnified the binding and withholding of the bots in the stomach wall [90]. Enteric microbots can accurately position and control withholding in required areas of the gastrointestinal tract of living mice. Tubiform magnesium microbots act as a vigorous nanobiotechnology tool for integrated and targeted drug delivery [45]. Magneto-aerotactic bacteria can be used to transport drug-loaded nanoliposomes into hypoxic regions of tumours. To stop haemorrhage through blood vessels in mice and pigs, thrombin was soaked up on adsorbent particles. In this approach, spermatozoon microbots were used as multifunctional surface micro-rollers for intravascular drug delivery. The drug release mechanism was achieved under visible light, near-infrared light, and magnetic fields. Drug Delivery capsules utilized a remote-controlled triggering mechanism to move and eject the drug actively for one time in a controlled amount at the targeted site [91]. Chemotactic bacteria were used to transfer potential drug-loaded nanoparticles using magnetic fields for a controlled and directional movement so that they can reach their specific sites before liberating the prospective drug cargo.

Nanorockets are another form of nanobot that proved to be highly resourceful for the direct delivery of drugs to the tumour cells. Nanorockets developed from layer-by-layer chitosan/alginate assembly are designed such that the inner layer of the framework can incorporate platinum nanoparticles stabilized with poly (diallyl-dimethylammonium-chloride) while the outer layer of the framework acting as a container loads doxorubicin [92]. These nanorockets have the potential to break in HeLa cells and discharge doxorubicin. The maximum speed nanorockets can propel is  $74 \mu\text{m/s}$  and can easily cover a distance of more than 30 cm. Like nanorockets, rolled microtubes of Ti/Fe/Pt alloy have immense potential for drug delivery. It entraps the drugs within the microtubes through laminar fluid motion. This design was pre-functionalized and optimized with surfactants and fuel



concentration to produce highly efficient nano-cargoes. These drug-loaded nano-cargoes can propel up to the speed of 275  $\mu\text{m/s}$  [92].

### 3.2. Biosensing and Isolation

The simplest and most widely used method for biosensing is fluorescence-based sensing owing to its background-free working principle. Merkoçi and co-workers demonstrated microarray-based immune sensing with microbots. Antibodies were labelled with the fluorophore and when they bound with the target molecules, the fluorescence intensity was measured. Capturing and transporting proteins through microbots can easily be detected in the presence of highly efficient fluorescent tags [93]. Furthermore, the fluorescence on-off strategy is also applicable for biosensing where de Ávila and co-workers demonstrated nanobots following ultrasound propulsion having a biosensing property against miRNA. It senses miRNA intracellularly and has real-time operations (Fig. 5f, 5g) [94]. Similarly, they also used the fluorophore fluorescein amine-coated microbots for the detection of nerve agents using stimulants like sarin and soman following the fluorescent on-off strategy. Microbots can be surface functionalized with fluorescent tags for protein detection that can easily perform capturing and transportation of specific proteins [93].

A new approach of biosensing was made where the identification of proteins and hybridized nucleic acids were done based on their motion. The functionalized particles have been bonded specifically to the micro/nanobots propelling in the presence of chemicals and their propulsion speed can also be lowered [95]. With this idea, Yu et al. developed a microbots-based sensor to identify the concentration of cancer biomarkers. Secondary antibody-modified microspheres were loaded on cancer biomarkers leading to an increase and decrease in the dimension and speed of the microbots respectively. By following this idea, it was easy to detect the cancer biomarkers within 5 min [96,97].

Another approach of target isolation via microbots can also be applied through lab-on-a-chip diagnostic devices. This approach has the advantages of capturing, integrating, releasing, transporting and detecting targets in micro-ranged channels or reservoirs [98,99]. Further, biosensing can also be done via electrogenerated chemiluminescence (ECL). These microswimmers were propelled by bipolar electrochemistry, exhibiting ECL and were also able to monitor the glucose concentration in phosphate buffer saline (PBS) solution [93].

Micro/nanobots generally fall within the size range of most biomolecules. So, it can perform better in terms of sensing and segregating cells as well as biomolecules from biological samples. Micro/nanobots are investigated to be able to capture and transport various biomolecules and cells based on donor-receptor interaction. *E. coli* expressing polysaccharides on the surface of cells are identified and isolated efficiently using

lectin bioreceptor functionalized catalytic nanobots of Au/Ni/Pt. Lectin receptors like ConA-based micro/nanobots can selectively pick the target bacterial pathogens like *S. cerevisiae* and *S. aureus* from a human urine sample [100]. Attempts are being made to develop nanobots for sensing applications using a sequence of oligonucleotides. For example, single-stranded DNA functionalized mesoporous silica nanobots for successfully isolating DNA and other functional oligonucleotides from biological fluid samples (Fig. 5d, 5e) [101]. Aptamers are a kind of functional oligonucleotides having a very high affinity towards proteins. So, these aptamers can be optimized and incorporated with nanobots for high-profile identification and isolation of various types of proteins and peptides from biological samples [102].

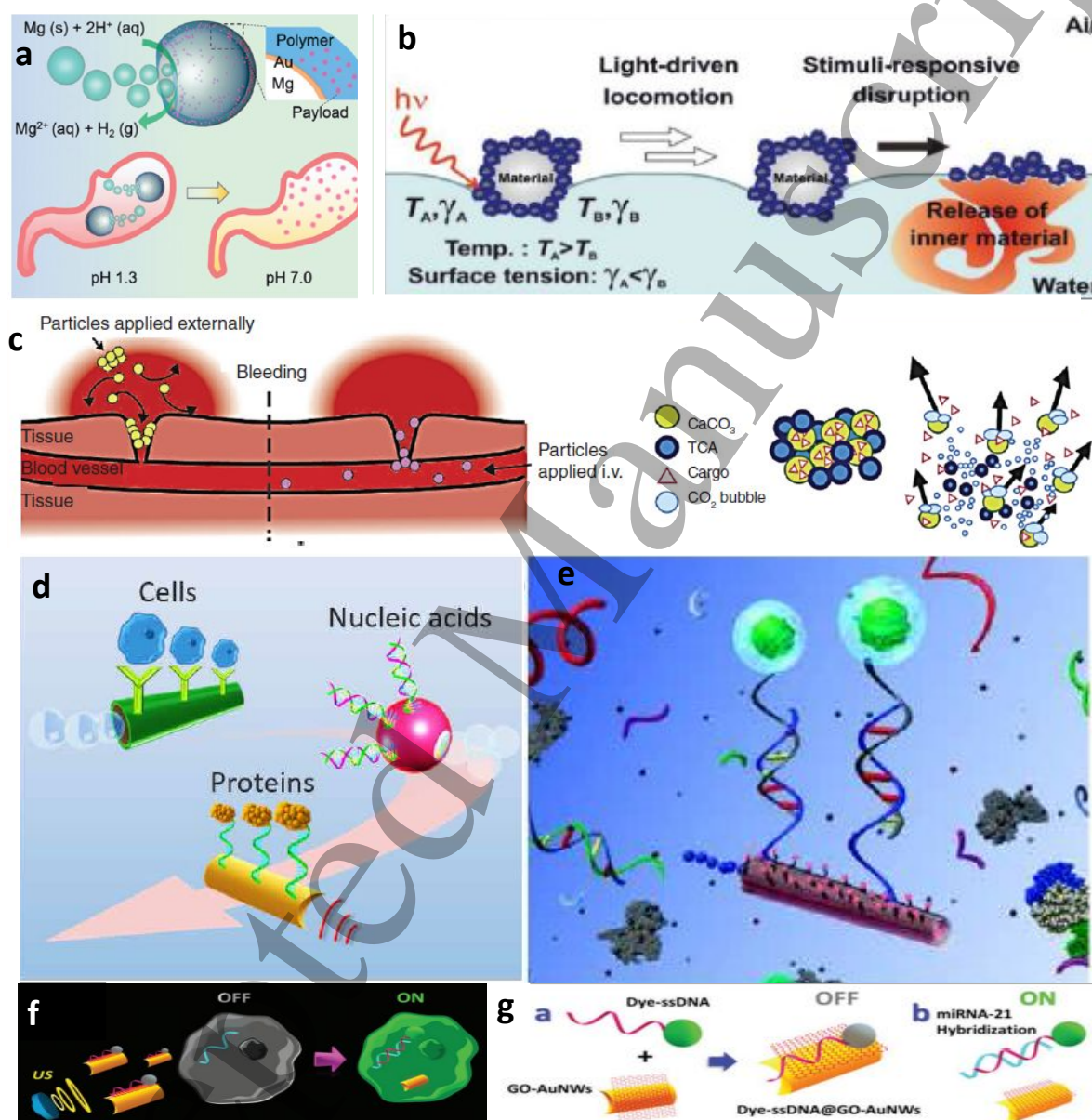
Gastric, pancreatic, and colorectal tumour cells overexpress carcinoembryonic antigens within the system. The anti-carcinoembryonic antigen monoclonal antibody-based micro or nanorockets are developed to capture these tumour cells. The monoclonal antibody has a high affinity for a carcinoembryonic antigen which becomes the prime mechanism for sensing, sorting, and isolation of tumour cells by micro or nanorockets. Nanobots are shown to be highly efficient and specific in sensing and isolation of biomolecules and cells. This capability of nanobots is explored further to develop integrated bioanalytical microchip devices [103].

The successful seizure and separation of target analytes in complex biological media by autonomously propelled microbots can be executed for several biosensing applications towards the exact diagnosis of diseases [104,105]. The sensing quality, mobility, robust binding and transfer capacities of micro/nanobots led to several ways for the identification and separation of biological targets like proteins, nucleic acids, and cancer cells in unprocessed body fluids [97,106,107]. Synthesis of functionalized microbots displayed successful isolation of sensitive and localized thrombin from biological samples [108]. Similarly, a tubular micro-rocket was synthesized with targeting ligands like antibodies for the recognition and isolation of target cancer cells.[109] The nanobots-based target isolation method could be integrated into lab-on-a-chip diagnostic devices, incorporating independent detection, capture, active transport, and release functions inside their discrete tapered micro-passages. This Micro-engine biosensing approach could perceive salient implementations for profiling miRNAs expression at a single-cell level at different clinical layouts [110,111].

Amongst all the noble metal-based nanoparticles, Au nanoparticles have persistently been used for biosensor implementations because of their biocompatibility, optical and electronic properties and comparatively simple production and modification [18]. Au nanoparticles can act as service tags when linked to secondary antibodies in our DNA strands. Au nanoparticles have the potential to transport electrons in a

wide range of electroactive biological species acting as nano-biosensor. Au nanoparticles can also act as electron shuttles [112,113]. Semiconductor nanocrystals called quantum dots (QDs) can be utilized as nanomaterials for biosensing applications. Magnetic nanoparticles could act as a possible substitute for fluorescently labelled biosensing devices. Systematic isolation of DNA strands in composite biological fluids could be attained in a quick and structured way, using silica or Au-coated shell nanoparticles. Magnetic labels were

of interest for biosensing applications as biological entities did not show any magnetic behaviour or susceptibility and therefore no interference happened during signal capturing. The ultra-highly sensitive magnetic-resistant biosensor was developed for detection. These *E. coli* were identified in skimmed milk samples using a Magneto-Geno sensing setup. Nanostructures can be used as electronic electrochemical transducers in biosensors [18].



**Fig. 5:** a) Au-coated Mg microbot with payload embedded with pH-sensitive polymer for gastric acid neutralization. (Adapted with permission from reference [84], Copyright 2017 Wiley); b) Light-driven micro/nanobots performing cargo delivery and release of payload at a targeted site under stimuli-responsive disruption. (Adapted with permission from reference [87], Copyright 2016 Wiley); c) Calcium carbonate ( $CaCO_3$ ) microbot autonomously moving under the impact of  $CO_2$  bubbles generated due to the chemical reaction of tranexamic acid in an aqueous solution for delivering thrombin, a haemostatic drug for wound healing. (Adapted with permission from reference [85], Copyright 2016 Elsevier). d) Micro/nanobots functionalized with different receptors for biosensing various analytes along with different cells, nucleic acids and proteins. (Adapted with

permission from reference [18], Copyright 2017 AAAS); e) Microrockets functionalized with single-stranded DNA for selective hybridization and nucleic acid isolation. (Adapted with permission from reference [101], Copyright 2011 American Chemical Society). f-g) Intracellular detection of miRNAs by US-propelled ssDNA@GO-functionalized gold nanobots and its fabrication process (Adapted with permission from reference [94], Copyright 2015 American Chemical Society).

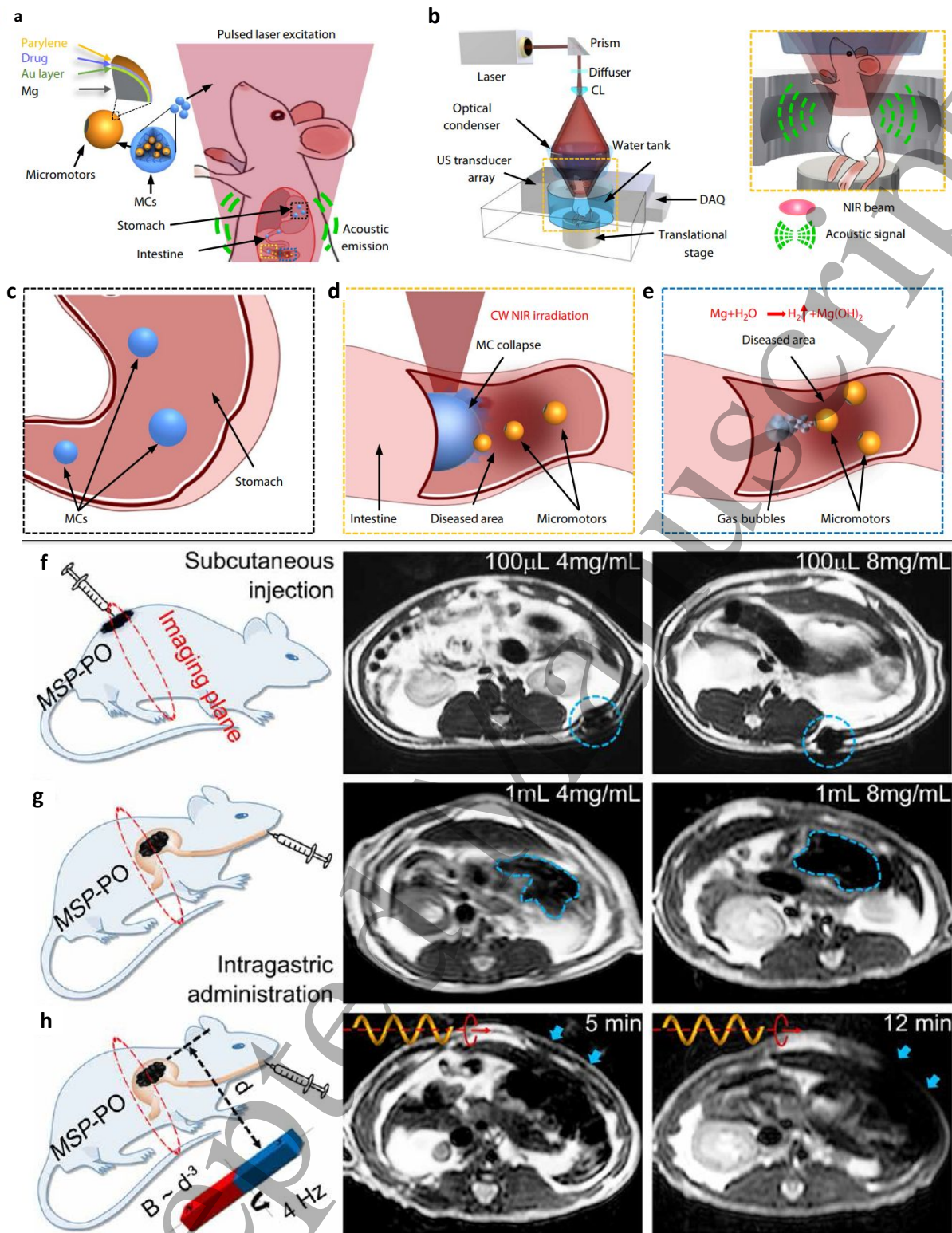
3.3. Bioimaging

Advancement in micro/nanobots is found to act as a boon for the field of bioimaging and medical imaging technology. Because there are several micro/nanobots available that could serve as a new class of imaging agents and/or contrast agents having minimal invasive methodology under natural physiological and pathological conditions. Actually, for the tracing of the movement of bots, and visualization of the route of propagation in real-time, the observable placing of autonomous micro/nanobots is crucial. Therefore, diverse imaging modalities like optical microscopy imaging, fluorescence imaging, magnetic resonance imaging, ultrasound imaging, photoacoustic computed tomography imaging, and computed tomography have been tried to spot micro/nanobots for the same. Additionally, it was evident that functionalized micro/nanobots have a high degree of sensitivity and better selectivity towards some of the pathological factors such as temperature, pH,  $H_2O_2$ , and cellular microenvironment. This provides an opportunity for micro/nanobots to potentially differentiate biological tissues via the above-mentioned imaging much more efficiently. The challenges to the current imaging system of obtaining the finest information on the vascular system have been overcome by the introduction of magnetic nanobots that served as a contrast agent for magnetic resonance imaging (MRI). For example, microbots developed from silica nanosphere converters coated with catalase were demonstrated as effective agents for magnetic resonance imaging. This catalase-coated nanosphere has a significant ability to break down inflammation related to  $H_2O_2$  into  $H_2O$  and  $O_2$ , forming micron-size bubbles. Thus, these nanospheres could also be used as bubble-based ultrasound imaging. A standard prototype of improved and efficient scanning probe

techniques was developed by engineering tiny robots with nano-optical elements, which might surpass the challenges faced by optical imaging devices [114].

Recently, microbot capsules were administered in mice and illuminated with near-infrared (NIR). This NIR triggered the autonomous propulsion of microbots and also facilitated real-time imaging of targeted regions. Microbot capsules when irradiated with continuous-wave (CW) NIR at the target site, the capsule collapsed and released the microbots (activating the movement of the bots). Autonomous movement of the microbots consequently promoted efficient retention and cargo delivery inside GI (Fig. 6a-e) [115]. Furthermore, for structural analysis of micro/nanobots of more than 200 nm and to trace their instantaneous movement practices, optical microscopy imaging was performed. Fluorescence imaging was expedited for the visualization and constant tracing of micro/nanobots that offers better scope to employ them for several biomedical applications [116]. For magnetic resonance imaging, nanobots were demonstrated to be extremely valuable [117]. In-vivo propulsion of magnetic microbots could be accurately managed due to the MRI visualization, and hence have likely been useful for biomedical applications like designated drug conveyance. Similarly, magnetic micro-swimmers developed from *Spirulina Platensis* for cross-sectional magnetic resonance (MR) imaging delivered in subcutaneous tissue/GI of mice followed by a recording of MR imaging that was used to probe inside the rat stomach and tissue (Fig. 6f-h) [118]. The studies showed that autonomous round nanobots of a high-refractive-index medium might work as portable micro-lenses to screen bigger regions. Polystyrene microspheres covered with Ni and Pt sheets were employed in a directed way to highlight point nanoscopic imaging [119].





**Fig. 6:** a) Schematic representation of the functioning of microbot bioimaging system based on photoacoustic computed tomography (PACT). Microbot Capsules (MCs) were administered in mice and illuminated with near-infrared (NIR). The NIR triggers the autonomous propulsion of microbots and also facilitates real-time imaging of targeted regions simultaneously. b) Schematic of PACT-based microbot bioimaging setup for bioimaging gastrointestinal tract (GI). c) Representation of coating on microbots that prevents its degradation inside GI. d) Microbot capsules when irradiated with continuous-wave (CW) NIR at

a target site, the capsule collapses and releases the microbots (activating the movement of the bots). e) Autonomous movement of the microbots promoted efficient retention and cargo delivery inside GI. (a-e) Adapted with permission from reference [115], Copyright 2019 AAAS. f) Magnetic micro-swimmers developed from *Spirulina Platensis* for cross-sectional magnetic resonance (MR) imaging delivered in the subcutaneous tissue of mice in two different concentrations. g) MR imaging through MSP when administered inside the stomach of mice in two different concentrations. h) MR imaging inside the rat stomach under a similar condition when subjected to actuation and steering through a rotating magnetic field. (f-h) Adapted with permission from reference [63], Copyright 2017 AAAS.

3.4. Biopsy and nano-surgery

Precision is the key to the successful surgery of complex tissues and organs. Minimal-invasive surgery is very challenging using current in-practice techniques and technologies. However, the use of micro/nanobots can overcome this challenge easily [120]. Magnetic nanojets and drillers are propelled and guided through the biological systems under external magnetic fields. An example of this is Ti/Cr/Fe and In GaAs/GaAs/CrPt micro-driller which can reach up to and drill in the HeLa cells. Mostly, nanobots are chemically driven such as using hydrogen peroxide or hydrazine as a propulsion mechanism which is toxic to the biological system, thus hindering most of the in vivo biomedical applications. So particular attention is given to the physically driven nanobots, having the ability to be modulated, controlled, and guided externally. An attempt has been made to modify and optimise these nanobots and micro-drillers. An iron oxide coating is done on Ti/Cr nanotubes before it is rolled up to obtain sharp tips micro-drillers. This magnetically functionalized the micro-drillers. This setup has been experimented with to drill porcine liver tissues. As the micro-driller reached the porcine liver tissue a rotational magnetic field is applied to have its centre at the liver tissue itself (Fig. 7e) [121]. This reorients the micro-drillers horizontally and vertically while moving towards the porcine liver tissues and ultimately starts the drilling process. The diameter of the drill remains the same as the diameter of the driller thus providing flexibility to surgical demands of adjusting drill diameter. Changing the diameter of magnetic micro-drillers will change the drill size in tissues and cells. Another excellent example of physically driven nanobots is ultrasound-assisted nano-surgery. Ultrasound-driven micro bullets are a very powerful nano-surgical tool (Fig. 7a) [18]. Under-applied ultrasound triggers the vaporization of perfluorocarbon loaded within micro-bullets producing a powerful thrust and generating ultrafast propulsion of micro-bullets. The speed of propulsion is very high, that is 6.3 m s<sup>-1</sup>, making deep piercing in lamb kidney tissues as reported in an experiment [122].

Nanotechnology has a significant perspective in the area of surgery. With surgical actions becoming increasingly less invasive, nanotechnology could have a huge impression whether it is non-invasive methods executed remotely or the occurrence of nanoscale tools for surgical procedures [123]. Anti-infective materials such as nanosilver could be very favourable, particularly in the growing infection problem of hospitals. Smart micro-sensors were implanted for the examination of the internal state and situation as they were executing the venous approach in adults [124]. Like natural minerals, nano-polymer scaffolds were prepared and utilized to make teeth and bone implants. Titanium (Ti) having great biocompatibility properties has often been used for orthopaedic implants. In Hernia surgery, mesh implant was made out of tetanized synthetics that lowered the scarring and postoperative pain compared to regular plastic meshes because of the biocompatibility of Ti. Semi-autonomous pre-programmed micro/nanobots were developed for executing them in several biomedical-related operations. The micro/nanobots follow ultrasound propulsion synchronized with a computer under the supervision of a surgeon making it applicable in pathological and diagnostic operations. It is used for on-site nanomanipulation works [123]. Miniaturized microscale sensors were used to examine and monitor biological signals like the release of proteins, and antibodies. In response to frequent cardiac ailments, nanotechnology assisted in divulging the process that has been involved in various cardiac diseases thus providing cardiac therapy. Nano-surgery enabled accurate ablation of cellular and subcellular structures without compromising cell viability and with minimal damage to surrounding cells. Nano-tweezers were utilized for the imaging and manipulation of nanosized objects. Untethered micro-robotic tools ranging from nano-drillers to micro-grippers and micro-bullets provide unique abilities for minimally invasive and precision surgery (Fig. 7b-6d) [125,18].

**Table 2:** Applications of different types of micro/nanobots in bioengineering or biomedical sciences with their respective advantages and disadvantages.

S. No.	Micro/Nanobots	Propulsion Method	Bioengineering Application	Advantage	Limitation	Ref.
1.	Gold Nanowires (AuNWs)	Acoustic	Drug Delivery	Easy delivery of doxorubicin drug to the <i>HeLa</i> cells with an average speed of 60 $\mu\text{m/s}$ .	-Biocompatibility of AuNWs in living Cells is not vividly explored.	[112]
2.	Microgrippers	Thermo-biotically	Precision Surgery	-Catch live fibroblast cells in a capillary tube. - Efficiently move out of capillary with caught cells. - Good for tissue biopsy.	-Grippers might become held up in tissue before arriving at the objective.	[125, 126]
3.	Tubular Implantable Microbots	Magnetic	Targeted Drug Delivery and Minimally Invasive Surgery	-Direction and incitation on various tissue types under high shear, can efficiently be explored	-NA	[127]
4.	Micro-bullets	Acoustic droplet vaporization	Precision Surgery	-Application under natural conditions is possible. - Empowers sutureless infusions into the eye. -Biocompatible films of microbots can effectively be filled with drugs and transferred.	-Hard to interface, customize and design, complex.	[18]
5.	Medibots	Magnetic	Precision Surgery and Drug Delivery	-Ultrafast development with paces of more than 6 m/s. - Adequately pushes to profound tissue infiltration, removal and annihilation	- Requires careful handling.	[128]
6.	Enteric Micromotor	pH-induced	Drug Delivery	- Can efficiently per multiple functions like cell cut followed by drug delivery.	-High Cost	[98]
7.	Tubular Microengines	Biological receptor	Biosensing and Isolation	- High specificity. - Controllable maintenance of desired fragments in the gastrointestinal (GI) tract of living mice. - Site-explicit GI conveyance.	-Future examinations are required to approve the conveyance effectiveness and restorative adequacy	[129]
8.	RBC-Mg Janus micromotor	Water-powered	detoxification	-High specificity.	-Complicated bot design.	[50,130]

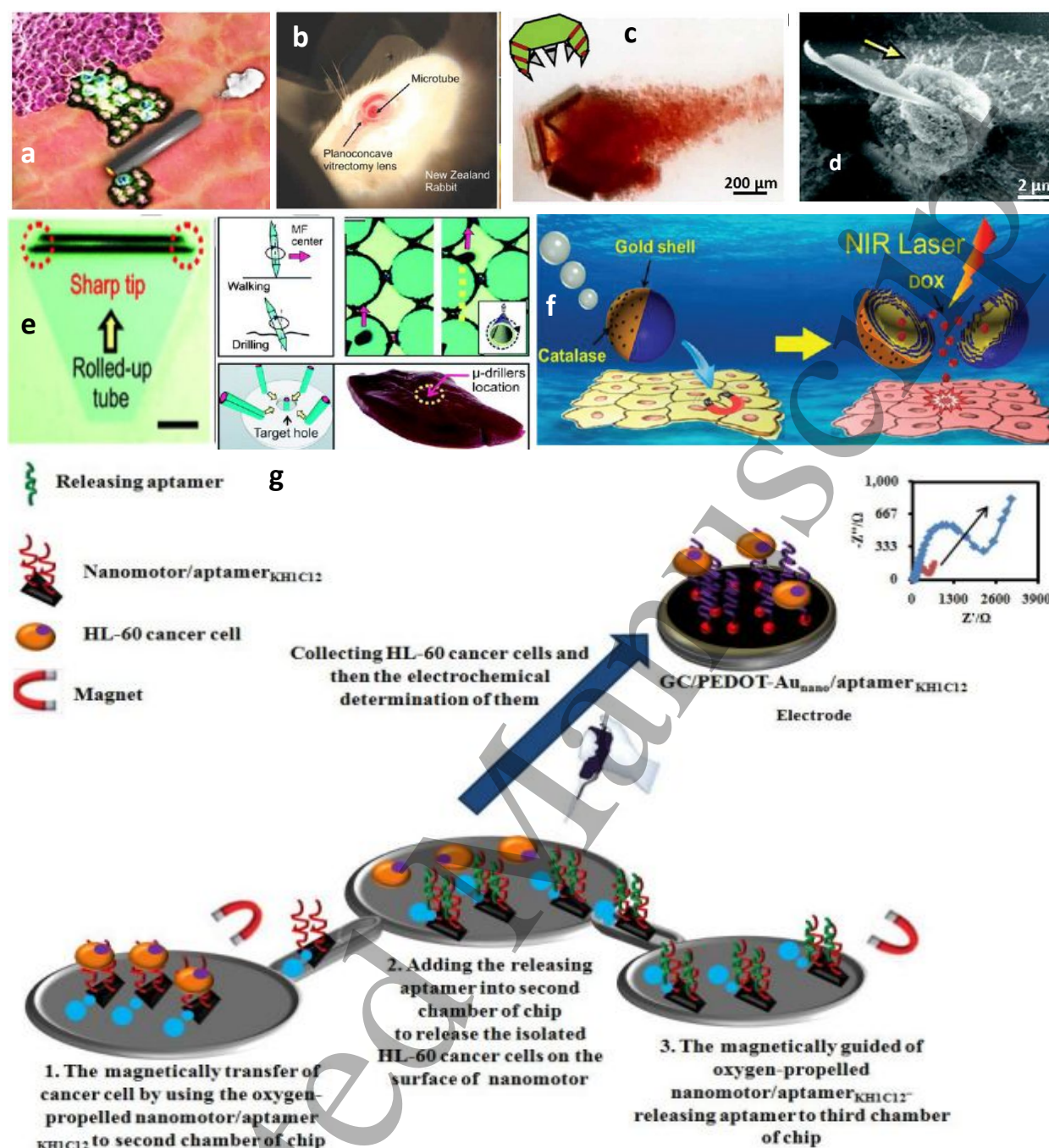
### 3.5. Isolation of cancer cells and cancer therapy

Micro/nanobots can also potentially be effective in cancer cell isolation and therapy through deep tumour penetration [131,109]. In an approach, Hortelão et al. demonstrated the



biocatalytic nanobot made up of mesoporous silica powered by enzyme urease. The core of the shell consisted of mesoporous silica with the anticancer drug (Dox) loaded on its surface. The self-propulsion was observed via optical tracking and dynamic light scattering. The success of this approach was attained when the drug-loaded was released at the targeted HeLa cells and the production of ammonia was observed at a high substrate concentration of urea substrate (Fig. 7f) [132]. After an incubation of 1, 4, 6, and 24 h, a higher content of Dox was detected inside HeLa cells. These enzyme-powered nanobots proved to be an efficient way for drug delivery and cancer therapy [1,133]. In a praiseworthy effort, a new approach was introduced where Au/PEDOT/Pt microbots were developed for the detection of miRNA-21 (cancer biomarker). Before getting into the intracellular studies, an anti-miRNA probe was made consisting of probe DNA/Au/PEDOT/Pt/MMs for the identification of the miRNA-21 target sequence as a proof of concept. Later, intracellular studies were done against 4 different cancer cells.

The DNA/Au/PEDOT/Pt microbots were used for analysis against A-549 (human lung carcinoma cell line), I (human breast cancer cell line), SJSA-1 (human osteosarcoma cell line) and, HT-29 (human colorectal adenocarcinoma cell line). It was observed that the microbots exhibited high potential in the treatment of breast cancer because of their cytotoxic effects and their anti-proliferative properties [59]. In 2017, Amouzadeh Tabrizi et al. demonstrated the MnO<sub>2</sub>-PEI/Ni/Au/aptamer-based nanobots. This technique was used to capture and isolate the HL-60 (cancer cells) from the human serum as the aptamer had a high affinity to capture. The electrochemical impedance spectroscopy technique was used to determine the concentration of cancer cells. The cancer cells were released on the surface of aptamer KH<sub>12</sub>/nanobots into the solution. A good response was observed by the APTA-sensor towards the concentration of HL-60 cells ranging from 2.5×10<sup>1</sup> to 5×10<sup>5</sup> cells mL<sup>-1</sup> with a low limit of detection of 250 cells mL<sup>-1</sup> (Fig. 7g) [134].



**Fig. 7:** a) Microbullets loaded with perfluorocarbon for tissue ablation propelling under acoustic droplet vaporization mechanism; b) Electroforming of implantable magnetic tubular microbots for eye surgery; c) Thermo-biochemically propelling Micro-grippers for catching live fibroblast cells; d) Autonomously moving nano-drillers performing surgery on single cell; (a-d) Adapted with permission from reference [18], Copyright 2017 AAAS. e) Schematics representing magnetically propelled microbots with sharp tips, that are capable of micro-drilling on porcine liver tissues; (Adapted with permission from reference [121], Copyright 2013 Royal Society of Chemistry). f) Schematics representing chemically propelled enzymatic Janus microbots loading doxorubicin and releasing onto the cancer cells using NIR laser; (Adapted with permission from reference [132], Copyright 2014 American Chemical Society). g) Nanobots/nanomotors for isolation of cancer cells via determination of HL-60; EIS (Adapted with permission from reference [134], Copyright 2018 Elsevier).

#### 4. Advantages of micro/nanobots in the biomedical field

These micro/nanobots are very helpful in targeted drug delivery for the treatment of various diseases. With the help of

guided motion, we can load the drugs on the micro/nanobots and deliver them to their targeted sites. They also possess sensing abilities thus; they have widely been used for biosensing purposes. To make these micro/nanobots more biocompatible, they have also been functionalized with different biomaterials to increase their activity and biocompatibility. Their lab-on-a-chip application aids the easy detection and identification of agents causing diseases. Apart from the above-mentioned advantages, the synthesis of micro/nanobots is cost-friendly and their controlled motion & controlled release of drugs has attracted scientists to use them as their tool in the biomedical field for biopsy, nano-surgery, isolation of cancer cells and their therapy as well.

### 5. Toxicological concerns of micro/nanobots in biomedical applications

Nanomaterials are the precursor for the development of most micro/nanobots. These nanomaterials usually trigger various toxic effects in different parts of the body such as respiratory, cardiac, and central nervous systems, etc depending on their composition and surface chemistry. The very small size of these nanomaterials alleviates the chances for their interaction with cells, tissues, organs, and macro or micro biomolecules [135]. Consequently, it causes cell membrane disruption, fibrillation, protein structure destabilization, enzyme inhibition, and thiol crosslinking. Inhibition of the enzyme generates abnormal processes in the metabolic pathway and the generation of a high amount of reactive oxygen species (ROS) [136]. Furthermore, it can lead to DNA damage and cell death. Therefore, these consequences are considered fatal as they might lead to the severe physiological interruption of biological function. Moreover, major factors responsible for the toxicity of nanomaterials depend on their structural and physicochemical properties such as size, charge, surface functionalization, coating, chemical composition, aggregation state, concentration, shell, and period of interaction with biological components. The suitable and judicious modification of these structural and physiological properties is the underlying principle for the development of micro or nanobots which would be biocompatible and environmentally friendly. Another toxicological concern for micro or nanobots is chemical fuel-based self-propulsion mechanisms. The most commonly used fuels are generally toxic such as hydrogen peroxide and hydrazine. These are biologically incompatible and might cause adverse biological impacts even at their lowest concentration. Therefore, it is extremely important to have a vivid knowledge of every factor as well as mechanisms that are involved in the toxicity of nanoscale materials and explore the possible ways to develop biocompatible and efficient micro or nanobots that are potential for disease diagnosis and treatment in various biomedical applications.

### 6. Conclusive remarks and future perspective

Micro/nanobots have made significant advancements in the past decade, leading towards providing overwhelming support and replacement for conventional theranostic techniques. Undoubtedly, the prospects of using these effective and efficient micro/nanobots in various fields are exceptional. Substantial progresses have been seen in the development of self-propulsion bots and devices for on-chip applications. However, in-vivo applications are still facing several major challenges as this area still requires rigorous testing on an animal model for its effectiveness. The composition, architecture and chemical properties are crucial parameters, which need to be optimized and standardized in a better way for real-life applications. Moreover, it is considerably found to be difficult to achieve single-cell precision when applied to a human model. Furthermore, multiple active and passive barriers in the human body are ignored mostly in the case of available bots that might create hindrances in the performance of the micro/nanobots for clinical applications. Monitoring the autonomous movement of micro/nanobots in blood vessels under pulsed blood flow is very difficult, thus simultaneous performing applications like drug delivery and biosensing was found to be challenging in several cases. Therefore, the proof-of-concept research still needs to be translated into real-world biomedical applications after cautiously optimizing bot physicochemical properties. Significant research has been going on in this direction to overcome the challenges associated with real-life applications. The introduction of biologically driven micro/nanobots is one such step towards the development of biocompatible and officious self-propulsion devices for numerous biomedical applications. Therefore, it is to be deemed that more efficient, programmable and smart micro/nanobots will take charge over many areas of conventional bioengineering applications in the near future.

### Abbreviations

H<sub>2</sub>O<sub>2</sub> - Hydrogen Peroxide; O<sub>2</sub> – Oxygen; H<sub>2</sub>O – Water; Pt – Platinum; Ag – Silver; MnO<sub>2</sub> - Manganese Dioxide; Mg – Magnesium; Al – Aluminium; CoFe<sub>2</sub>O<sub>4</sub> - Ferromagnetic Cobalt Ferrite; Pd – Palladium; NaOH - Sodium Hydroxide; DEHPA - di(2-Ethylhexyl) phosphoric acid; PPy-Cd - poly(pyrrole)-cadmium; CdSe-Au-CdSe – Cadmium Selenide-Gold; Ni – Nickel; CoPt – Cobalt Platinum; TiO<sub>2</sub> - Titanium Dioxide; AgCl - Silver Monochloride; HMSNPs - hollow mesoporous silica nanoparticles; GO<sub>x</sub> - Glucose Oxidase; MSD – mean square displacement; DLS – dynamic light scattering; NiTi – nickel titanium; Na<sub>2</sub>Ti<sub>6</sub>O<sub>13</sub> - fibrous titanate; Sr – Strontium; Zn – Zinc; Fe – Iron; ECL - electrogenerated chemiluminescence; PBS – phosphate buffered saline; Cr – Chromium; Dox – Doxorubicin; MCF-7 – Michigan Cancer Foundation – 7; EIS - electrochemical impedance spectroscopy technique; ROS – reactive oxygen species; DNA – deoxyribonucleic acid; siRNA – small

interfering ribonucleic acid; miRNA – micro ribonucleic acid; MRI – Magnetic Resonance Imaging, pDNA – plasmid deoxy ribonucleic acid; ZnO – Zinc Oxide; JHMSNP-Janus Hollow Mesoporous Silica Nanoparticles; GA-Glutaraldehyde; CLSM-Confocal Laser Scanning Microscope; ECL-Electrogenerated Chemiluminescence; PBS-Phosphate-buffered Saline; PEDOT-poly (3,4-ethylenedioxythiophene); EIS-Electrochemical Impedance Spectroscopy; PEI-Polyethyleneimine;

### Acknowledgements

The authors acknowledge the financial support from DBT, Govt. of India (SAN. No. BT/PR40544/COD/139/14/2020).

### Conflict of interest

The authors declare no conflict of interest.

### References

- [1] Hortelão, A. C.; Patiño, T.; Perez-Jiménez, A.; Blanco, A.; Sánchez, S. Enzyme-Powered Nanobots Enhance Anticancer Drug Delivery. *Advanced Functional Materials* 2018, 28 (25), 1705086.
- [2] Fournier-Bidoz, S.; Arsenault, A. C.; Manners, I.; Ozin, G. A. Synthetic Self-Propelled Nanorotors. *Chem. Commun.* 2005, No. 4, 441–443.
- [3] Gautam, P. K.; Shivapriya, P. M.; Banerjee, S.; Sahoo, A. K.; Samanta, S. K. Biogenic Fabrication of Iron Nanoadsorbents from Mixed Waste Biomass for Aqueous Phase Removal of Alizarin Red S and Tartrazine: Kinetics, Isotherm, and Thermodynamic Investigation. *Environmental Progress & Sustainable Energy* 2020a, 39 (2), e13326.
- [4] Shivalkar, S.; Gautam, P. K.; Chaudhary, S.; Samanta, S. K.; Sahoo, A. K. Recent Development of Autonomously Driven Micro/Nanobots for Efficient Treatment of Polluted Water. *Journal of Environmental Management* 2021a, 281, 111750.
- [5] Shivalkar, S.; Gautam, P. K.; Verma, A.; Maurya, K.; Sk, M. P.; Samanta, S. K.; Sahoo, A. K. Autonomous Magnetic Microbots for Environmental Remediation Developed by Organic Waste Derived Carbon Dots. *J Environ Manage* 2021b, 297, 113322.
- [6] Gautam, P. K.; Shivalkar, S.; Banerjee, S. Synthesis of M. Oleifera Leaf Extract Capped Magnetic Nanoparticles for Effective Lead [Pb (II)] Removal from Solution: Kinetics, Isotherm and Reusability Study. *Journal of Molecular Liquids* 2020b, 305, 112811.
- [7] Zhang J, Chen Z, Kankala RK, Wang SB, Chen AZ. Self-propelling micro-/nano-motors: Mechanisms, applications, and challenges in drug delivery. *International Journal of Pharmaceutics*. 2021 Mar 1;596:120275.
- [8] Shivalkar S, Chowdhary P, Afshan T, Chaudhary S, Roy A, Samanta SK, et al. Nanoengineering of biohybrid micro/nanobots for programmed biomedical applications. *Colloids and Surfaces B: Biointerfaces*. 2023 Feb 1;222:113054.
- [9] Liang Z, Tu Y, Peng F. Polymeric Micro/Nanomotors and Their Biomedical Applications. *Advanced Healthcare Materials*. 2021;10(18):2100720.
- [10] Wan M, Li T, Chen H, Mao C, Shen J. Biosafety, Functionalities, and Applications of Biomedical Micro/nanomotors. *Angewandte Chemie International Edition*. 2021;60(24):13158–76.
- [11] Chen Y, Xu B, Mei Y. Design and Fabrication of Tubular Micro/Nanomotors via 3D Laser Lithography. *Chemistry – An Asian Journal*. 2019;14(14):2472–8.
- [12] Liu C, Huang J, Xu T, Zhang X. Powering bioanalytical applications in biomedicine with light-responsive Janus micro-/nanomotors. *Microchim Acta*. 2022 Feb 23;189(3):116.
- [13] Zhuang R, Zhou D, Chang X, Mo Y, Zhang G, Li L. Alternating Current Electric Field Driven Topologically Defective Micro/nanomotors. *Applied Materials Today*. 2022 Mar 1;26:101314.
- [14] Chang X, Feng Y, Guo B, Zhou D, Li L. Nature-inspired micro/nanomotors. *Nanoscale*. 2022 Jan 6;14(2):219–38.
- [15] Purcell, E. M. Life at Low Reynolds Number. *American Journal of Physics* 1977, 45 (1), 3–11.
- [16] Lee, T.-C.; Alarcón-Correa, M.; Miksch, C.; Hahn, K.; Gibbs, J. G.; Fischer, P. Self-Propelling Nanomotors in the Presence of Strong Brownian Forces. *Nano Lett.* 2014, 14 (5), 2407–2412.
- [17] Dey, K. K.; Senapati, K. K.; Phukan, P.; Basu, S.; Chattopadhyay, A. Stable Magnetic Chemical Locomotive with Pd Nanoparticle Incorporated Ferromagnetic Oxide. *J. Phys. Chem. C* 2011, 115 (26), 12708–12715.
- [18] Li, J.; Esteban-Fernández de Ávila, B.; Gao, W.; Zhang, L.; Wang, J. Micro/Nanorobots for Biomedicine: Delivery, Surgery, Sensing, and Detoxification. *Science Robotics* 2017, 2 (4), eaam6431.
- [19] Hu, M.; Ge, X.; Chen, X.; Mao, W.; Qian, X.; Yuan, W.-E. Micro/Nanorobot: A Promising Targeted Drug Delivery System. *Pharmaceutics* 2020, 12 (7), 665.
- [20] Gautam, P. K.; Shivalkar, S.; Samanta, S. K. Environmentally Benign Synthesis of Nanocatalysts: Recent Advancements and Applications. In *Handbook of Nanomaterials and Nanocomposites for Energy and Environmental Applications*; Kharisova, O. V., Martínez, L. M. T., Kharisov, B. I., Eds.; Springer International Publishing: Cham, 2020c; pp 1–19.

- [21] Wang, J.; Manesh, K. M. Motion Control at the Nanoscale. *Small*2010, 6 (3), 338–345.
- [22] Guix, M.; Mayorga-Martinez, C. C.; Merkoçi, A. Nano/Micromotors in (Bio)Chemical Science Applications. *Chem. Rev.*2014, 114 (12), 6285–6322.
- [23] Wang, W.; Li, S.; Mair, L.; Ahmed, S.; Huang, T. J.; Mallouk, T. E. Acoustic Propulsion of Nanorod Motors inside Living Cells. *Angew Chem Int Ed Engl*2014, 53 (12), 3201–3204.
- [24] Wang, Y.; Hernandez, R. M.; Bartlett, D. J.; Bingham, J. M.; Kline, T. R.; Sen, A.; Mallouk, T. E. Bipolar Electrochemical Mechanism for the Propulsion of Catalytic Nanomotors in Hydrogen Peroxide Solutions. *Langmuir*2006, 22 (25), 10451–10456.
- [25] Wang, H.; Zhao, G.; Pumera, M. Beyond Platinum: Bubble-Propelled Micromotors Based on Ag and MnO<sub>2</sub> Catalysts. *J. Am. Chem. Soc.*2014, 136 (7), 2719–2722.
- [26] Wang, H.; Gu, X.; Wang, C. Self-Propelling Hydrogel/Emulsion-Hydrogel Soft Motors for Water Purification. *ACS Appl. Mater. Interfaces*2016, 8 (14), 9413–9422.
- [27] Khezri, B.; Novotný, F.; Moo, J. G. S.; Nasir, M. Z. M.; Pumera, M. Confined Bubble-Propelled Microswimmers in Capillaries: Wall Effect, Fuel Deprivation, and Exhaust Product Excess. *Small*2020, 16 (27), 2000413.
- [28] Hu L, Wang N, Tao K, Miao J, Kim YJ. Circular steering of gold–nickel–platinum micro-vehicle using singular off-centre nanoengine. *International Journal of Intelligent Robotics and Applications*. 2021 Mar;5(1):79-88.
- [29] Reddy NK, Clasen C. Self-propelling micro-disks. *Korea-Australia rheology journal*. 2014 Feb;26(1):73-9.
- [30] Soler L, Martínez-Cisneros C, Swiersy A, Sánchez S, Schmidt OG. Thermal activation of catalytic microjets in blood samples using microfluidic chips. *Lab on a Chip*. 2013;13(22):4299-303.
- [31] Gibbs JG, Zhao YP. Autonomously motile catalytic nanomotors by bubble propulsion. *Applied Physics Letters*. 2009 Apr 20;94(16):163104.
- [32] Paxton WF, Kistler KC, Olmeda CC, Sen A, St. Angelo SK, Cao Y, Mallouk TE, Lammert PE, Crespi VH. Catalytic nanomotors: autonomous movement of striped nanorods. *Journal of the American Chemical Society*. 2004 Oct 20;126(41):13424-31.
- [33] Paxton, W. F.; Sen, A.; Mallouk, T. E. Motility of Catalytic Nanoparticles through Self-Generated Forces. *Chemistry*2005, 11 (22), 6462–6470.
- [34] Wang H, Pumera M. Fabrication of micro/nanoscale motors. *Chemical reviews*. 2015 Aug 26;115(16):8704-35.
- [35] Sánchez S, Soler L, Katuri J. Chemically powered micro-and nanomotors. *Angewandte Chemie International Edition*. 2015 Jan 26;54(5):1414-44.
- [36] Dey, K. K.; Bhandari, S.; Bandyopadhyay, D.; Basu, S.; Chattopadhyay, A. The PH Taxis of an Intelligent Catalytic Microbot. *Small*2013, 9 (11), 1916–1920.
- [37] Ban, T.; Yamagami, T.; Nakata, H.; Okano, Y. PH-Dependent Motion of Self-Propelled Droplets Due to Marangoni Effect at Neutral PH. *Langmuir*2013, 29 (8), 2554–2561.
- [38] Chang, S. T.; Paunov, V. N.; Petsev, D. N.; Velev, O. D. Remotely Powered Self-Propelling Particles and Micropumps Based on Miniature Diodes. *Nature Mater*2007, 6 (3), 235–240.
- [39] Liu L, Gao J, Wilson DA, Tu Y, Peng F. Fuel-Free Micro-/Nanomotors as Intelligent Therapeutic Agents. *Chemistry – An Asian Journal*. 2019;14(14):2325–35.
- [40] Lv J, Xing Y, Xu T, Zhang X, Du X. Advanced micro/nanomotors for enhanced adhesion and tissue penetration. *Applied Materials Today*. 2021 Jun 1;23:101034
- [41] Shivalkar, S.; Singh, S. Solid Freeform Techniques Application in Bone Tissue Engineering for Scaffold Fabrication. *Tissue Eng Regen Med*2017, 14 (3), 187–200.
- [42] Jang, B.; Hong, A.; Alcantara, C.; Chatzipirpiridis, G.; Martí, X.; Pellicer, E.; Sort, J.; Harduf, Y.; Or, Y.; Nelson, B. J.; Pané, S. Programmable Locomotion Mechanisms of Nanowires with Semihard Magnetic Properties Near a Surface Boundary. *ACS Appl. Mater. Interfaces*2019, 11 (3), 3214–3223.
- [43] Chen, X.-Z.; Hoop, M.; Shamsudhin, N.; Huang, T.; Özkale, B.; Li, Q.; Siringil, E.; Mushtaq, F.; Di Tizio, L.; Nelson, B. J.; Pané, S. Hybrid Magnetoelectric Nanowires for Nanorobotic Applications: Fabrication, Magnetoelectric Coupling, and Magnetically Assisted In Vitro Targeted Drug Delivery. *Advanced Materials*2017, 29 (8), 1605458.
- [44] Medina-Sánchez, M.; Xu, H.; Schmidt, O. G. Micro- and Nano-Motors: The New Generation of Drug Carriers. *Therapeutic Delivery*2018, 9 (4), 303–316.
- [45] Sitti, M.; Ceylan, H.; Hu, W.; Giltinan, J.; Turan, M.; Yim, S.; Diller, E. Biomedical Applications of Untethered Mobile Milli/Microrobots. *Proc IEEE Inst Electr Electron Eng*2015, 103 (2), 205–224.
- [46] Servant, A.; Qiu, F.; Mazza, M.; Kostarelos, K.; Nelson, B. J. Controlled In Vivo Swimming of a Swarm of Bacteria-Like Microrobotic Flagella. *Advanced Materials*2015, 27 (19), 2981–2988.
- [47] Mou, F.; Li, Y.; Chen, C.; Li, W.; Yin, Y.; Ma, H.; Guan, J. Single-Component TiO<sub>2</sub> Tubular Microengines with Motion Controlled by Light-Induced Bubbles. *Small*2015, 11 (21), 2564–2570.
- [48] Liu, L.; Gao, J.; Wilson, D. A.; Tu, Y.; Peng, F. Fuel-Free Micro-/Nanomotors as Intelligent Therapeutic Agents. *Chem Asian J*2019, 14 (14), 2325–2335.
- [49] Moo, J. G. S.; Presolski, S.; Pumera, M. Photochromic Spatiotemporal Control of Bubble-Propelled Micromotors by a Spiropyran Molecular Switch. *ACS Nano*2016, 10 (3), 3543–3552.
- [50] Wu, Z.; Li, T.; Li, J.; Gao, W.; Xu, T.; Christianson, C.; Gao, W.; Galamyk, M.; He, Q.; Zhang, L.; Wang, J. Turning Erythrocytes into Functional Micromotors. *ACS Nano*2014, 8 (12), 12041–12048.



- [51] Wu Z, Ávila BEF de, Martín A, Christianson C, Gao W, Thamphiwatana SK, et al. RBC micromotors carrying multiple cargos towards potential theranostic applications. *Nanoscale*. 2015 Aug 6;7(32):13680–6.
- [52] Gao C, Lin Z, Wang D, Wu Z, Xie H, He Q. Red Blood Cell-Mimicking Micromotor for Active Photodynamic Cancer Therapy. *ACS Appl Mater Interfaces*. 2019 Jul 3;11(26):23392–400.
- [53] Pantarotto, D.; Browne, W. R.; Feringa, B. L. Autonomous Propulsion of Carbon Nanotubes Powered by a Multienzyme Ensemble. *Chem. Commun.* 2008, No. 13, 1533–1535.
- [54] Goswami, U.; Sahoo, A. K.; Chattopadhyay, A.; Ghosh, S. S. In Situ Synthesis of Luminescent Au Nanoclusters on a Bacterial Template for Rapid Detection, Quantification, and Distinction of Kanamycin-Resistant Bacteria. *ACS Omega* 2018, 3 (6), 6113–6119.
- [55] Sahoo, A. K.; Sharma, S.; Chattopadhyay, A.; Ghosh, S. S. Quick and Simple Estimation of Bacteria Using a Fluorescent Paracetamol Dimer–Au Nanoparticle Composite. *Nanoscale* 2012, 4 (5), 1688–1694.
- [56] Stanton MM, Park BW, Miguel-López A, Ma X, Sitti M, Sánchez S. Biohybrid Microtube Swimmers Driven by Single Captured Bacteria. *Small*. 2017;13(19):1603679.
- [57] Stanton MM, Park BW, Vilela D, Bente K, Faivre D, Sitti M, et al. Magnetotactic Bacteria Powered Biohybrids Target *E. coli* Biofilms. *ACS Nano*. 2017 Oct 24;11(10):9968–78.
- [58] Wu Z, Lin X, Si T, He Q. Recent Progress on Bioinspired Self-Propelled Micro/Nanomotors via Controlled Molecular Self-Assembly. *Small*. 2016 Jun;12(23):3080–93.
- [59] YurdabakKaraca, G.; Kuralay, F.; BingolOzakpinar, O.; Uygun, E.; Koc, U.; Ulusoy, S.; BosgelmezTinaz, G.; Oksuz, L.; Uygun Oksuz, A. Catalytic Au/PEDOT/Pt Micromotors for Cancer Biomarker Detection and Potential Breast Cancer Treatment. *Applied Nanoscience* 2021.
- [60] Tu, Y.; Peng, F.; André, A. A. M.; Men, Y.; Srinivas, M.; Wilson, D. A. Biodegradable Hybrid Stomatocyte Nanomotors for Drug Delivery. *ACS Nano* 2017, 11 (2), 1957–1963. <https://doi.org/10.1021/acsnano.6b08079>.
- [61] Lee, S.; Kim, J.; Kim, J.; Hoshir, A. K.; Park, J.; Lee, S.; Kim, J.; Pané, S.; Nelson, B. J.; Choi, H. A Needle-Type Microrobot for Targeted Drug Delivery by Affixing to a Microtissue. *Advanced Healthcare Materials* 2020, 9 (7), 1901697.
- [62] Walker, D.; Käs Dorf, B. T.; Jeong, H.-H.; Lieleg, O.; Fischer, P. Enzymatically Active Biomimetic Micropropellers for the Penetration of Mucin Gels. *Sci Adv* 2015, 1 (11), e1500501.
- [63] Yan, X.; Zhou, Q.; Vincent, M.; Deng, Y.; Yu, J.; Xu, J.; Xu, T.; Tang, T.; Bian, L.; Wang, Y.-X. J.; Kostarelos, K.; Zhang, L. Multifunctional Biohybrid Magnetite Microrobots for Imaging-Guided Therapy. *Science Robotics* 2017, 2 (12), eaq1155.
- [64] Medina-Sánchez, M.; Schwarz, L.; Meyer, A. K.; Hebenstreit, F.; Schmidt, O. G. Cellular Cargo Delivery: Toward Assisted Fertilization by Sperm-Carrying Micromotors. *Nano Lett.* 2016, 16 (1), 555–561.
- [65] Ahmed, D.; Baasch, T.; Jang, B.; Pane, S.; Dual, J.; Nelson, B. J. Artificial Swimmers Propelled by Acoustically Activated Flagella. *Nano Lett.* 2016, 16 (8), 4968–4974.
- [66] Miyako, E.; Kono, K.; Yuba, E.; Hosokawa, C.; Nagai, H.; Hagihara, Y. Carbon Nanotube–Liposome Supramolecular Nanotrains for Intelligent Molecular-Transport Systems. *Nat Commun* 2012, 3 (1), 1226.
- [67] You M, Chen C, Xu L, Mou F, Guan J. Intelligent Micro/nanomotors with Taxis. *Acc Chem Res*. 2018 Dec 18;51(12):3006–14.
- [68] Singh DP, Uspal WE, Popescu MN, Wilson LG, Fischer P. Photogravitactic Microswimmers. *Advanced Functional Materials*. 2018;28(25):1706660.
- [69] Ren L, Zhou D, Mao Z, Xu P, Huang TJ, Mallouk TE. Rheotaxis of Bimetallic Micromotors Driven by Chemical–Acoustic Hybrid Power. *ACS Nano*. 2017 Oct 24;11(10):10591–8.
- [70] Schattling PS, Ramos-Docampo MA, Salgueiriño V, Städler B. Double-Fueled Janus Swimmers with Magnetotactic Behavior. *ACS Nano*. 2017 Apr 25;11(4):3973–83.
- [71] Peng F, Tu Y, Men Y, van Hest JCM, Wilson DA. Supramolecular Adaptive Nanomotors with Magnetotaxis Behavior. *Advanced Materials*. 2017;29(6):1604996.
- [72] Guo, J.; Lin, Y. One-dimensional micro/nanomotors for biomedicine: delivery, sensing and surgery. *Biomater Transl*. 2020, 1(1), 18–32.
- [73] Magdanz, V.; Sanchez, S.; Schmidt, O. G. Development of a Sperm-Flagella Driven Micro-Bio-Robot. *Advanced Materials* 2013, 25 (45), 6581–658.
- [74] Zhang L, Abbott JJ, Dong L, Kratochvil BE, Bell D, Nelson BJ. Artificial bacterial flagella: Fabrication and magnetic control. *Applied Physics Letters*. 2009 Feb 13;94(6):064107.
- [75] Mhanna R, Qiu F, Zhang L, Ding Y, Sugihara K, Zenobi-Wong M, et al. Artificial Bacterial Flagella for Remote-Controlled Targeted Single-Cell Drug Delivery. *Small*. 2014;10(10):1953–7.
- [76] Li T, Li J, Morozov KI, Wu Z, Xu T, Rozen I, et al. Highly Efficient Freestyle Magnetic Nanoswimmer. *Nano Lett*. 2017 Aug 9;17(8):5092–8.
- [77] Li T, Li J, Zhang H, Chang X, Song W, Hu Y, et al. Magnetically Propelled Fish-Like Nanoswimmers. *Small*. 2016;12(44):6098–105.
- [78] Sanchez, S.; Solovev, A. A.; Mei, Y.; Schmidt, O. G. Dynamics of Biocatalytic Microengines Mediated by



- Variable Friction Control. *J. Am. Chem. Soc.* 2010, 132 (38), 13144–13145.
- [79] Wang Q, Jin D, Wang B, Xia N, Ko H, Ip BYM, et al. Reconfigurable Magnetic Microswarm for Accelerating tPA-Mediated Thrombolysis Under Ultrasound Imaging. *IEEE/ASME Transactions on Mechatronics*. 2022 Aug;27(4):2267–77.
- [80] Ji F, Jin D, Wang B, Zhang L. Light-Driven Hovering of a Magnetic Microswarm in Fluid. *ACS Nano*. 2020 Jun 23;14(6):6990–8.
- [81] Krishna, G.; Mary, L. R.; Jerome, K. Nanobots for Biomedical Applications. In *Proceedings of the 2019 9th International Conference on Biomedical Engineering and Technology; ICBET' 19; Association for Computing Machinery: New York, NY, USA, 2019; pp 270–279.* <https://doi.org/10.1145/3326172.3326189>.
- [82] Martel, S.; Mohammadi, M.; Felfoul, O.; Lu, Z.; Pouponneau, P. Flagellated Magnetotactic Bacteria as Controlled MRI-Trackable Propulsion and Steering Systems for Medical Nanorobots Operating in the Human Microvasculature. *Int J Rob Res* 2009, 28 (4), 571–582. <https://doi.org/10.1177/0278364908100924>.
- [83] Peng, F.; Tu, Y.; van Hest, J. C. M.; Wilson, D. A. Self-Guided Supramolecular Cargo-Loaded Nanomotors with Chemotactic Behavior towards Cells. *Angew Chem Int Ed Engl* 2015, 54 (40), 11662–11665.
- [84] Li, J.; Angsantikul, P.; Liu, W.; Esteban-Fernández de Ávila, B.; Thamphiwatana, S.; Xu, M.; Sandraz, E.; Wang, X.; Delezuk, J.; Gao, W.; Zhang, L.; Wang, J. Micromotors Spontaneously Neutralize Gastric Acid for PH-Responsive Payload Release. *Angewandte Chemie International Edition* 2017, 56 (8), 2156–2161.
- [85] Baylis, J. R.; Chan, K. Y. T.; Kastrup, C. J. Halting Hemorrhage with Self-Propelling Particles and Local Drug Delivery. *Thromb Res* 2016, 141 Suppl 2, S36–39.
- [86] Hansen-Bruhn, M.; de Ávila, B. E.-F.; Beltrán-Gastélum, M.; Zhao, J.; Ramírez-Herrera, D. E.; Angsantikul, P.; Vesterager Gothelf, K.; Zhang, L.; Wang, J. Active Intracellular Delivery of a Cas9/SgRNA Complex Using Ultrasound-Propelled Nanomotors. *Angew Chem Int Ed Engl* 2018, 57 (10), 2657–2661.
- [87] Paven, M.; Mayama, H.; Sekido, T.; Butt, H.-J.; Nakamura, Y.; Fujii, S. Light-Driven Delivery and Release of Materials Using Liquid Marbles. *Advanced Functional Materials* 2016, 26 (19), 3199–3206.
- [88] Esteban-Fernández de Ávila, B.; Angell, C.; Soto, F.; Lopez-Ramirez, M. A.; Báez, D. F.; Xie, S.; Wang, J.; Chen, Y. Acoustically Propelled Nanomotors for Intracellular siRNA Delivery. *ACS Nano* 2016, 10 (5), 4997–5005.
- [89] Singh, A. V.; Sitti, M. Targeted Drug Delivery and Imaging Using Mobile Milli/Microrobots: A Promising Future Towards Theranostic Pharmaceutical Design. *Curr Pharm Des* 2016, 22 (11), 1418–1428.
- [90] Gao, W.; Dong, R.; Thamphiwatana, S.; Li, J.; Gao, W.; Zhang, L.; Wang, J. Artificial Micromotors in the Mouse's Stomach: A Step toward in Vivo Use of Synthetic Motors. *ACS Nano* 2015, 9 (1), 117–123.
- [91] Esteban-Fernández de Ávila, B.; Ramírez-Herrera, D. E.; Campuzano, S.; Angsantikul, P.; Zhang, L.; Wang, J. Nanomotor-Enabled PH-Responsive Intracellular Delivery of Caspase-3: Toward Rapid Cell Apoptosis. *ACS Nano* 2017, 11 (6), 5367–5374.
- [92] Abdelmohsen LKEA, Peng F, Tu Y, Wilson DA. Micro- and nano-motors for biomedical applications. *J Mater Chem B*. 2014 Apr 3;2(17):2395–408.
- [93] Kong, L.; Guan, J.; Pumera, M. Micro- and Nanorobots Based Sensing and Biosensing. *Current Opinion in Electrochemistry* 2018, 10, 174–182.
- [94] Esteban-Fernández de Ávila, B.; Martín, A.; Soto, F.; Lopez-Ramirez, M. A.; Campuzano, S.; Vázquez-Machado, G. M.; Gao, W.; Zhang, L.; Wang, J. Single Cell Real-Time miRNAs Sensing Based on Nanomotors. *ACS Nano* 2015, 9 (7), 6756–6764.
- [95] Morales-Narváez, E.; Guix, M.; Medina-Sánchez, M.; Mayorga-Martínez, C. C.; Merkoçi, A. Micromotor Enhanced Microarray Technology for Protein Detection. *Small* 2014, 10 (13), 2542–2548.
- [96] Kim, K.; Guo, J.; Liang, Z.; Fan, D. Artificial Micro/Nanomachines for Bioapplications: Biochemical Delivery and Diagnostic Sensing. *Advanced Functional Materials* 2018, 28 (25), 1705867.
- [97] Yu, X.; Li, Y.; Wu, J.; Ju, H. Motor-Based Autonomous Microsensor for Motion and Counting Immunoassay of Cancer Biomarker. *Anal. Chem.* 2014, 86 (9), 4501–4507.
- [98] Li, J.; Thamphiwatana, S.; Liu, W.; Esteban-Fernández de Ávila, B.; Angsantikul, P.; Sandraz, E.; Wang, J.; Xu, T.; Soto, F.; Ramez, V.; Wang, X.; Gao, W.; Zhang, L.; Wang, J. Enteric Micromotor Can Selectively Position and Spontaneously Propel in the Gastrointestinal Tract. *ACS Nano* 2016, 10 (10), 9536–9542.
- [99] García, M.; Orozco, J.; Guix, M.; Gao, W.; Sattayasamitsathit, S.; Escarpa, A.; Merkoçi, A.; Wang, J. Micromotor-Based Lab-on-Chip Immunoassays. *Nanoscale* 2013, 5 (4), 1325–1331.
- [100] Felfoul, O.; Mohammadi, M.; Taherkhani, S.; de Lanauze, D.; Zhong Xu, Y.; Loghin, D.; Essa, S.; Jancik, S.; Houle, D.; Lafleur, M.; Gaboury, L.; Tabrizian, M.; Kaou, N.; Atkin, M.; Vuong, T.; Batist, G.; Beauchemin, N.; Radzioch, D.; Martel, S. Magneto-Aerotactic Bacteria Deliver Drug-Containing Nanoliposomes to Tumour Hypoxic Regions. *Nature Nanotech* 2016, 11 (11), 941–947.
- [101] Kagan, D.; Campuzano, S.; Balasubramanian, S.; Kuralay, F.; Flechsig, G.-U.; Wang, J. Functionalized Micromachines for Selective and Rapid Isolation of Nucleic Acid Targets from Complex Samples. *Nano Lett.* 2011, 11 (5), 2083–2087.
- [102] Nguyen, K. V.; Minter, S. D. DNA-Functionalized Pt Nanoparticles as Catalysts for Chemically Powered Micromotors: Toward Signal-on Motion-Based DNA Biosensor. *Chem. Commun.* 2015, 51 (23), 4782–4784.
- [103] Wang, J. Cargo-Towing Synthetic Nanomachines:

- Towards Active Transport in Microchip Devices. *Lab Chip* 2012, 12 (11), 1944–1950.
- [104] Campuzano, S.; Ávila, B. E.-F. de; Yáñez-Sedeño, P.; Pingarrón, J. M.; Wang, J. Nano/Microvehicles for Efficient Delivery and (Bio)Sensing at the Cellular Level. *Chem. Sci.* 2017, 8 (10), 6750–6763.
- [105] Duan, W.; Wang, W.; Das, S.; Yadav, V.; Mallouk, T. E.; Sen, A. Synthetic Nano- and Micromachines in Analytical Chemistry: Sensing, Migration, Capture, Delivery, and Separation. *Annual Review of Analytical Chemistry* 2015, 8 (1), 311–333.
- [106] Campuzano, S.; Orozco, J.; Kagan, D.; Guix, M.; Gao, W.; Sattayasamitsathit, S.; Claussen, J. C.; Merkoçi, A.; Wang, J. Bacterial Isolation by Lectin-Modified Microengines. *Nano Lett.* 2012, 12 (1), 396–401. <https://doi.org/10.1021/nl203717q>.
- [107] Majumdar, M.; Shivalkar, S.; Pal, A.; Verma, M. L.; Sahoo, A. K.; Roy, D. N. Chapter 15 - Nanotechnology for Enhanced Bioactivity of Bioactive Compounds. In *Biotechnological Production of Bioactive Compounds*; Verma,
- [108] Holzinger, M.; Le Goff, A.; Cosnier, S. Nanomaterials for Biosensing Applications: A Review. *Frontiers in Chemistry* 2014, 2, 63. <https://doi.org/10.3389/fchem.2014.00063>.
- [109] Balasubramanian, S.; Kagan, D.; Jack Hu, C.-M.; Campuzano, S.; Lobo-Castañón, M. J.; Lim, N.; Kang, D. Y.; Zimmerman, M.; Zhang, L.; Wang, J. Micromachine-Enabled Capture and Isolation of Cancer Cells in Complex Media. *Angewandte Chemie International Edition* 2011, 50 (18), 4161–4164. <https://doi.org/10.1002/anie.201100115>.
- [110] Gao, W.; Wang, J. Synthetic Micro/Nanomotors in Drug Delivery. *Nanoscale* 2014, 6 (18), 10486–10494.
- [111] S. Shivalkar, A.K. Sahoo, Bio-Molecules Sensing Using Physical and Microfluidics Devices, in *MEMS Applications in Biology and Healthcare*, AIP Publishing LLC, 2021; pp. 6-1-6–36.
- [112] García-Gradilla, V.; Sattayasamitsathit, S.; Soto, F.; Kuralay, F.; Yardımcı, C.; Wiitala, D.; Galarnyk, M.; Wang, J. Ultrasound-Propelled Nanoporous Gold Wire for Efficient Drug Loading and Release. *Small* 2014, 10 (20), 4154–4159. <https://doi.org/10.1002/sml.201401013>.
- [113] Restrepo-Pérez, L.; Soler, L.; Martínez-Cisneros, C.; Sánchez, S.; Schmidt, O. G. Biofunctionalized Self-Propelled Micromotors as an Alternative on-Chip Concentrating System. *Lab Chip* 2014, 14 (16), 2914–2917.
- [114] Olson, E. S.; Orozco, J.; Wu, Z.; Malone, C. D.; Yi, B.; Gao, W.; Eghtedari, M.; Wang, J.; Mattrey, R. F. Toward in Vivo Detection of Hydrogen Peroxide with Ultrasound Molecular Imaging. *Biomaterials* 2013, 34 (35), 8918–8924.
- [115] Wu, Z.; Li, L.; Yang, Y.; Hu, P.; Li, Y.; Yang, S.-Y.; Wang, L. V.; Gao, W. A Microrobotic System Guided by Photoacoustic Computed Tomography for Targeted Navigation in Intestines in Vivo. *Science Robotics* 2019, 4 (32), eaax0613.
- [116] Gao, C.; Wang, Y.; Ye, Z.; Lin, Z.; Ma, X.; He, Q. Biomedical Micro-/Nanomotors: From Overcoming Biological Barriers to In Vivo Imaging. *Advanced Materials* 2021, 33 (6), 2000512.
- [117] Abdelmohsen, L. K. E. A.; Peng, F.; Tu, Y.; Wilson, D. A. Micro- and Nano-Motors for Biomedical Applications. *J. Mater. Chem. B* 2014, 2 (17), 2395–2408.
- [118] Yan, X.; Zhou, Q.; Vincent, M.; Deng, Y.; Yu, J.; Xu, J.; Xu, T.; Tang, T.; Bian, L.; Wang, Y.-X. J.; Kostarelos, K.; Zhang, L. Multifunctional Biohybrid Magnetite Microrobots for Imaging-Guided Therapy. *Science Robotics* 2017, 2 (12), eaq1155. <https://doi.org/10.1126/scirobotics.aq1155>.
- [119] Safdar, M.; Khan, S. U.; Jänis, J. Progress toward Catalytic Micro- and Nanomotors for Biomedical and Environmental Applications. *Advanced Materials* 2018, 30 (24), 1703660.
- [120] Gultepe, E.; Randhawa, J. S.; Kadam, S.; Yamanaka, S.; Selaru, F. M.; Shin, E. J.; Kalloo, A. N.; Gracias, D. H. Biopsy with Thermally-Responsive Untethered Microtools. *Adv Mater* 2013, 25 (4), 514–519.
- [121] Xi, W.; Solovev, A. A.; Ananth, A. N.; Gracias, D. H.; Sanchez, S.; Schmidt, O. G. Rolled-up Magnetic Microdrillers: Towards Remotely Controlled Minimally Invasive Surgery. *Nanoscale* 2013, 5 (4), 1294–1297.
- [122] Kagan, D.; Benchimol, M. J.; Claussen, J. C.; Chuluun-Erdene, E.; Esener, S.; Wang, J. Acoustic Droplet Vaporization and Propulsion of Perfluorocarbon-Loaded Microbullets for Targeted Tissue Penetration and Deformation. *Angewandte Chemie International Edition* 2012, 51 (30), 7519–7522.
- [123] Singh, S.; Singh, A. Current Status of Nanomedicine and Nanosurgery. *Anesth Essays Res* 2013, 7 (2), 237–242.
- [124] Scott, N. R. Nanotechnology and Animal Health. *Rev Sci Tech* 2005, 24 (1), 425–432.
- [125] Diller, E.; Sitti, M. Three-Dimensional Programmable Assembly by Untethered Magnetic Robotic Micro-Grippers. *Advanced Functional Materials* 2014, 24 (28), 4397–4404.
- [126] Leong, T. G.; Randall, C. L.; Benson, B. R.; Bassik, N.; Stern, G. M.; Gracias, D. H. Tetherless Thermobiochemically Actuated Microgrippers. *PNAS* 2009, 106 (3), 703–708.
- [127] Chatzipirpiridis, G.; Ergeneman, O.; Pokki, J.; Ullrich, F.; Fusco, S.; Ortega, J. A.; Sivaraman, K. M.; Nelson, B. J.; Pané, S. Electroforming of Implantable Tubular Magnetic Microrobots for Wireless Ophthalmologic Applications. *Adv Healthc Mater* 2015, 4 (2), 209–214.
- [128] Srivastava, S. K.; Medina-Sánchez, M.; Koch, B.; Schmidt, O. G. Medibots: Dual-Action Biogenic Microdaggers for Single-Cell Surgery and Drug Release. *Advanced Materials* 2016, 28 (5), 832–837.
- [129] Orozco, J.; Campuzano, S.; Kagan, D.; Zhou, M.; Gao, W.; Wang, J. Dynamic Isolation and Unloading of

- Target Proteins by Aptamer-Modified Microtransporters. *Anal. Chem.* 2011, 83 (20), 7962–7969.
- [130] Wu, Z.; Li, J.; de Ávila, B. E.-F.; Li, T.; Gao, W.; He, Q.; Zhang, L.; Wang, J. Water-Powered Cell-Mimicking Janus Micromotor. *Advanced Functional Materials* 2015, 25 (48), 7497–7501.
- [131] Andhari, S. S.; Wavhale, R. D.; Dhobale, K. D.; Tawade, B. V.; Chate, G. P.; Patil, Y. N.; Khandare, J. J.; Banerjee, S. S. Self-Propelling Targeted Magneto-Nanobots for Deep Tumor Penetration and PH-Responsive Intracellular Drug Delivery. *Sci Rep* 2020, 10 (1), 4703.
- [132] Wu, Y.; Lin, X.; Wu, Z.; Möhwald, H.; He, Q. Self-Propelled Polymer Multilayer Janus Capsules for Effective Drug Delivery and Light-Triggered Release. *ACS Appl. Mater. Interfaces* 2014, 6 (13), 10476–10481.
- [133] Banerjee, S.; Sahoo, A. K.; Chattopadhyay, A.; Ghosh, S. S. Hydrogel Nanocarrier Encapsulated Recombinant IκBα as a Novel Anticancer Protein Therapeutics. *RSC Adv.* 2013, 3 (33), 14123–14131.
- [134] Amouzadeh Tabrizi, M.; Shamsipur, M.; Saber, R.; Sarkar, S. Isolation of HL-60 Cancer Cells from the Human Serum Sample Using MnO<sub>2</sub>-PEI/Ni/Au/Aptamer as a Novel Nanomotor and Electrochemical Determination of Thereof by Aptamer/Gold Nanoparticles-Poly(3,4-Ethylene Dioxythiophene) Modified GC Electrode. *Biosens Bioelectron* 2018, 110, 141–146.
- [135] S. Shivalkar, A. Verma, V. Singh, A.K. Sahoo, Dermatological Delivery of Nanodrugs, in *Nanotechnology in Medicine*, John Wiley & Sons, Ltd, 2021: pp. 259–280.
- [136] A. Verma, S. Shivalkar, M.P. Sk, S.K. Samanta, A.K. Sahoo, Nanocomposite of Ag nanoparticles and catalytic fluorescent carbon dots for synergistic bactericidal activity through enhanced reactive oxygen species generation, *Nanotechnology*. 31 (2020) 405704.
- [137] Wang H, Pumera M. Fabrication of micro/nanoscale motors. *Chemical reviews*. 2015 Aug 26;115(16):8704–35.
- [138] Sánchez S, Soler L, Katuri J. Chemically powered micro-and nanomotors. *Angewandte Chemie International Edition*. 2015 Jan 26;54(5):1414–44.
- [139] Hu L, Wang N, Lim YD, Miao J. Chemical reaction dependency, magnetic field and surfactant effects on the propulsion of disk-like micromotor and its application for *E. coli* transportation. *Nano Select*. 2020 Oct;1(4):432–42.
- [140] Hu L, Miao J, Grüber G. Temperature effects on disk-like gold-nickel-platinum nanoswimmer's propulsion fuelled by hydrogen peroxide. *Sensors and Actuators B: Chemical*. 2017 Feb 1;239:586–96.
- [141] Hu L, Tao K, Miao J, Grüber G. Hydrogen-peroxide-fuelled platinum–nickel–SU-8 microrocket with steerable propulsion using an eccentric nanoengine. *RSC advances*. 2016;6(104):102513–8.
- [142] Wang W, Duan W, Ahmed S, Mallouk TE, Sen A. Small power: Autonomous nano-and micromotors propelled by self-generated gradients. *Nano Today*. 2013 Oct 1;8(5):531–54.

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/373554751>

# Synergistic action of nano silica and w/b ratio on accelerated durability performance of concrete

Article in International Journal of Advanced Technology and Engineering Exploration · September 2023

DOI: 10.19101/IJATEE.2023.10101121

CITATIONS

0

READS

16

3 authors:



Satish Kumar Chaudhary

National Institute of Technology Patna

7 PUBLICATIONS 17 CITATIONS

SEE PROFILE



Ajay KUMAR Sinha

National Institute of Technology Patna

23 PUBLICATIONS 62 CITATIONS

SEE PROFILE



Praveen Anand

National Institute of Technology Patna

5 PUBLICATIONS 19 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Visco-frictional Damping [View project](#)



retrofitting, jacketing of members [View project](#)

## Synergistic action of nano silica and w/b ratio on accelerated durability performance of concrete

Satish Kumar Chaudhary<sup>1</sup>, Ajay Kumar Sinha<sup>2</sup> and Praveen Anand<sup>3\*</sup>

Assistant Engineer, Road Construction Department (RCD), Government of Bihar, Bihar, India<sup>1</sup>

Professor, Department of Civil Engineering, National Institute of Technology, Patna, Bihar, India<sup>2</sup>

Research Scholar, Department of Civil Engineering, National Institute of Technology, Patna, Bihar, India<sup>3</sup>

Received: 07-February-2023; Revised: 23-August-2023; Accepted: 26-August-2023

©2023 Satish Kumar Chaudhary et al. This is an open access article distributed under the Creative Commons Attribution (CC BY) License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

*Cement, which is a primary component of concrete and the manufacture of which generates a significant amount of CO<sub>2</sub>, has an adverse effect on the environment. The environmental effect of manufacturing cement may be decreased by focusing on improving their durability criteria. Sufficient number of nanoparticles may be incorporated to the concrete mix to change the cementitious materials' nano-structure and increase concrete's durability as well. The present article examines the synergistic influence of nano silica (NS) and water-binder (w/b) ratio on concrete subjected to aggressive chemical environment and thus examines the durability of the mix. To study the tendency of binary mix concrete with NS subjected to aggressive environment; acid test, sulphate test and chloride resistance test have been conducted in laboratory. To meet the research objectives the concrete specimen was exposed to 5% sulphuric acid, 5% sodium sulphate and 5% sodium chloride respectively for 60 days. To perform this study, eighteen mix proportion were prepared for M35 concrete by six partial replacement level of cement with NS, viz. 0, 0.5, 1.0, 1.5, 2.0 and 2.5% and three dissimilar w/b ratio, viz. 0.36, 0.40 and 0.44. The percentage drop in weight and compressive strength of concrete specimen after the chemical attack has been measured. The test results showed that there is significant effect of NS and w/b ratio on improving the resistance of concrete towards chemical attack. But the effect of NS inclusion was more prominent than w/b ratio. To gain extra insights for the durability of the mix, flexural tensile test and split tensile strength test was performed. To investigate the morphology and properties of the mix at microscopic level scanning electron microscope (SEM) test was also performed.*

### Keywords

*Nano silica, Acid resistance, Sulphate attack, Chloride resistance, Mix proportion.*

### 1.Introduction

As civil engineering is under constant technological advancements with increasing number of high-rise structures and long-span bridges etc., concrete with greater compressive strength has always been in focus. In certain instances, additional durability features like low water penetrability, resistance to acid, sulphate, and chloride attack, and workability are desirable in addition to compressive strength. Due to its extensive use in constructions, buildings, industries, road bridges, and airport terminals, concrete is one of the most scrutinised materials. Cement plays a significant role among the various materials utilised in the manufacturing of concrete due to its size and adhesiveness.

Concrete without any mineral admixture is not capable to withstand environmental impact. It is vulnerable to the ingress of ions and fluid in extreme environment, due to porous microstructure. In coastal region the ingress of chloride ion in concrete structure through pores lead to corrosion of steel and finally damage of structure [1]. The local environment around the concrete plays a significant role in changing the properties and quality of concrete. In these environments, presence of sulphate either in the ground water or in the soil or in sea water affects the quality of concrete and may prove to be harmful for the underground structures [2, 3]. Concrete infrastructure is susceptible to deterioration due to environmental conditions over time. Common deterioration factors include salt corrosion to reinforced concrete structures, freezing and thawing in cold climates, carbonation from carbon dioxide,

\*Author for correspondence

and chemical attacks from acid solutions. Acid rain contributes to concrete building corrosion. Acid rain is also responsible for degradation of concrete structure constructed in vicinity of industrial area or polluted area by weathering action [4]. The concrete structure, which are constructed in the coastal zone or water retaining structure are more susceptible to aggressive environment. Some common types of chemical attacks which influence the durability of concrete, such as acid attack, sulphate attack, chloride attack and carbonation [5–7]. These drawbacks associated with the cement motivated the need to study advanced mix proportions that would help in increasing the strength as well as the durability of the concrete. To counteract these drawbacks and to produce concrete with improved qualities, several materials known as supplemental cementitious materials (SCMs) are often blended to the concrete mix. Some SCMs include rice husk, fly ash, blast furnace slag, silica fume and nano silica (NS). The SCMs are a waste product, and substituting cement with these materials not only results in high-strength and long-lasting concrete, but is also beneficial for the environment by enhancing sustainability in the construction. Nanoscale research on hydration products (calcium hydroxide, ettringite, mono-sulphate, un hydrated particles, and air spaces) is vital for addressing durability difficulties in concrete and promoting sustainability. Research has been conducted in this area in recent years. Recently, nanotechnology has been employed or proposed for use in a variety of applications, and it has attracted more popularity as a technology for building materials, with both possible benefits and problems highlighted [8–11].

One of the most popular SCMs that is currently being widely used is NS. NS are nanoparticle whose small dosage can significantly improve the concrete properties. With the incorporation of NS along with the micro fibres in the concrete mix, variety of properties on the concrete performance can be investigated. Addition of NS to the concrete is found to be effective in dipping the porosity and also protects the concrete against sulphate attack [12–14]. Researchers found that adding NS improved the compressive and tensile strength of concrete also, particularly in early use. They observed that NS improved cement paste microstructure. It was also concluded that NS concrete resists permeability better [15, 16].

Concrete mixes generated by adding NS as a binder are found to have lower workability as compared to

the plain mixes, thus reducing the initial and final setting time [17, 18]. The available literatures also suggests that even small dosage of NS in the concrete mix are effective in increasing the mechanical properties. It is found that by the addition of 4 percentage (by mass of cement) of NS, 70% enhancement in the compressive strength can be achieved [19].

In this research, the main objective was to focus on the effect of NS and water-binder (w/b) ratio on acid, sulphate and chloride resistance of concrete have been studied. Furthermore, flexural strength test, split tensile strength test have been conducted followed by microstructural investigation of the concrete mix by scanning electron microscope (SEM). The impact of w/b and NS on concrete's resistance to acids, sulphates, and chlorides was also investigated.

The article is organised as follows. Literature review is thoroughly discussed in section 2. Experimental sequence, methodology along with material properties are discussed in section 3. Section 4 presents all the results obtained from the experiments conducted. An extensive discussion of the tests is presented in section 5. A separate discussion section including the outcome and limitations of the study is added in section 6. Lastly, the overall conclusion along with the scope for future work is presented in section 7.

## 2.Literature review

Due to the severe harm that acids add to concrete buildings, acid attacks on hardened concrete have always attracted researchers on improving the acid resistance of the concrete. Due to the expansion of industrial and urban areas, which causes acid media to come into contact with concrete structures, the main reason of acid attacks is the spread of acidic sources [20]. The impact of acid attack on building material has been recognised and studied over centuries. Cement and concrete are negatively impacted by acid precipitation with a pH level in the range of 3.0 to 5.0. Acid attack may be generated from a variety of sources, including silage effluents, acidic waste water, and acidic rain [20–22]. Comparison of biogenic as well as chemical acid attack for evaluating the last stage of corrosion on flexural microstructure of concrete have been addressed [23]. Concrete's alkaline nature makes it extremely susceptible to acidic attack. The acids attack various cement matrix hydration products, resulting in their breakdown and a corresponding decline in the mechanical characteristics of concrete.



Very few of the hydraulic cements, regardless of their kind are able to provide satisfactory and durable resistance to acid attack [24].

Most researchers agree that chloride and sulfate deterioration is the biggest issue for concrete structures, particularly in maritime conditions. Sulphate assault occurs when calcium hydroxide interacts with sulphate ions, converting alumina-containing hydrates to high-sulfate ettringite. When chloride ions contact steel, they de-passivate the surrounding area, causing corrosion in the presence of water/air. Corrosion products are larger than the original steel, causing concrete expansion and spalling. The sulphate attack is a significant threat to the serviceability and durability of concrete. According to some reports, the chemical reactions between sulphates and the hydrated phases of concrete are responsible for causing the sulphate ions to diffuse into the concrete and thus degrade the quality of concrete [25–28]. The preliminary results indicates that sulfur-dioxide contributes significantly in the deterioration of concrete especially in the regions of high pollution. An expanded numerical technique is also available which uses Fick's law and reaction kinetics to estimate concrete deterioration in sodium sulphate solution [29, 30]. The high dilution of chloride and sulphate ions in the marine environment makes it particularly aggressive. Sulphate ions damages the concrete by producing expansive ettringite and gypsum, while chloride ions weaken concrete structures by starting the corrosion of the reinforced steel [31, 32]. One of the most crucial factors that must be taken into consideration while discussing the durability of concrete is chloride attack. As the facts suggests, chloride attack primarily results in reinforcement corrosion significantly. According to statistics, reinforcing corrosion accounts for more than 40% of structural failures [33]. The leaching of calcium hydroxides and the creation of porosity are two other processes that chloride is known to activate from. Complex reactions are involved in calcium silicate hydrate (CSH). Concrete is negatively impacted by the de-calcification effect of NaCl, the creation of porous CSH, and the leaching of Ca (OH)<sub>2</sub> [34, 35]. However, the synergistic impact of mineral admixtures with nanomaterials and their creation process are still unclear. NS has been widely used as a partial substitute for cement and also as a durability modifier for concrete [36]. The synergistic impact of NS and fly ash on hydration process, mechanical property, hardened paste pore size distribution, pozzolanic activity and synergetic effect generation

method were explored [37,38]. Research investigated how NS affects the hydration parameters of binary, ternary, and quaternary mixed mortar and cement paste incorporating nano sized admixtures such fly ash and colloidal nano silica (CNS) [39–41]. The NS demonstrates a higher pozzolanic response even at an early stage, which improves water penetration and chloride resistance [42]. From the extensive literature survey conducted, it was found that though the inclusion of NS as SCM significantly improves the properties of concrete and enhances the durability of concrete as well. However, the reports on the optimum dosage of NS along with optimum percentage of w/b ratio is scarce. Also, most of the researches was focused on attempting to analyse single or double parameters affecting the durability of concrete. This motivated the need to make an extensive effort to analyse more parameters to find the importance of varying percentage of NS along with varying the w/b ratio and determine an optimum dosage combination.

### 3. Materials and methods

#### 3.1 Materials

43 Grade ordinary portland cement (OPC), NS, zone III fine aggregate (FA), coarse aggregate (CA) of 20mm gradation, and super plasticizer (Structuro 203) were the constituent materials employed for this investigation (polycarboxylic based). The OPC 43 employed in this investigation, which complies with [43], has a specific gravity (G) of 3.15, a fineness of 0.225m<sup>2</sup>/g, and a soundness of 0.8% (Autoclave expansion). The consistency of OPC was noted to be 28% while the initial and final setting time was 60 minutes and 275 minutes respectively. The soundness value and the bulk density of OPC was found to be 2.5mm and 1200kg/m<sup>3</sup> respectively. The NS utilised in the experiment has a regular particle size of 30 to 50 nm and was acquired from the Nano Research Lab in Jamshedpur. *Table 1* displays the chemical make-up of the NS and cement employed in this experiment.

The FA was comprised of sand obtained from the Sone River, and sieve testing showed that it was in zone III of the classification system [44]. As CA, locally accessible crumpled stone "with a maximum graded size of 20 mm has been used." The sieve analysis of a sample of CA verified that it was 20 mm in size, graded according to Indian standards (IS) [44]. *Table 2* and *Table 3* show the physical characteristics of FA and CA respectively.

### 3.2 Mix proportion

Following are the steps used during the mixing process. Firstly, 5 minutes were spent on dry mixing of NS and cementitious ingredients, such as cement. Secondly, the specimens were continuously dry mixed with the addition of aggregates, superplasticizer and sand for 3 minutes to achieve a complete blended mix. Eighteen M35 concrete mixes were made according to IS code [45] using six partial replacements of cement with NS: 0%, 0.5%, 1.0%, 1.5%, 2.0%, and 2.5% by weight of cement and three different w/b ratios: 0.36, 0.40, and 0.44. *Figure 1* presents the samples of the casted specimen. The decrement in water cement ratio to increase the strength reduces workability, however high-rate-water-reducer (HRWR) enables it up to 0.36. Further, IS code [46] restricted the cement content (cementitious material) in a mix up to  $450\text{kg/m}^3$ . The mix proportion for M35 concrete with three different w/b ratios has been summarized in *Table 4*. Maximum strength requires proper concrete mixing. First of all, premixing NS powder with half of mixing water and then continues mixing to achieve a uniform dispersion of nano particles then mixing the powder mineral (cement, silica fume) and aggregates for one minute in a drum mixture and adding NS, which was mixed in water, to drum mixture and continue mixing for one minute and finally adding the mixture of remaining half water and super plasticizer to the composition and continue mixing for five minutes to accomplish a fully blended mix.



**Figure 1** Casted cube specimens

### 3.3 Methodology

Due to the lack of widely accepted procedures for testing acid, sulphate, and chloride resistance, tests have been carried out in this study based on the results of previous studies and work that has been documented. Cubic concrete specimens of 150 millimetres cubed were casted in the mould so that the impact of chemical assault could be evaluated. Eighteen cubes were casted for each mix type, hence a total of 324 cubes have been casted for this study. After twenty-four hours, the samples were taken out of the mould and allowed to cure in fresh water at normal temperature for twenty-eight days. The cubes were dried for 2–3 hours after curing. Similarly, acid, sulphate, and chloride resistance tests were also performed. To analyse the impact of acid attack on each mix type, the weight of all six surface dry cubes was obtained and the average value was found out, then compressive strength of three cubes was found and the average values was found. Then after the remaining three cubes were submerged in a bucket of 5% sulphuric acid at pH 1.5–2.5 for 60 days. pH was measured in the Laboratory with the help of pH metres. After 60 days, the cubes were removed from solution, dried, and weighed to determine compressive strength. Similarly, the percentage change in weight and compressive strength of cube specimen were found due to sulphate and chloride attack for each mix type.

To conduct flexural strength test, prism specimens of dimension  $100\text{mm} \times 100\text{mm} \times 500\text{mm}$  prepared for split tensile strength were tested in lab at a curing age of 28 days. The variation in flexural strength with change in percentage NS content and w/b ratio is discussed in the result section. To conduct split tensile strength, every variation of percentage nano-silica content, cylindrical specimen of diameter 150mm and height 300mm were casted and subjected to testing after 28 days of water curing.

**Table 1** Chemical structure of OPC 43 and NS

Type of sample	% (By mass)							
	$\text{K}_2\text{O}+\text{Na}_2\text{O}$	$\text{SiO}_2$	$\text{SO}_3$	$\text{Al}_2\text{O}_3$	$\text{Fe}_2\text{O}_3$	CaO	MgO	Ignition loss
OPC	1.09	22.11	3.46	5.2	3.45	64.34	2.61	1.45
NS	-	95	-	0.02	0.05	0.08	0.1	2.34

**Table 2** Physical property of FA

Parameters	Fineness modulus (FM)	G	Water absorption
Value	2.49	2.66	1.36%

**Table 3** Physical property of CA

Parameters	Impact value	Crushing value	Specific gravity(G)	Water absorption
Value	29%	24%	2.72	0.76%

**Table 4** Mix ratio of M35 concrete

Mix	%NS	w/b	Cement(Kg)	NS(Kg)	HRWR(Kg)	Water(Kg)	FA(Kg)	CA(Kg)
L <sub>A</sub>	0.0	-	425	0	5.1	153	750.38	1156.74
L <sub>B</sub>	0.5	-	422.87	2.13	5.1	153	750.38	1156.74
L <sub>C</sub>	1.0	-	420.75	4.25	5.1	153	750.38	1156.74
L <sub>D</sub>	1.5	0.36	418.62	6.38	5.1	153	750.38	1156.74
L <sub>E</sub>	2.0	-	416.50	8.50	5.1	153	750.38	1156.74
L <sub>F</sub>	2.5	-	414.37	10.63	5.1	153	750.38	1156.74
I <sub>A</sub>	0.0	-	412.50	0	3.3	165	757.05	1132.59
I <sub>B</sub>	0.5	-	410.44	2.06	3.3	165	757.05	1132.59
I <sub>C</sub>	1.0	-	408.38	4.12	3.3	165	757.05	1132.59
I <sub>D</sub>	1.5	0.40	406.31	6.19	3.3	165	757.05	1132.59
I <sub>E</sub>	2.0	-	404.25	8.25	3.3	165	757.05	1132.59
I <sub>F</sub>	2.5	-	402.19	10.31	3.3	165	757.05	1132.59
H <sub>A</sub>	0.0	-	382	0	2.29	168	779.05	1131.31
H <sub>B</sub>	0.5	-	380.09	1.91	2.29	168	779.05	1131.31
H <sub>C</sub>	1.0	-	378.18	3.82	2.29	168	779.05	1131.31
H <sub>D</sub>	1.5	0.44	376.27	5.73	2.29	168	779.05	1131.31
H <sub>E</sub>	2.0	-	374.36	7.64	2.29	168	779.05	1131.31
H <sub>F</sub>	2.5	-	372.45	9.55	2.29	168	779.05	1131.31

#### 4.Results

This section presents the findings from the experimental program after testing the concrete specimens. A comparative study was carried out for the varying percentage of NS and change in w/b ratio on properties of concrete. Percentage decrease in weight and compressive strength of the samples after

the acid, sulphate and chloride attack has been summarized in *Table 5*. The result of flexural strength tests and average split tensile strength for all mixes has been presented in *Table 6*. *Figure 2* and *Figure 3* shows the cube specimen sample after acid and chloride attack respectively.

**Figure 1** Cube specimens after acid attack**Figure 2** Cube specimens after sulphate attack**Table 5** Percentage change in weight and strength of specimen made of different mixes

Mix	Acid attack		Sulphate attack		Chloride attack	
	% wt. loss	% strength loss	% wt. loss	% strength loss	% wt. loss	% strength loss
L <sub>A</sub>	8.76	19.52	6.62	15.86	4.25	10.16
L <sub>B</sub>	7.92	15.88	5.81	11.89	3.93	8.12
L <sub>C</sub>	6.47	12.32	4.75	8.37	3.44	6.65
L <sub>D</sub>	5.42	9.56	3.96	7.51	3.05	5.54
L <sub>E</sub>	4.51	8.78	3.74	7.38	2.94	5.38
L <sub>F</sub>	4.32	8.64	3.62	7.26	2.86	5.25

Mix	Acid attack		Sulphate attack		Chloride attack	
	% wt. loss	% strength loss	% wt. loss	% strength loss	% wt. loss	% strength loss
I <sub>A</sub>	8.82	19.88	6.73	16.25	4.37	10.38
I <sub>B</sub>	7.95	16.05	5.92	12.05	4.05	8.27
I <sub>C</sub>	6.58	12.51	4.81	8.48	3.51	7.06
I <sub>D</sub>	5.45	9.64	4.05	7.62	3.16	5.82
I <sub>E</sub>	4.63	8.91	3.82	7.52	3.04	5.65
I <sub>F</sub>	4.46	8.72	3.73	7.35	2.95	5.48
H <sub>A</sub>	8.9	20.12	6.81	16.41	4.45	10.52
H <sub>B</sub>	8.05	16.31	6.05	12.21	4.13	8.41
H <sub>C</sub>	6.65	12.64	4.89	8.59	3.62	7.16
H <sub>D</sub>	5.51	9.85	4.13	7.81	3.25	6.04
H <sub>E</sub>	4.58	9.05	3.91	7.72	3.16	5.74
H <sub>F</sub>	4.53	8.97	3.84	7.57	3.07	5.66

**Table 6** result of flexural strength tests and average split tensile strength for all mixes

Mix	Flexural Strength	% Increase in flexural strength	Split tensile strength	% Increase in split tensile strength
L <sub>A</sub>	5.21	0	3.31	0
L <sub>B</sub>	5.74	10.28	3.48	5.14
L <sub>C</sub>	6.39	22.61	4.11	24.17
L <sub>D</sub>	6.85	31.45	4.42	33.53
L <sub>E</sub>	6.94	33.28	4.53	36.86
L <sub>F</sub>	7.06	33.57	4.62	39.58
I <sub>A</sub>	5.1	0	3.27	0
I <sub>B</sub>	5.57	9.14	3.39	3.67
I <sub>C</sub>	6.15	20.68	3.95	20.79
I <sub>D</sub>	6.78	32.88	4.35	33.03
I <sub>E</sub>	6.89	35.21	4.42	35.17
I <sub>F</sub>	6.94	36.16	4.59	40.37
H <sub>A</sub>	4.95	0	3.19	0
H <sub>B</sub>	5.5	11.21	3.38	5.96
H <sub>C</sub>	6	21.33	3.76	17.87
H <sub>D</sub>	6.48	30.96	4.05	26.96
H <sub>E</sub>	6.66	34.64	4.21	31.97
H <sub>F</sub>	6.69	35.18	4.36	36.68

## 5. Discussion

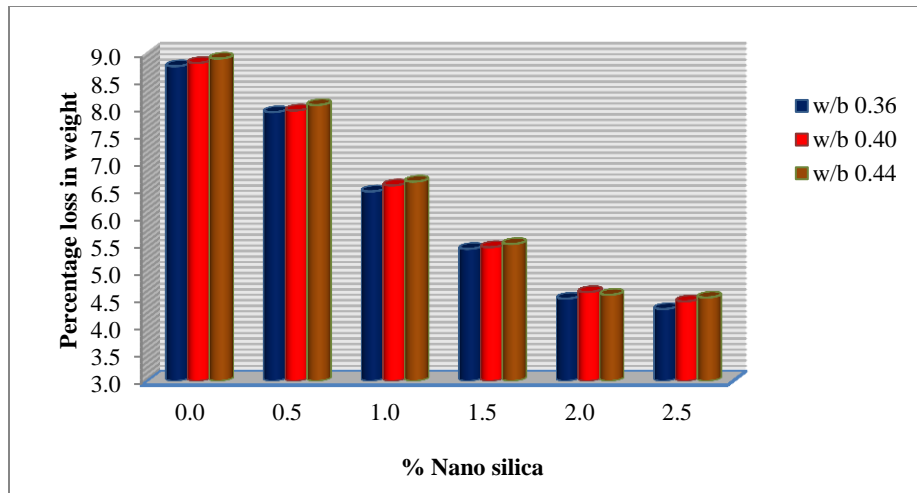
This section discusses the experimental results in detail. The investigation undertaken to analyze the durability of the concrete mix was followed by a sequence of experiments whose results has been discussed.

### 5.1 Acid resistance test result

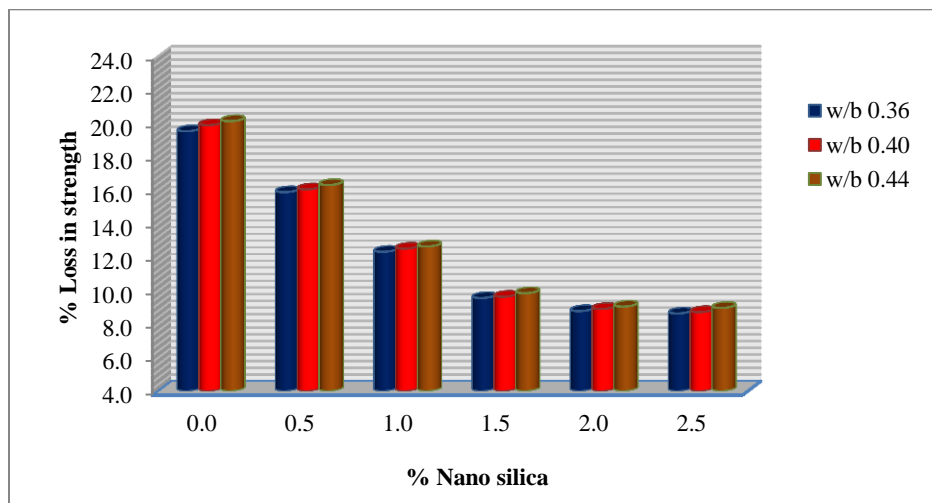
Figure 4 illustrates the % weight loss of cube samples exposed to acid attack and having varied amounts of NS and w/b ratio. The variance in percentage weight loss demonstrates that as the amount of NS grows, weight loss decreases, also, for the same amount of NS, as the w/b ratio rises, the percentage weight loss rises. It is evident from the results that the control mix's percentage weight loss is the highest compared to all other mixes, and that as the amount of NS rises, the percentage weight loss

declines. The large specific surface area of NS, which leads to enhanced pozzolanic activity and because of its size in nanometers; it works as filler that makes concrete denser, is responsible for the concrete's good resistance to acid attack.

Figure 5 provides a visual representation of the proportion of cube specimens that had a reduction in strength as a result of acid assault. These specimens had varied percentages of NS and w/b ratios. The graph demonstrates that a rise in the percentage of NS results in a decrease in the percentage of strength loss, whereas an increase in the w/b ratio results in substantial increase in the percentage strength loss. The pozzolanic nature of NS concrete may be responsible for its superior performance against acid attack.



**Figure 4** % Loss in weight with varying % of NS and w/b ratio due to acid attack

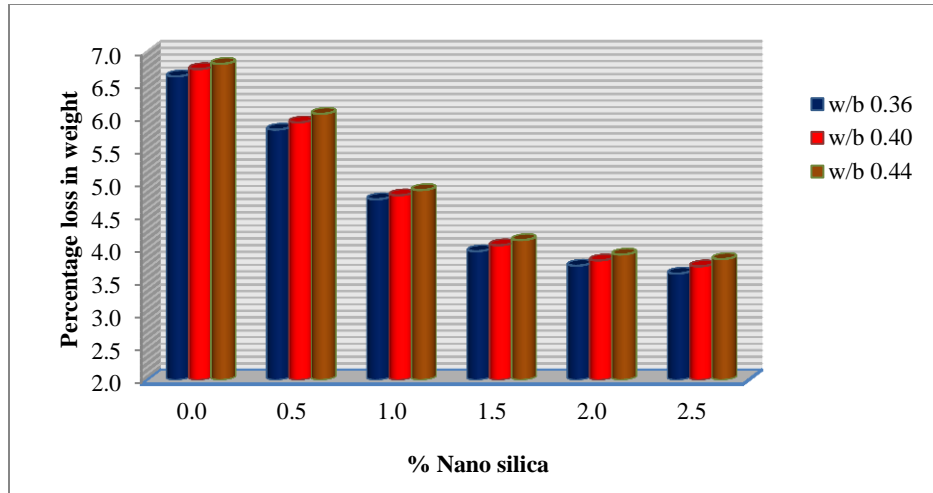


**Figure 5** Loss in strength with varying % of NS and w/b ratio due to acid attack

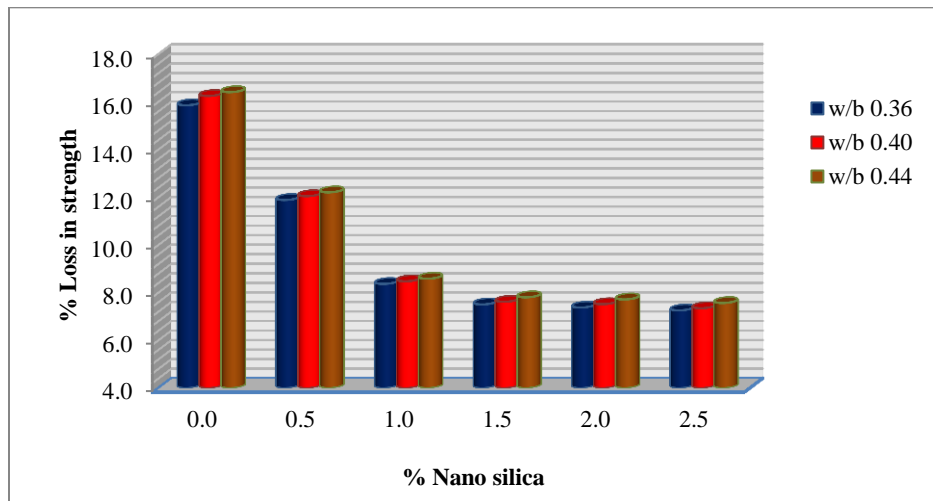
## 5.2 Sulphate resistance test result

Figure 6 presents a visual representation of the proportion of weight loss that occurs as a result of sulphate attack on cube specimens with varying percentage of NS. The variance in percentage weight loss demonstrates that the amount of weight lost decreases as the percentage of NS in the mixture rises, but the amount of weight lost increases for the same percentage of NS when the weight-to-volume ratio is increased. The findings indicate that the control mix suffers the biggest percentage of weight loss when compared to the other mixes, and that this percentage of weight loss reduces as the proportion of NS in the mix increases. Due of its high pozzolanic nature, NS associates with these calcium hydroxide crystals to generate CSH gel, which is why concrete containing NS exhibits good resistance to sulphate attack. The  $\text{Ca}(\text{OH})_2$  crystal shrinks in size

and quantity, and the C-S-H gel fills in the spaces to increase the density of the interfacial transition zone. In Figure 7, a graphic representation of the percentage loss in strength caused by sulphate attack on cube specimens with varied percentages of NS and w/b ratio is displayed. The graph indicates that the amount of strength loss experienced for a certain percentage of NS drops as the w/b ratio increases, but the amount of strength loss experienced for the same amount of NS surges when the w/b ratio increases. As can be seen from the results, the percentage strength loss is lowest for the control mix as compared to the other mixes, and it grows as the percentage of NS does. The better performance of NS concrete against sulphate attack may be attributed to the higher consumption of  $\text{Ca}(\text{OH})_2$  due to the additional pozzolanic reaction by the NS at early age.



**Figure 6** % Loss in weight with varying % of NS and w/b ratio due to sulphate attack



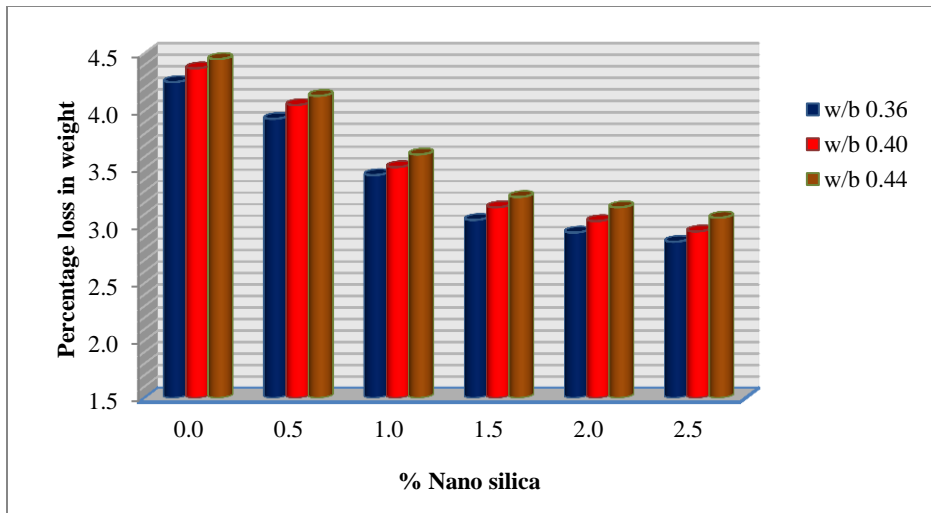
**Figure 7** % Loss in strength with varying % of NS and w/b ratio due to sulphate attack

### 5.3 Chloride resistance test result

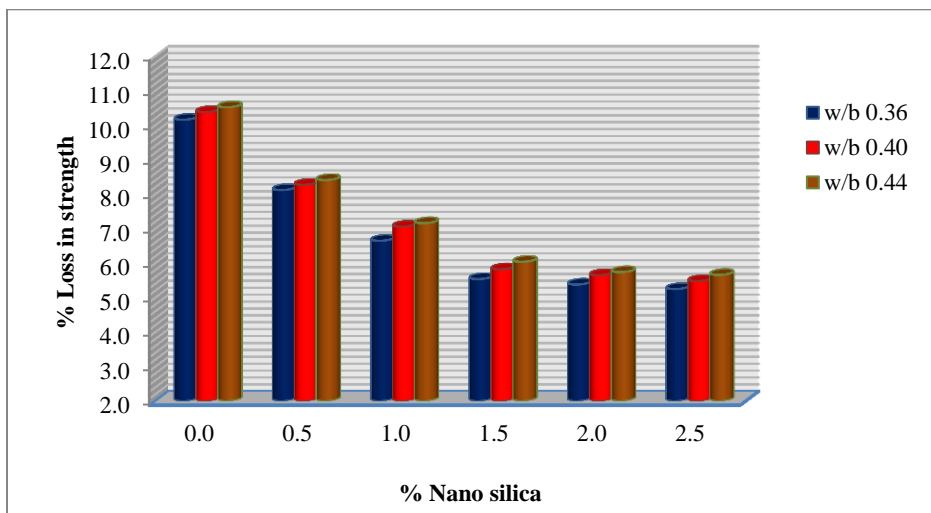
In *Figure 8*, a graphical representation of the proportion of weight loss that may be attributed to chloride attack on cube specimens with varied percentages of NS and w/b ratios has been shown. The fluctuation in the % weight loss reveals that as the quantity of NS increases, the weight loss lowers; conversely, for the same amount of NS, as the ratio of w/b rises, the percentage weight loss rises. This is shown by the fact that the weight loss decreases. The findings indicate that the control mix suffers the biggest percentage of weight loss when compared to the other mixes, and that this percentage of weight loss reduces as the proportion of NS in the mix increases. NS enhanced concrete demonstrates strong resistance to the corrosive effects of chloride, it can be attributed to the presence of reactive silica, which combines with  $\text{Ca(OH)}_2$  (a byproduct of cement

hydration) in finely divided form and resulting in leaching of  $\text{Ca(OH)}_2$  and converting them into CSH gel, due to which there are fewer bleed channels and permeability reduced. So, the entrance of anions is minimal. *Figure 9* illustrates the percentage strength loss caused by chloride attack on cube specimens with various amounts of NS and w/b ratio. The graph demonstrates that for a certain percentage of NS, strength loss decreases as the w/b ratio rises, while for the same amount of NS, strength loss increases as the w/b ratio rises. As can be seen from the results, the percentage strength loss is lowest for the control mix with respect to the other mixes, and it grows as the proportion of NS does. The better performance of NS concrete against chloride attack may be attributed to its pozzolanic character, due to which it combines with free lime and increasing structural strength over time.





**Figure 8** % Loss in weight with varying % of NS and w/b ratio due to chloride attack



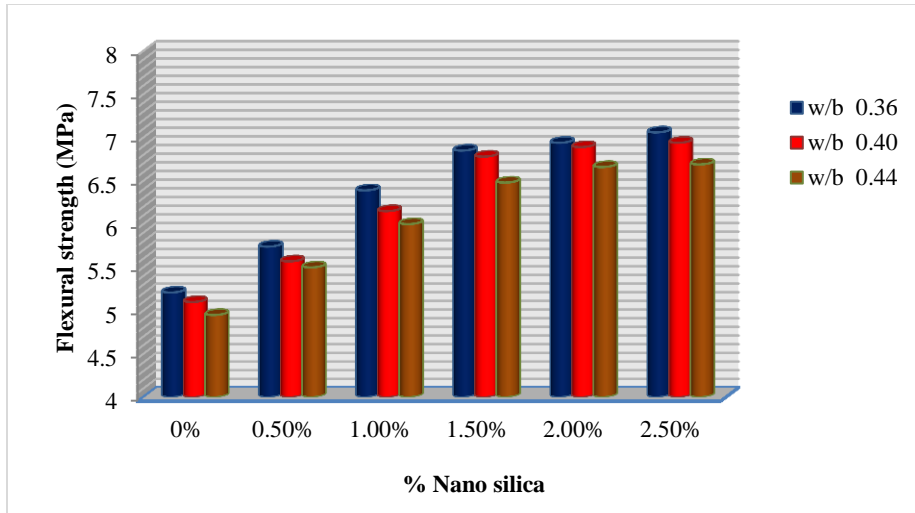
**Figure 9** % Loss in strength with varying % of NS and w/b ratio due to chloride attack

#### 5.4 Flexural strength test results

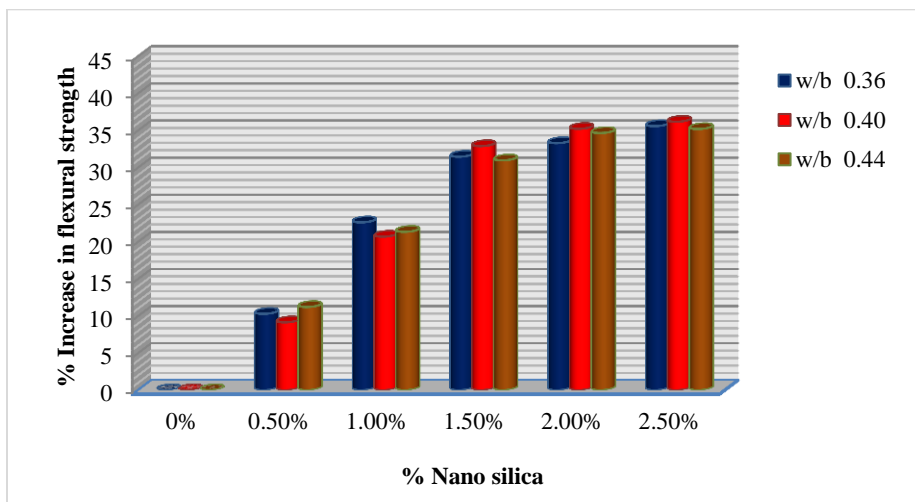
Figure 10 shows the variation in flexural strength with change in percentage NS content and w/b ratio. Figure 11 shows the percentage increase in flexural strength as compare to control mix for different w/b ratio with varying percentage of NS at 28 days. It is evident that all of the concrete specimens containing nanoparticles exhibited flexural strengths greater than those of the control specimens, which may be attributed to the pozzolanic reaction and filler effects of NS. The graph demonstrates that flexural strength values rose with increasing % of NS content. The main effect plots for the 28 days flexural strength have been shown in Figure 12.

#### 5.5 Split tensile strength test results

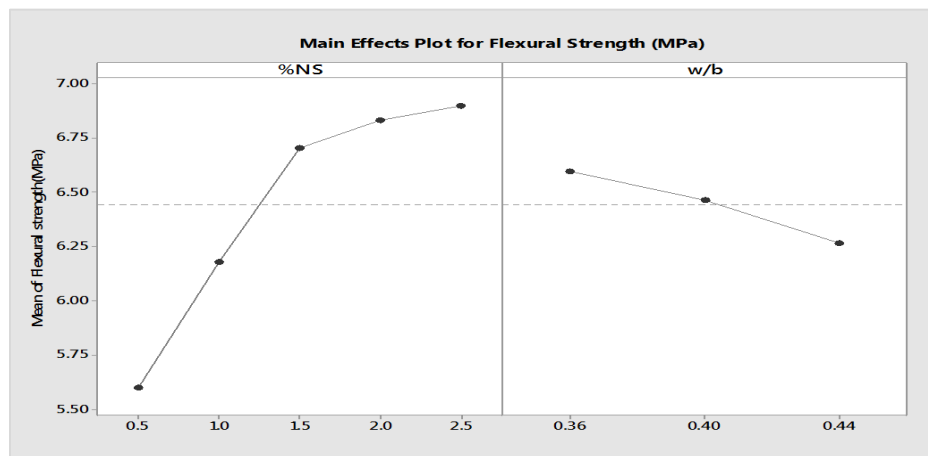
Figure 13 shows the variation in split tensile strength with change in percentage NS content and w/b ratio. Figure 14 shows the percentage increase in split tensile strength as compare to control mix for different w/b ratio with varying percentage of NS at 28 days. The graph shows that as percentage of NS increases, the split tensile strength also shows increasing trend. The main effect plots for the 28 days split tensile strength have been shown in Figure 15. The increased binding property of finely divided NS as a result of strong pozzolanic reaction and cement paste aggregate interfacial refinement resulting to better bond strength is the source of the higher split tensile strength for mixes containing NS.



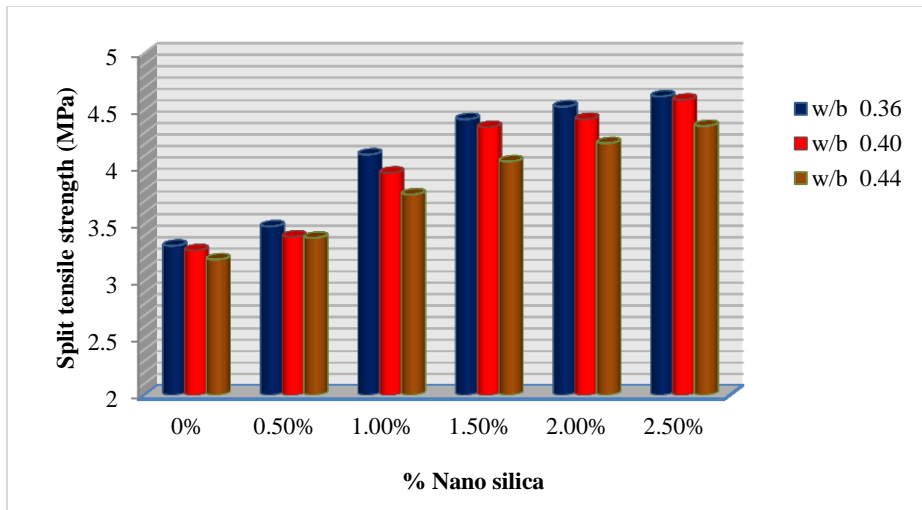
**Figure 10** Effect of % NS and w/b ratio on flexural strength



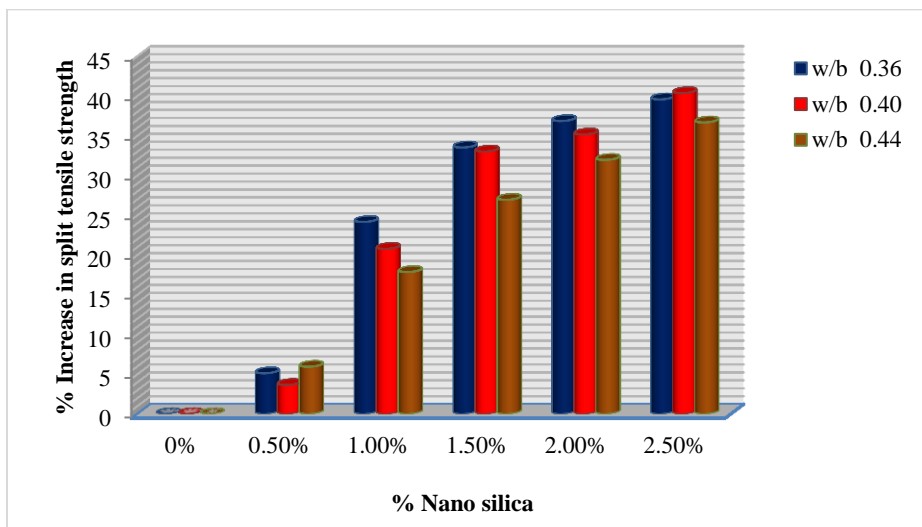
**Figure 11** Increase in flexural strength as compare to control mix



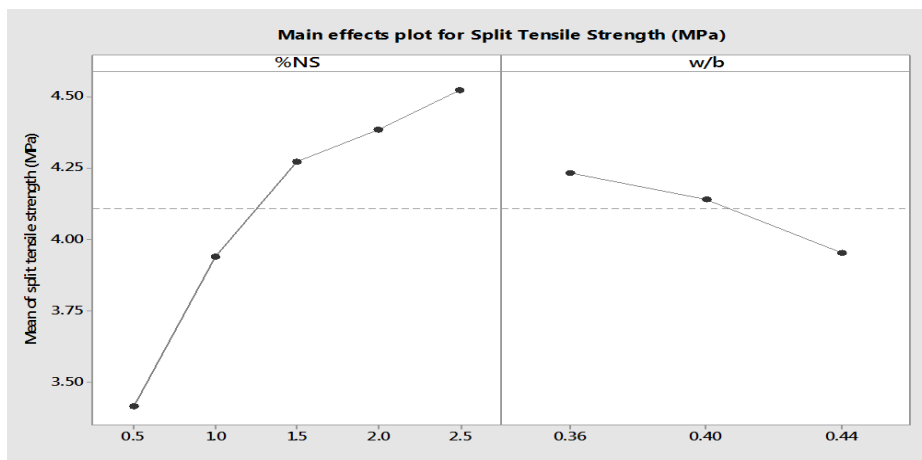
**Figure 12** Main effect plot for the 28 days flexural strength



**Figure 13** Effect of % NS and w/b ratio on split tensile strength



**Figure 14** Increase in split-tensile strength as compare to control mix



**Figure 15** Main effect plots for the 28 days split tensile strength

### 5.6 Results of the SEM test

Figure 16 shows the SEM image of Mix having Binary blending with NS at 7 days of curing. The SEM image shows condensed microstructure and a good dispersion of NS cluster throughout the entire surface of hydrated cement products. Due to high pozzolanic nature of NS, it reacts with portlandite crystals and start converting them into C-S-H gel at early age. NS particles have a higher specific surface area, which provides high chemical reactivity and these particles behaving as a nucleation centers, consequences in early hydration of cementitious

materials. These observations are well in accordance with the enhanced mechanical strength result obtained at 7 days. Figure 15 shows the SEM image of mix having binary blending with NS at 28 days of curing. It can be detected that the most of cluster of NS particle, which was found at 7 days curing period reacts with CSH crystals and convert them into C-S-H gel, due to which a dense and compacted microstructure has been observed at 28 days of curing. The improvement of microstructure was also justified by the improved mechanical strength and durability of concrete.

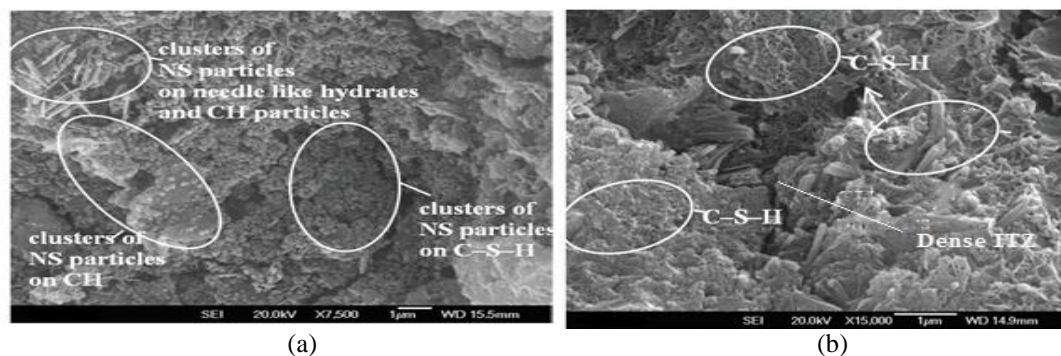


Figure 16 SEM image of Mix having binary blending with NS at (a) 7 days (b) 28 days

### 6. Discussion

Adding NS to concrete can alter its rheological properties. Replacement of NS can improve the mechanical and durability aspects of concrete. As per the investigation conducted, it can be summarised that the blended mix with NS showed significant improvement in terms of mechanical strength and durability criteria as well. Figure 4 and Figure 5, Figure 6 and Figure 7, Figure 8 and Figure 9 for acid test, sulphate test and chloride test respectively show that when the percentage of NS goes up, the percentage of strength loss goes down, but when the ratio of weight to body size goes up, the percentage of strength loss goes up a lot. It is possible that the fact that NS concrete is pozzolanic is what makes it so resistant to acid attack. Due to finer particle size and high specific area, NS shows very high pozzolanic activity, it not only works as a filler material, but also as an activator to pozzolanic reaction, which were evident from the SEM study of NS blended HSC. Figure 10 and Figure 11 shows the percentage increase in flexural strength as compare to control mix for different w/b ratio with varying percentage of NS at 28 days. Similarly, Figure 13 and Figure 14 shows the percentage increase in split tensile strength as compare to control

mix for different w/b ratio with varying percentage of NS at 28 days Split tensile strength test and flexural strength test also indicated improvement in the mechanical strength of the blended mix. Similarly, the results obtained from the acid resistance test, sulphate resistance test and chloride resistant test also shows the improved resistance of concrete toward the chemical attack. SEM image as shown in Figure 16 also indicates the improvement in the microstructure of the concrete after the inclusion of NS in the blended mix. The results of this study are based upon the six tests that have been undertaken. However, due to unavailability of the instruments tests like rapid chloride penetration test (RCPT), gas permeability test and chloride diffusion test including Nord-test method could not be performed and may be considered for further research.

A complete list of abbreviations is shown in Appendix I.

### 7. Conclusion and future work

As per the experimental findings, it can be stated that the percentage reduction in compressive strength as well as the weight of the cube specimen after chemical attack increases as the percentage of NS in

the concrete mix decreases. However, the rate of reduction in percentage weight and strength decreases after 1.5% NS inclusion. Therefore, the optimal dose for the best resistance to chemical attack is 1.5% NS replacement level. Additionally, it may be inferred that when the w/b ratio decreases, so does the percentage of strength and weight loss. However, the concrete becomes unworkable when the w/b ratio is reduced by more than 0.36. NS inclusion in concrete shows tremendous improvement in strength and durability properties of concrete, but due to its higher cost its use is restricted to high strength concrete (HSC) and important concrete structure only. Additionally, significant enhancement in split tensile strength and flexural strength of concrete were observed with increasing percentage of NS, but the rate of percentage increase in strength was maximum for 1.5% NS. Furthermore, the experimental setup in this study focused on the effect of NS, however, other durability indicators have still not been investigated which becomes the limitation of this study. Investigating the effect of using quaternary blending with ground granulated blast furnace slag (GGBS), alccofines and NS on the strength and durability of HSC will help in providing a clear understanding of the possible mix proportions.

### Acknowledgment

None.

### Conflicts of interest

The authors have no conflicts of interest to declare.

### Author's contribution statement

**Dr. Satish Kumar Chaudhary:** Writing, experimental setup and data collection. **Dr. Ajay Kumar Sinha:** Conceptualising the research and supervision. **Dr. Praveen Anand:** Writing draft and data collection.

### References

- [1] Chaudhary SK, Sinha AK. Effect of silica fume on permeability and microstructure of high strength concrete. *Civil Engineering Journal*. 2020; 6(9):1697-703.
- [2] Mangi SA, Ibrahim MH, Jamaluddin N, Arshad MF, Jaya RP. Short-term effects of sulphate and chloride on the concrete containing coal bottom ash as supplementary cementitious material. *Engineering Science and Technology, an International Journal*. 2019; 22(2):515-22.
- [3] Wei Y, Chai J, Qin Y, Li Y, Xu Z, Li Y, et al. Effect of fly ash on mechanical properties and microstructure of cellulose fiber-reinforced concrete under sulfate dry-wet cycle attack. *Construction and Building Materials*. 2021; 302:124207.

- [4] Yu J, Qiao H, Zhu F, Wang X. Research on damage and deterioration of fiber concrete under acid rain environment based on GM (1, 1)-Markov. *Materials*. 2021; 14(21):1-16.
- [5] Nadir HM, Ahmed A. The mechanisms of sulphate attack in concrete—a review. *Modern Approaches on Material Science*. 2022; 5(2):658-70.
- [6] Khan MI, Sayyed MA, Ali MM. Examination of cement concrete containing micro silica and sugarcane bagasse ash subjected to sulphate and chloride attack. *Materials Today: Proceedings*. 2021; 39:558-62.
- [7] Metalssi OO, Touhami RR, Barberon F, De LJB, Roussel N, Divet L, et al. Understanding the degradation mechanisms of cement-based systems in combined chloride-sulfate attack. *Cement and Concrete Research*. 2023; 164:107065.
- [8] Sakr MR, Bassuoni MT. Effect of nano-based coatings on concrete under aggravated exposures. *Journal of Materials in Civil Engineering*. 2020; 32(10):04020284.
- [9] Idrees M, Akbar A, Ashraf S. Potential of pyrogenic nano silica to enhance the service life of concrete. *Journal of Materials in Civil Engineering*. 2023; 35(5):04023051.
- [10] Manthana SL, Boddepalli KR. Effect of tile aggregate and flyash on durability and mechanical properties of self-compacting concrete. *Journal of Building Pathology and Rehabilitation*. 2022; 7(1):68.
- [11] Tayeh BA, Alyousef R, Alabduljabbar H, Alaskar A. Recycling of rice husk waste for a sustainable concrete: a critical review. *Journal of Cleaner Production*. 2021; 312:127734.
- [12] Xu X, Jin Z, Yu Y, Li N. Impact properties of ultra high performance concrete (UHPC) cured by steam curing and standard curing. *Case Studies in Construction Materials*. 2022; 17(2022):1-17.
- [13] Costa VC, De SJFG, Thomas S, Toledo FRD, De CSL, Thode FS, et al. Nanotechnology in concrete: a bibliometric review. *Brazilian Journal of Experimental Design, Data Analysis and Inferential Statistics*. 2021; 1(1):100-13.
- [14] Ghafoori N, Batilov I, Najimi M. Resistance to sulfate attack of mortars containing colloidal nano silica and silica fume. *Journal of Materials in Civil Engineering*. 2020; 32(12):06020019.
- [15] Tuan NA, Nga NT, Khai LT, Van TN, Vuong VQ. Combination of additives to characteristics of concrete in marine works. *Magazine of Civil Engineering*. 2022; 112(4):11204.
- [16] Zhang A, Ge Y, Du S, Wang G, Chen X, Liu X, et al. Durability effect of nano-SiO<sub>2</sub>/Al<sub>2</sub>O<sub>3</sub> on cement mortar subjected to sulfate attack under different environments. *Journal of Building Engineering*. 2023; 64:105642.
- [17] Hamada HM, Alattar AA, Yahaya FM, Muthusamy K, Tayeh BA. Mechanical properties of semi-lightweight concrete containing nano-palm oil clinker powder. *Physics and Chemistry of the Earth, Parts a/b/c*. 2021; 121:102977.

- [18] Hamada HM, Alya'a A, Yahaya FM, Muthusamy K, Tayeh BA, Humada AM. Effect of high-volume ultrafine palm oil fuel ash on the engineering and transport properties of concrete. *Case Studies in Construction Materials*. 2020; 12(2020):1-12.
- [19] Mostafa SA, Ahmed N, Almeshal I, Tayeh BA, Elgamal MS. Experimental study and theoretical prediction of mechanical properties of ultra-high-performance concrete incorporated with nanorice husk ash burning at different temperature treatments. *Environmental Science and Pollution Research*. 2022; 29(50):75380-401.
- [20] Ren J, Zhang L, Walkley B, Black JR, San NR. Degradation resistance of different cementitious materials to phosphoric acid attack at early stage. *Cement and Concrete Research*. 2022; 151:106606.
- [21] Albattat RA, Jamshidzadeh Z, Alasadi AK. Assessment of compressive strength and durability of silica fume-based concrete in acidic environment. *Innovative Infrastructure Solutions*. 2020; 5:1-7.
- [22] De BN, Debruyckere M, Van ND, De BB. Concrete attack by feed acids: accelerated tests to compare different concrete compositions and technologies. *Materials Journal*. 1997; 94(6):546-54.
- [23] Erbektas AR, Isgor OB, Weiss WJ. Comparison of chemical and biogenic acid attack on concrete. *ACI Materials Journal*. 2020; 117(1):255-64.
- [24] Chang ZT, Song XJ, Munn R, Marosszeky M. Using limestone aggregates and different cements for enhancing resistance of concrete to sulphuric acid attack. *Cement and Concrete Research*. 2005; 35(8):1486-94.
- [25] Zhou Y, Tian H, Sui L, Xing F, Han N. Strength deterioration of concrete in sulfate environment: an experimental study and theoretical modeling. *Advances in Materials Science and Engineering*. 2015; 2015:1-14.
- [26] Collepardi M. Ettringite formation and sulfate attack on concrete. Fifth CANMET/ACI international conference on recent advances in concrete technology. 2001; Sp-200:21-38.
- [27] Ait-mokhtar K, Millet O, editors. *Structure design and degradation mechanisms in coastal environments*. ISTE; 2015.
- [28] Abhilash PP, Nayak DK, Sangoju B, Kumar R, Kumar V. Effect of nano-silica in concrete; a review. *Construction and Building Materials*. 2021; 278:122347.
- [29] Wang H, Chen Z, Li H, Sun X. Numerical simulation of external sulphate attack in concrete considering coupled chemo-diffusion-mechanical effect. *Construction and Building Materials*. 2021; 292:123325.
- [30] Yin GJ, Shan ZQ, Miao L, Tang YJ, Zuo XB, Wen XD. Finite element analysis on the diffusion-reaction-damage behavior in concrete subjected to sodium sulfate attack. *Engineering Failure Analysis*. 2022; 137:106278.
- [31] Sun D, Cao Z, Huang C, Wu K, De SG, Zhang L. Degradation of concrete in marine environment under coupled chloride and sulfate attack: a numerical and experimental study. *Case Studies in Construction Materials*. 2022; 17:e01218.
- [32] Maes M, De BN. Resistance of concrete and mortar against combined attack of chloride and sodium sulphate. *Cement and Concrete Composites*. 2014; 53:59-72.
- [33] Shetty MS, Jain AK. *Concrete technology (Theory and Practice)*, 8e. S. Chand Publishing; 2019.
- [34] Ryan M. Sulphate attack and chloride ion penetration, their role in concrete durability. QCL group technical note, Australia. Search In. 1999.
- [35] Shannag MJ, Shaia HA. Sulfate resistance of high-performance concrete. *Cement and Concrete Composites*. 2003; 25(3):363-9.
- [36] Tabish M, Zaheer MM, Baqi A. Effect of nano-silica on mechanical, microstructural and durability properties of cement-based materials: a review. *Journal of Building Engineering*. 2022; 105676.
- [37] Liu X, Ma B, Tan H, Zhang T, Mei J, Qi H, et al. Effects of colloidal nano-SiO<sub>2</sub> on the immobilization of chloride ions in cement-fly ash system. *Cement and Concrete Composites*. 2020; 110:103596.
- [38] Zhang S, Niu D, Luo D. Enhanced hydration and mechanical properties of cement-based materials with steel slag modified by water glass. *Journal of Materials Research and Technology*. 2022; 21:1830-42.
- [39] Singh A, Mehta PK, Kumar R. Recycled coarse aggregate and silica fume used in sustainable self-compacting concrete. *International Journal of Advanced Technology and Engineering Exploration*. 2022; 9(96):1581-96.
- [40] Wang C, Zhang M, Wang Q, Zhang R, Pei W, Zhou Y. Influence of nano-silica on the performances of concrete under the negative-temperature curing condition. *Cold Regions Science and Technology*. 2021; 191:103357.
- [41] Oh T, Chun B, Lee SK, Lee W, Banthia N, Yoo DY. Substitutive effect of nano-SiO<sub>2</sub> for silica fume in ultra-high-performance concrete on fiber pull-out behavior. *Journal of Materials Research and Technology*. 2022; 20:1993-2007.
- [42] Yunchao T, Zheng C, Wanhui F, Yumei N, Cong L, Jieming C. Combined effects of nano-silica and silica fume on the mechanical behavior of recycled aggregate concrete. *Nanotechnology Reviews*. 2021; 10(1):819-38.
- [43] IS 8112:1989, Ed., Indian Standard 43 – Grade Ordinary Portland cement Specifications. Bureau of Indian Standards.
- [44] IS 383, "Specification for Coarse and fine aggregates from natural sources for concrete," Bur. Indian Stand., 2016.
- [45] IS 10262:2009, Ed., Indian standards recommended Guidelines for concrete mix design, 2009th ed. Bureau of Indian Standards.
- [46] IS 456. Code of practice for plain and reinforced concrete. Bureau of Indian Standards. 2000.





**Dr. Satish Kumar Chaudhary** currently serves as an Assistant Engineer at RCD, Bihar. With a combination of 5 years of field and research experience, he acquired his B.Tech degree from NCE in 2012 and went on to pursue an M.Tech in Structural Engineering from NIT Patna in 2016. He successfully completed his PhD from NIT Patna as well. His primary research interests revolve around concrete, Micro silica, and Nano silica. He holds memberships in esteemed organizations such as the All India Council for Technical Skill Development (AICTSD) and the International Association of Engineers (IAENG). Notably, he possesses extensive expertise in quality monitoring of Road Design and MATERIALs. Email: satish.ce16@nitp.ac.in



**Dr. Ajay Kumar Sinha** currently holds the position of Professor in the Civil Engineering Department at the National Institute of Technology Patna. With an impressive 37 years of teaching and research experience, he earned his B.Tech degree from IIT BHU in 1986. Furthering his academic journey, he completed his M.E. in Earthquake Engineering from IIT Roorkee in 1989, and later accomplished his PhD from Delhi College of Engineering, University of Delhi. Dr. Sinha's research interests encompass a wide range of areas including Seismic Resistant Structures, Vulnerability Assessment and Retrofitting of Structures, Structural Health Monitoring, and Reliability Engineering. Notably, he serves as the Center Director cum Nodal Officer of the Earthquake Safety Clinic and Center at NIT Patna. He is also a distinguished member of the Earthquake Committee of BSDMA, GoB, Patna. With an impressive academic portfolio, Dr. Sinha has contributed significantly to the field of civil engineering. He has published over 155 research papers in both national and international journals and conferences. Additionally, his mentorship extends to supervising 7 PhD and 60 ME students, with an additional 10 PhD candidates currently under his guidance. Email: aksinha@nitp.ac.in



**Dr. Praveen Anand** earned his B.Tech degree in Civil Engineering in 2015 from West Bengal University of Technology, India. He pursued his M.Tech with a specialization in structural engineering from the National Institute of Technology, Patna. In 2022, he successfully completed his Ph.D. at the National Institute of Technology, Patna. Dr. Anand has maintained an affiliation with the Earthquake Safety and Clinic Centre (EQSC) at NIT Patna. He also served as a Guest Assistant Professor at Government Engineering College, Jehanabad. His primary area of interest revolves around Retrofitting and Strengthening, as well as Vulnerability Assessment and Seismic Analysis of Structures. Email: praveen.ce16@nitp.ac.in

### Appendix I

S. No.	Abbreviation	Description
1	CA	Coarse Aggregate
2	CNS	Colloidal Nano Silica
3	CSH	Calcium Silicate Hydrate
4	FA	Fine Aggregate
5	FM	Fineness Modulus
6	G	Specific Gravity
7	GGBS	Ground Granulated Blast Furnace Slag
8	HRWR	High-Rate-Water-Reducer
9	HSC	High Strength Concrete
10	IS	Indian Standards
11	NS	Nano Silica
12	OPC	Ordinary Portland Cement
13	RCPT	Rapid Chloride Penetration Test
14	SCM	Supplemental Cementitious Materials
15	SEM	Scanning Electron Microscope
16	w/b	Water-Binder



Article

# Synthesis and Wear Behaviour Analysis of SiC- and Rice Husk Ash-Based Aluminium Metal Matrix Composites

Sameen Mustafa <sup>1,\*</sup>, Julfikar Haider <sup>2,\*</sup>, Paolo Matteis <sup>3</sup> and Qasim Murtaza <sup>4</sup>

<sup>1</sup> Faculty of Engineering, Free University of Bozen-Bolzano, 39100 Bolzano, Italy

<sup>2</sup> Department of Engineering, Manchester Metropolitan University, Manchester M1 5GD, UK

<sup>3</sup> Department of Applied Science and Technology, Politecnico di Torino, 10129 Turin, Italy; paolo.matteis@polito.it

<sup>4</sup> Department of Mechanical Engineering, Delhi Technological University, Delhi 110042, India; qasimmurtaza@gmail.com

\* Correspondence: smustafa@unibz.it (S.M.); j.haider@mmu.ac.uk (J.H.)

**Abstract:** Research efforts seek to develop aluminium alloy composites to enhance the poor tribological performance of aluminium alloy base matrix. In this research, a hybrid metal matrix composite (HMMC) was developed by reinforcing an aluminium alloy (AA8011) with SiC and rice husk ash (RHA) using a stir casting technique. RHA was prepared by the cracking of rice husk, which is abundantly available in the Indian subcontinent. The samples were cast by keeping the amount of RHA constant at 2.5 wt.% and varying the amount of SiC from 0.0 wt.% to 8 wt.%. The samples were machined to manufacture pins for wear tests (at ambient temperature, 100 °C, and 200 °C) and hardness measurement. The microstructures of the cast samples were analysed using an X-ray diffractometer (XRD) and a scanning electron microscope (SEM), along with energy-dispersive X-ray spectroscopy (EDS). It was observed that the composites with greater reinforcement of SiC exhibited improved hardness and wear resistance, but the coefficient of friction increased with the addition of RHA and SiC, and the wear performance deteriorated with an increase in the operating temperature. The contribution of RHA alone to the improvement in wear performance was marginal compared to the pure alloy. It was also confirmed that the reinforced composites could be a better option for automotive applications to replace aluminium alloys.

**Keywords:** metal matrix composite (MMC); SiC; rice husk ash (RHA); stir casting; wear



**Citation:** Mustafa, S.; Haider, J.; Matteis, P.; Murtaza, Q. Synthesis and Wear Behaviour Analysis of SiC- and Rice Husk Ash-Based Aluminium Metal Matrix Composites. *J. Compos. Sci.* **2023**, *7*, 394. <https://doi.org/10.3390/jcs7090394>

Academic Editor: Francesco Tornabene

Received: 17 August 2023

Revised: 6 September 2023

Accepted: 13 September 2023

Published: 15 September 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Composite materials amount to approximately 12.5% of the overall industry of engineering materials [1]. Composites have replaced bronze and cast-iron, as they exhibit low density and higher mechanical strength compared to these materials. Despite these enhanced material properties, composites are extensively being researched to enhance their low wear and seizure resistance [2]. Over the past decades, numerous studies have explored and reported on the wear behaviour of composites. In this study, the focus is on investigating the wear behaviour of aluminium metal matrix composites (AMMC) reinforced with SiC and rice husk ash (RHA).

Aluminium and aluminium alloys are finding increasing applications in various technical fields. They are employed in aviation, aerospace, military, automotive, and electronic industries. To improve the tribological and mechanical characteristics of aluminium alloys, different reinforcements are added, and corresponding composites are formed [3,4]. AMMCs reinforced with materials such as SiC, B<sub>4</sub>C, and other ceramics are increasingly being utilised in the automotive, aerospace, underwater, and transportation industries. The addition of reinforcements such as ceramics and ash improve tribological and mechanical properties like strength, stiffness, impact, wear, and abrasion resistance. In recent years, the addition of fillers (graphite, fly ash, material fibres from wheat husk ash, and jute

ash) to AMMCs is being studied to obtain good toughness levels [5]. If the base metal alloy is mixed with reinforcing particles such as SiC or B<sub>4</sub>C and fibre particles such as ash, the resulting product is called a hybrid metal matrix composite (HMMC) [6]. For applications in the automotive and aerospace industries, lightweight and high-performance materials are required. The lightweight material is produced by developing composites with hard reinforcement particulates in the matrix of soft material [7]. Singh et al. reported that, for application in the automobile, aerospace, and aircraft industries, lightweight and high-performance materials are required [7].

The addition of SiC can increase both the mechanical strength and the wear performance of AMMCs. Additionally, it has been reported that hybrid composites exhibit superior wear properties of worn surfaces as compared to pure Al alloys [8]. SiC particles possess excellent compatibility with aluminium matrices, and they can be obtained at a low cost. A major application of AMMCs involves moving and sliding parts. Therefore, the investigation of the tribological properties of these materials is of significant importance for us to determine the behaviour of composite materials during service application [9]. The applied load and sliding distance during the wear test affects the coefficient of friction (COF) to a greater extent, and an increased silicon carbide content offers better wear resistance in the HMMC [10]. It has been found that the wear resistance of the HMMC is higher than that of the corresponding base alloy. The addition of hard reinforcements, such as B<sub>4</sub>C and fly ash, shows a greater improvement in wear performance [11,12].

AA8011 is a type of AMMC that exhibits higher ductility and malleability than most of the other AMMCs [13]. It contains a rare mix of desirable properties such as low weight, low maintenance, and good corrosion resistance. Studies are conducted to enhance the strength of AA8011 by reinforcing it with ceramic particles to make it more durable and increase the scope of its application [14]. Furthermore, it has been observed that the wear decreases when the amount of reinforcement element is increased [15,16]. In addition, it has been determined that the COF increases with the sliding velocity [17].

The addition of ceramics such as titanium diboride can significantly improve the wear behaviour of AMMCs due to the formation of a mechanically mixed layer of Fe<sub>2</sub>O<sub>3</sub>. [17]. Karthikumar et al. prepared Aluminium 8011 MMC by stir casting using TiB<sub>2</sub> as a reinforcement. The study concluded that the maximum hardness of 55.03% was exhibited in the case of 8% by weight of TiB<sub>2</sub> reinforcement. Also, maximum % elongation was found in the case of 4% by weight of TiB<sub>2</sub> [18].

Studies have reported that the reinforcement of AMMCs with Al<sub>2</sub>O<sub>3</sub> increases the wear resistance of the HMMC as compared to the corresponding matrix material, AL-6061 [19]. When the operating temperature increases, the wear rate decreases constantly with the increase in sliding velocity. Further, the increase in the reinforcement volume fraction increases the wear resistance of the composites [20,21].

According to a study, for every ton of paddy processed, an average mill produces 200 kg of rice husk and 40 kg of RHA [22]. Thus, the total amount of rice husk produced in India is estimated to be about 24 million tons, while the RHA production is estimated to be around 4.8 million tons per year. Since rice is a diet staple in India, the choice of a strengthening particle such as RHA provides a possible sustainable option for countries like India. This could help with utilising the waste obtained from rice harvesting.

Although research is being carried out to study the enhancement of the properties of AA8011 MMC, the aspect of characterising the physical wear of AA8011 at different temperatures and its surface morphology has not often been studied yet in terms of the wear on the AMMC. The exploration of modifying Al alloys with two different kinds of fillers, i.e., typical ceramic (SiC) particles and plant-based waste (RHA), might be of engineering significance. This study explores the wear behaviour of an HMMC (AA8011 reinforced with SiC and RHA) through experimental testing. The HMMC samples were prepared using the stir casting technique, which is the most suited for producing reinforced AMMC.

The principal contribution of this research lies in its pioneering utilisation of RHA as a reinforcement material, which not only demonstrates its viability for enhancing the tribological properties of aluminium alloy composites, but also underscores its potential to significantly reduce the waste generated from rice harvesting. In this context, the most efficient utilisation of natural resources can be pursued [23]. Additionally, the development of a hybrid composite, blending RHA and SiC with an aluminium alloy base matrix, allows for the prospect of achieving a well-balanced amalgamation of mechanical properties. This approach, involving systematic variation in SiC content, facilitates a meticulous examination of its impact on the composite's characteristics. This could substantially contribute to mitigating the ecological footprint by valorising an abundant agricultural by-product. Consequently, the dissemination of this research through publication stands to advance the body of knowledge in the field of materials science and engineering, with far-reaching implications for the development of sustainable, high-performance composite materials and reducing agricultural waste stemming from rice harvesting practices.

## 2. Materials and Methods

### 2.1. Preparation of Cast MMC Samples

To prepare the HMMC, first, the aluminium alloy was melted in a stir casting furnace and the blended powder mix (SiC and RHA) was fed into the melt. The stir casting process was selected to produce metal matrix composites due to its inherent advantages, despite its drawbacks relative to alternative casting techniques. This choice was primarily driven by factors such as cost-effectiveness, material flexibility, and the ability to achieve homogeneous mixing. While it exhibits limitations such as potential porosity and limited control over reinforcement distribution, these drawbacks can often be managed effectively, making stir casting a pragmatic choice in scenarios where cost efficiency, rapid prototyping, and ease of implementation are paramount.

After thorough mixing, the melt was poured into a cylindrical cast-iron die to prepare samples. Freshly obtained rice husk contains a large amount of moisture and oil; thus, it does not burn, but chars into black, brittle material when torched. Therefore, to prepare a fine RHA powder, the rice husk was charred overnight using the indigenous method. The burnt husk was then milled to a fine powder using a ball mill. SEM and EDS of the milled ash were then carried out to identify the powder's morphology and chemical composition.

Commercially available SiC powder was used in the experiment. To obtain a homogeneous mix of RHA and SiC, the milled ash and SiC powder were blended in the ball mill for at least 30 min. Before blending, SiC powder was carefully measured using a precision balance according to the different weight percentages required for the samples. The measurement of RHA was kept constant at 2.5% by weight for each cast sample. The samples' nomenclature and their respective compositions are given in Table 1.

**Table 1.** Nomenclature and composition of samples.

Sample Group Number	Sample ID	wt.% of AA-8011	wt.% of SiC	wt.% of RHA
S1	8SiC/2.5RHA	89.5	8	2.5
S2	6SiC/2.5RHA	91.5	6	2.5
S3	4SiC/2.5RHA	93.5	4	2.5
S4	2SiC/2.5RHA	95.5	2	2.5
S5	0SiC/2.5RHA	97.5	0	2.5
S6	0SiC/0RHA	100	0	0

The work-hardened AA8011 alloy was purchased in the form of sheets with a composition of 98.7% Al, 0.7% Fe, and 0.6% Si. The sheets were cut into smaller plates, each weighing around 20 g. Looking at the size of the die, the aluminium alloy required to fill

the die was calculated as approximately 380 g. The precise weight of the alloy was placed into the crucible by carefully choosing the correct number of smaller plates. The alloy was subsequently melted in a furnace, maintaining a constant temperature of 900 °C for a minimum of 20 min. In parallel, the RHA/SiC mixture was preheated in a muffle furnace at 350 °C. This step served two purposes: ensuring the complete dryness of the mixture and minimizing the temperature disparity between the mix and the molten metal.

The heated powder mix was then fed into the stir casting furnace, and stirring was carried out using a mechanical stirrer for at least 20 min. Since the experimentation was carried out during the winter, with ambient temperatures being lower than 20 °C, the die had to be pre-heated to minimise the directional solidification and heat transfer during the casting process. The final sample was obtained as a solid cylinder. A total of 18 samples were prepared, with 3 samples for each material group.

The cast samples were cut into smaller parts using a hacksaw. The cut pieces were then precision-turned on a universal lathe in order to reach a final diameter value of  $6 \pm 0.05$  mm, and the lengths of the pins were kept at  $50 \pm 2$  mm. It is to be noted that the pins needed to be prepared through conventional machining. This was due to the fact that blow holes are very common in gravity-cast samples, and that the in-fusion of non-conducting particles of ceramic and ash make it almost impossible to cut the samples using a wire-cut EDM process.

## 2.2. Characterisation of Samples

The castings were cut into small discs to measure their hardnesses by a Rockwell C Hardness (RHN) tester. For each sample, hardness was measured at 5 different points, and the average for each sample was recorded. A load of 100 kgf and an indentation ball diameter of approximately 1.6 mm were chosen for the hardness measurements. For SEM and EDS analysis (EDS, Jeol, JSM-6510LV, Tokyo, Japan), the cast samples were further cut into small, plate-like pieces, and one surface of each sample was ground to a fine finish. The structural analysis of the samples was performed on an XRD machine (Shimadzu LabX, Kyoto, Japan), which had a monochromatic Cu source of  $K\alpha$  wavelength of 1.540 Å. The measurement conditions for XRD were as follows: voltage = 40.0 (kV), current = 30.0 (mA), scan range = 10.00–80.00 (deg), scan speed = 6.00 (deg/min), and sampling pitch = 0.02 (deg). The scanning time for each sample was around 11 min. The diffraction patterns were plotted using OriginPro software version 9.7.

## 2.3. Wear Test Procedure

The dry friction wear test was conducted using a pin-on-disc tribometer able to conduct tests at elevated temperatures, either by heating the pin or the disc. The selection of a pin-on-disc tribometer for the wear analysis of aluminium composites was underpinned by its capacity to closely emulate real-world wear conditions, affording a comprehensive understanding of material performance. This apparatus offers meticulous control over experimental parameters, encompassing load, sliding speed, and environmental factors, thus ensuring the reproducibility of wear tests and facilitating the systematic evaluation of diverse composite formulations [24]. Moreover, pin-on-disc tribometers furnish quantitative data on wear rates and coefficients of friction, crucial for the rigorous assessment of aluminium composites and their suitability for specific industrial contexts [25,26]. The examination of worn surfaces and wear debris enables the discernment of wear mechanisms, thereby guiding material enhancement endeavours. Furthermore, the adaptability of these tribometers to varying materials and conditions, coupled with their cost-effectiveness and expeditious testing capabilities, renders them a pragmatic choice for elucidating the tribological behaviour of aluminium composites and advancing their performance characteristics in numerous engineering applications [8]. In this case, the pins were heated at different temperatures and the disc temperature was kept constant. The pin was inserted into the holder, which was heated by an electrical coil, and the temperature of the pin was measured by a thermocouple attached to it. The friction pair was chosen to be the

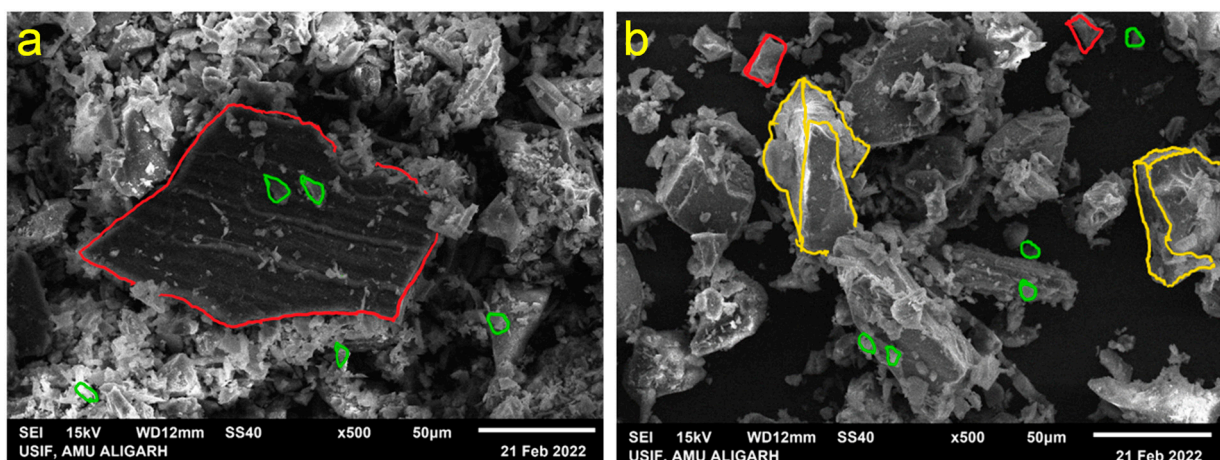


AA-8011 HMMC pin and high carbon EN-31 (100Cr6) steel disc. The disc material had a much greater hardness level than the HMMC. This type of coupling was chosen intentionally to observe the wear behaviour of the casted samples only. The wear test parameters were kept constant during each test, with a sliding speed of 300 rpm, a track diameter of 80 mm, and an applied load of 30 N. Each set of pins was tested at varying temperatures of 22 °C (ambient), 100 °C, and 200 °C. The choice of this experimental set-up bears significant relevance to several industrial domains, including the automotive, aerospace, and manufacturing sectors. Within the automotive industry, it serves as a valuable tool for assessing the wear and friction characteristics of composite materials like Al/SiC/Gr under conditions that simulate the temperature fluctuations encountered by engine components during operation [12,27]. Such insights are instrumental in the development of lightweight and high-performance materials for applications such as pistons and cylinder liners. Similarly, in the aerospace sector, this tribomechanical system aids in evaluating the performances of composite materials in critical components subjected to diverse temperature conditions, ensuring the safety and reliability of the aerospace systems [3,19]. These references underscore the significance of studying the tribological behaviour of the composite materials under fluctuating temperature conditions across different industrial applications. The disc was not heated, and its temperature was kept constant at  $22 \pm 2$  °C. A total of 18 samples (3 samples from each group) were tested on the tribometer. Each composition was tested at three different temperatures. The results for frictional force (N) and sliding distance (m) were recorded automatically by the tribometer's software. The pins were also weighed (using a precision balance with an accuracy of 0.0001 g) before and after performing the wear tests, and the weight differences were recorded in order to calculate the volumetric wear.

### 3. Results and Discussion

#### 3.1. Microstructure and Composition

From the SEM observation, it could be inferred that the RHA was a homogenous mixture of bimodal particle size, as shown in Figure 1a. The bigger particles are marked with red, and the smaller ones with green. Particle size was measured in accordance with the scale shown in SEM micrographs, and then mean value was computed. Smaller ash particles adhered to bigger particles with homogenous blending. The smaller particles varied in size from 7 µm to 15.6 µm, while the larger particle sizes were in the range of 46.8 µm to 156.2 µm.



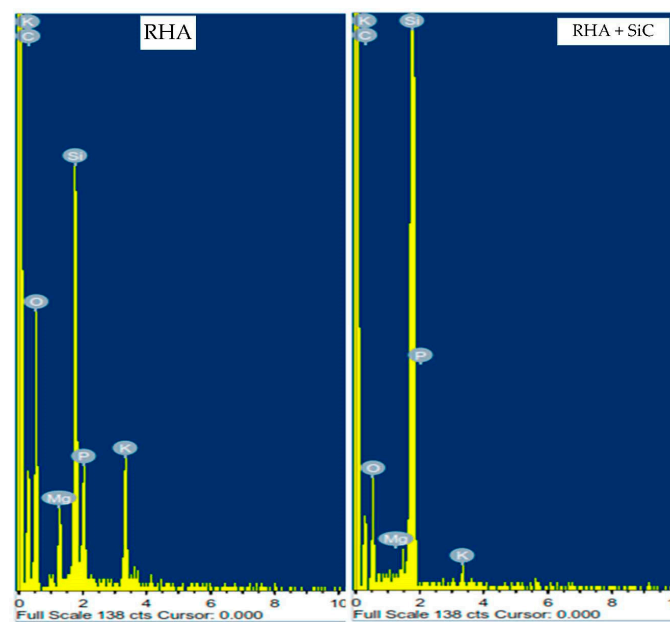
**Figure 1.** SEM images of (a) ash and (b) the ash and SiC powder mix at  $\times 500$ .

An SEM micrograph of the powder mix (RHA+SiC) is shown in Figure 1b. It was clearly seen that the wedge-like particles (SiC) were uniformly dispersed, along with other irregularly shaped particles, as desirable. The abrasive particles can be seen marked in



yellow, while the larger- and smaller-sized ash particles are marked with red and green, respectively. The SiC particles were clearly distinguishable from the RHA particles based on size and shape.

The EDS spectra of the powders, as well as the metal samples, were analysed. The specimens were tested for all the possible elements, and no peaks were omitted. Figure 2 shows the spectrum and composition chart of the RHA. Similarly, the spectra for the other samples were also obtained. The results obtained for the powdered RHA and powder mix are summarised in Table 2. Higher percentages of silicon and carbon in the powder mix compared to the RHA indicated the presence of SiC. This supports the findings in the SEM images.



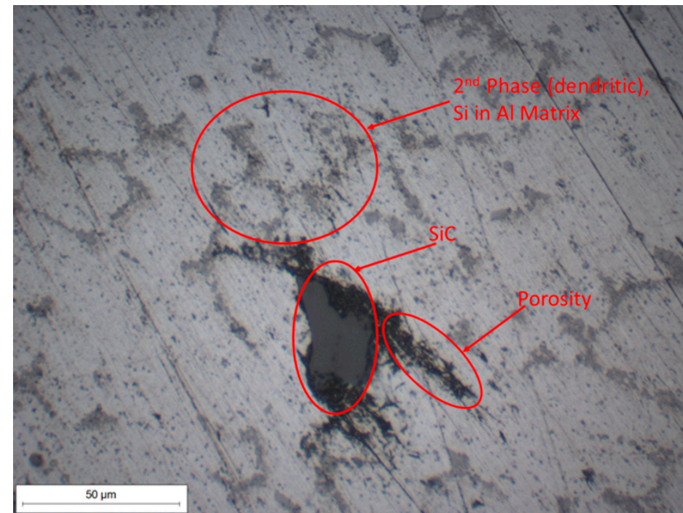
**Figure 2.** EDS elemental distributions for RHA and RHA+SiC powder mix.

**Table 2.** EDS compositions of pure RHA and a mix of RHA+SiC powders.

Elements	Weight Percentage	
	RHA	RHA+SiC
C	24.50	27.09
O	48.61	21.54
Mg	2.63	0.07
Si	13.17	50.30
P	4.88	0.14
K	6.20	0.87

A small axial cross-section from the mostly highly reinforced sample was mounted, ground, and polished, then observed under an optical microscope (Figure 3). It can easily be observed that the hybrid composites contained two phases. One phase is light grey in colour, and the other phase can be seen as dark grey dendritic phases dispersed throughout the matrix. These two phases can indicate the base metal matrix and Si dissolved in HMMC, respectively. To support this evidence, the EDS spectrum of the respective composite was sufficient, as it showed a peak of elemental Si. During preheating and mixing in the molten MMC, silica present in the ash might have turned into silicon and dissolved into the metal matrix to form a second phase of Si in Al. On the other hand, inclusions of SiC particles can be seen as very prominent, irregularly shaped dark grey particles embedded in large

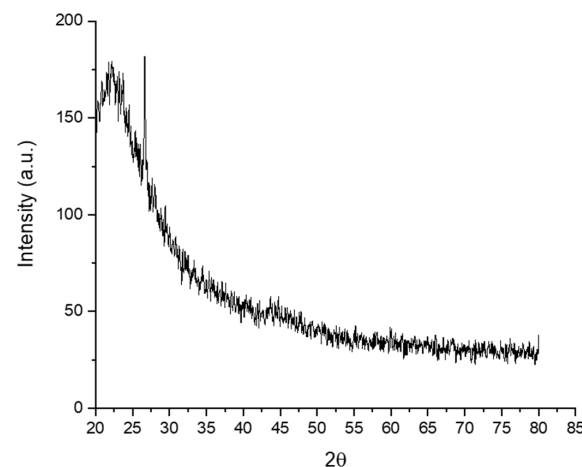
groups or pockets throughout the sample. The ash was amorphous; therefore, it was more likely to dissolve in the liquid aluminium than the SiC powder. However, the dissolving of some SiC could not be completely ruled out.



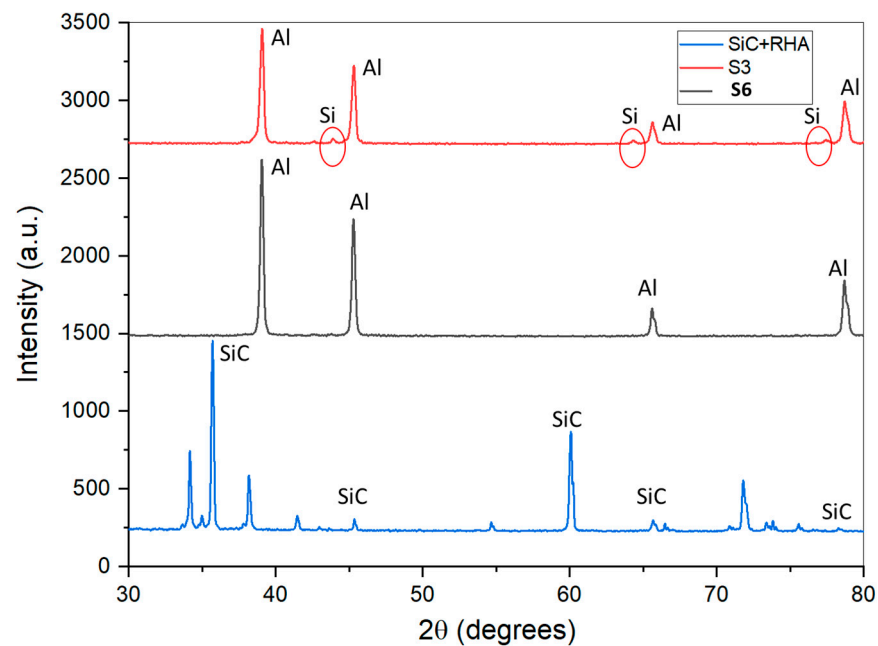
**Figure 3.** Axial cross-section of 8SiC/2.5RHA, observed under a microscope.

### 3.2. XRD Analysis

The spectrum of RHA shown in Figure 4 resembles the spectrum of an amorphous material, and is a typical graph of amorphous silica with no pronounced peaks [28,29]. The XRD spectra of pure AA8011, HMMC with 4% SiC reinforcement, and powder mix are shown together in Figure 5. It was observed that the spectrum of pure Al alloy (S6) was in accordance with JCPDS card no. 65-2869 [30], showing Al peaks. The spectrum of the powder mix (RHA+SiC) was also in accordance with JCPDS card no. 29-1129, with peaks of SiC. The HMMC spectrum showed additional small peaks compared to the spectrum of the original Al alloy. These small peaks in the S3 spectrum were, however, not present in the RHA+SiC spectrum. If the weak peaks in S3 represented SiC phase coming from the powders, much higher peaks should have been found in the very same angles in the RHA+SiC spectrum, but they were not present. This indicated that the small peaks could come from the silicon phase identified in the optical microscopic images, while the amount of the SiC phase is probably too low to be found by XRD. This is certainly an interesting observation and requires further investigation.



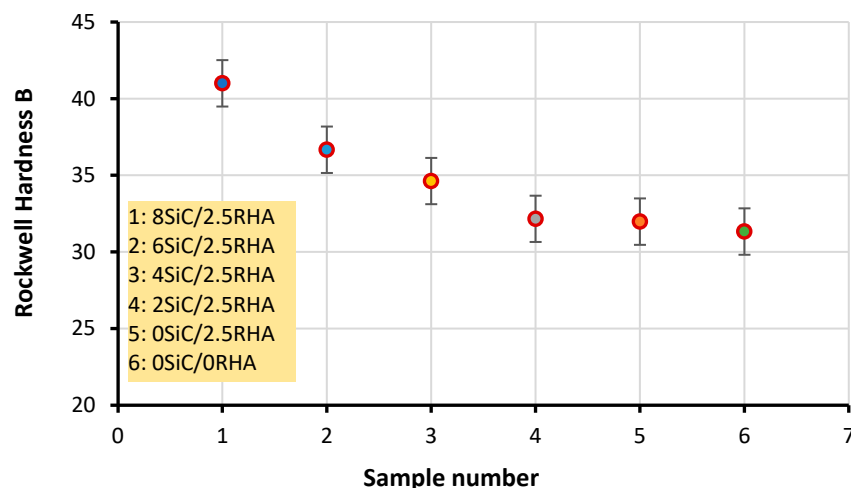
**Figure 4.** XRD spectrum of RHA.



**Figure 5.** XRD spectra of pure AA8011 alloy (S6), HMMC with 4SiC/2.5RHA (S3), and RHA+SiC powder mix.

### 3.3. Hardness

The variation in Rockwell hardness across different samples is shown in Figure 6. It was clearly evidenced that the reinforcement had a positive effect on the hardness. The samples with the highest percentage by weight of reinforcement were the hardest, while the original Al alloy displayed the lowest hardness. The effect of RHA when added to the pure alloy (S5) is not quite clear, as there was no significant difference between S5 and S6. The slight increase in hardness can be explained by the fact that the RHA dissolved in the liquid AA8011 alloy and caused chemical alloying and second-phase precipitation. However, this explanation needs to be interpreted with caution.



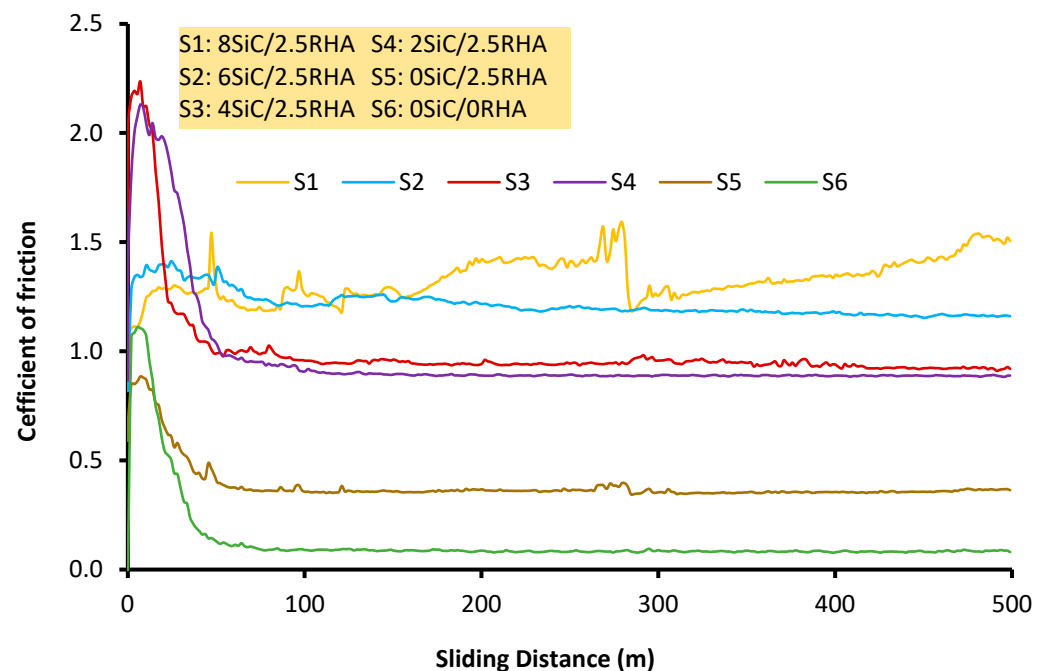
**Figure 6.** Variation in hardness, with sample numbers.

### 3.4. Wear Tests Results

#### 3.4.1. Coefficient of Friction

The plot of COF against the sliding distance, in meters, at 200 °C is shown in Figure 7. It was noticed that, for the first few rotations, the COF rose exponentially to a peak value due to the higher friction force. This explains the opposing behaviour of dry friction and

transition from static to kinematic friction. After this transition, COF attained a lower asymptotic value, possibly due to the steady state being reached. For Sample 5, the COF peak value decreased and the asymptotic value at the steady state increased compared to that of Sample S6. The former might be attributed to the lubrication effect of RHA, and the latter might have resulted from the RHA's abrasiveness. However, COF significantly increased with the amount of SiC particles, possibly due to its abrasive effect compared to the samples (S5 and S6) without any SiC. However, it should be noted that, compared to the pure alloy, the 8SiC/2.5RHA (S1) reinforced composite showed significantly higher COF.



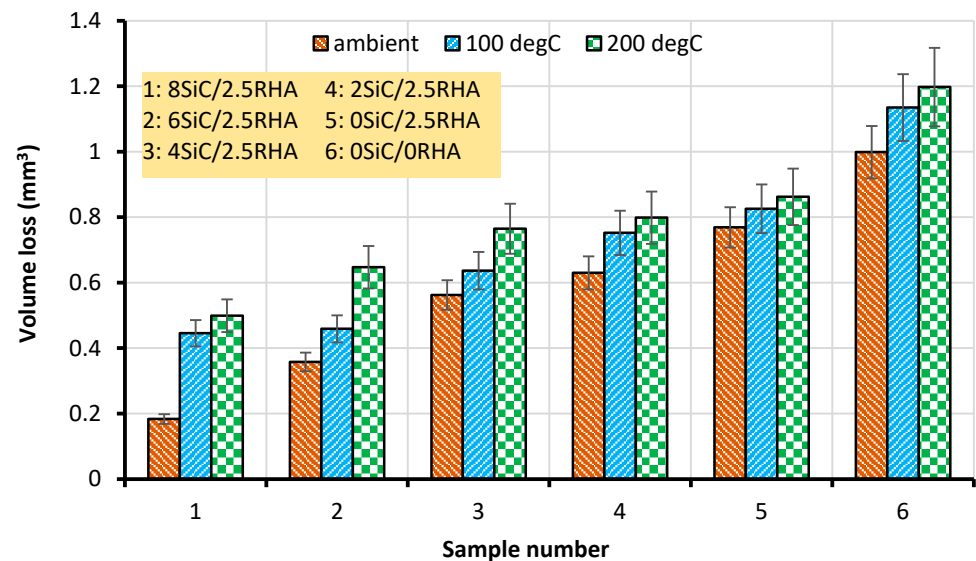
**Figure 7.** Variation in COF with the sliding distance at 200 °C.

The values of the coefficient of friction, as reported by researchers for AMMCs, usually fall between 0.25–0.65 depending upon testing parameters such as load, speed, etc. [31–34]. In this case, however, the COF values for samples S1, S2 and S3 were considerably higher than these values. This could be explained by the high amount of Si reinforcement added to the base alloy and the Si coming from both of the abrasive particles. RHA could add to this extraordinary high coefficient of friction.

### 3.4.2. Wear Volume Loss and Specific Wear

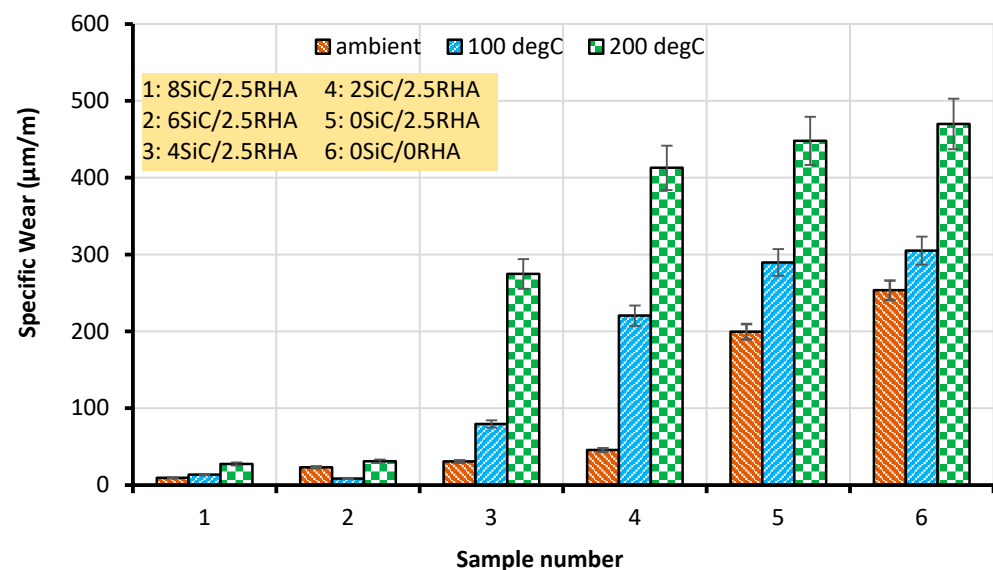
Figure 8 depicts the variation in the loss of material with composition and varying temperatures. The volume loss was measured in all the samples after the same sliding distance of 500 m. This was calculated by carefully weighing each sample before and after the wear test, then dividing it by the average density of the HMMC. The volumetric wear is usually calculated as mm<sup>3</sup> per meter, and it also decreases with increases in the reinforcement [33,34]. It can clearly be observed that the eroded volume increases almost linearly with decreasing reinforcement. As expected, the material loss was at its maximum in the sample without any reinforcement. This implies that addition of abrasive particles of SiC and RHA improves the wear behaviour of the HMMCs. The addition of RHA reduced the wear volume to a certain extent when compared to the pure alloy. The HMMC with 8 wt.% SiC showed the lowest wear volume due to the improvement in hardness compared to the mixed powder. Another factor that affects the wear is the operating temperature. At higher temperatures, the composite suffers severe volume loss, which may be due to

delamination. This is because higher temperatures render the alloys soft, and the wear behaviour transitions from adhesive to delamination.



**Figure 8.** Variation in the volume loss of different samples at different ambient temperatures.

Specific wear was calculated as the total amount of worn pin material ( $\mu\text{m}$ ) over the total sliding distance (m). Figure 9 shows a plot of the variation in specific wear ( $\mu\text{m}/\text{m}$ ), with composition and operating temperatures. The wear increased with increasing temperatures and with decreasing percentages of reinforcement. It can be inferred from the plot that at 200°C, for the composite with no reinforcement, the wear was most severe, while the best tribological performance was exhibited by the composite with the highest reinforcement at an ambient temperature. With the addition of only RHA, the specific wear did not decrease significantly when compared to the non-reinforced Al alloy. However, the HMMC with 8.0 wt.% SiC reduced the specific wear by nearly a factor of 20 at 200 °C.



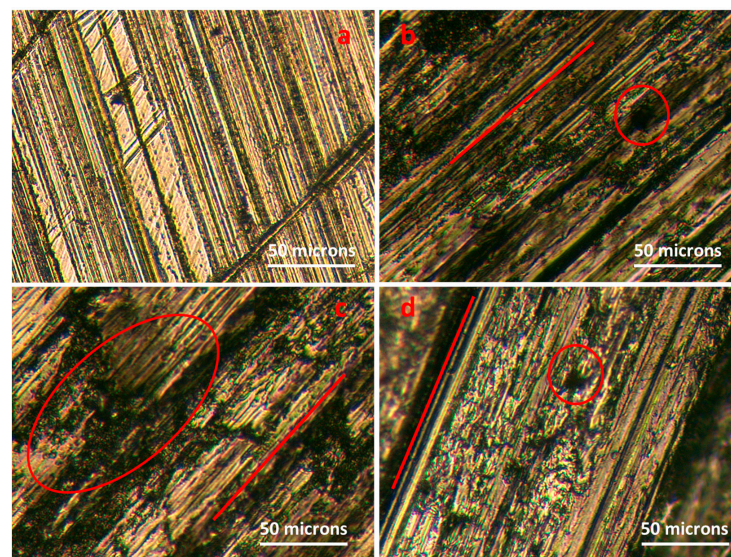
**Figure 9.** Variation in specific wear at different temperatures.

### 3.4.3. Worn Surface of the Disc

For studying wear in tribological systems, a tribometer is used. A tribometer is a pair of a pin and a disc, where the pin is a softer material sliding over the much harder disc.



Thus, it is important to analyse the surfaces of both the pin and the disc to gain better insight into the wear process. The wear track of the disc was observed under an optical microscope at a magnification of  $\times 50$ , and the images are shown in Figure 10. The wear tracks show that fine grooves were developed on the surface of the disc. The losses of material from the disc surface are marked by red enclosures, and the sliding directions are marked with red lines. In Figure 10c, some small particles of the pin can be seen inside the red oval. This can be due to the erosion of the pin surface at a high temperature of 200 °C. The delamination of the pin surfaces may have caused the Al particles to scatter on the wear track, and high temperature could have caused the welding of the pin particles on the wear track due to the softening of the pin material. However, the presence of the pin material on the wear track cannot be confirmed by EDX analysis due to the large size of the disk. However, the wear of the pin track was much less severe than that observed on the worn pin surfaces. This was because the disc material was much harder than the pin material.



**Figure 10.** Optical microscopic images of the wear track of the disc at  $\times 50$  when testing the wear of 8SiC/2.5RHA at 200 °C: (a) ground disc surface before wear; images of the wear track at (b) position 1, (c) position 2, and (d) position 3.

#### 3.4.4. Worn Surface Morphology of the Pin

The worn surfaces of the pins tested at 200 °C are shown in Figure 11. The maximum loss of material due to wear was observed in the unreinforced AA-8011, and minimum wear was observed in the 8SiC/2.5RHA HMMC. This result is aligned with the improvement in the hardness of the samples which was observed with greater reinforcement. Some interesting findings regarding the wear surface morphology are discussed below:

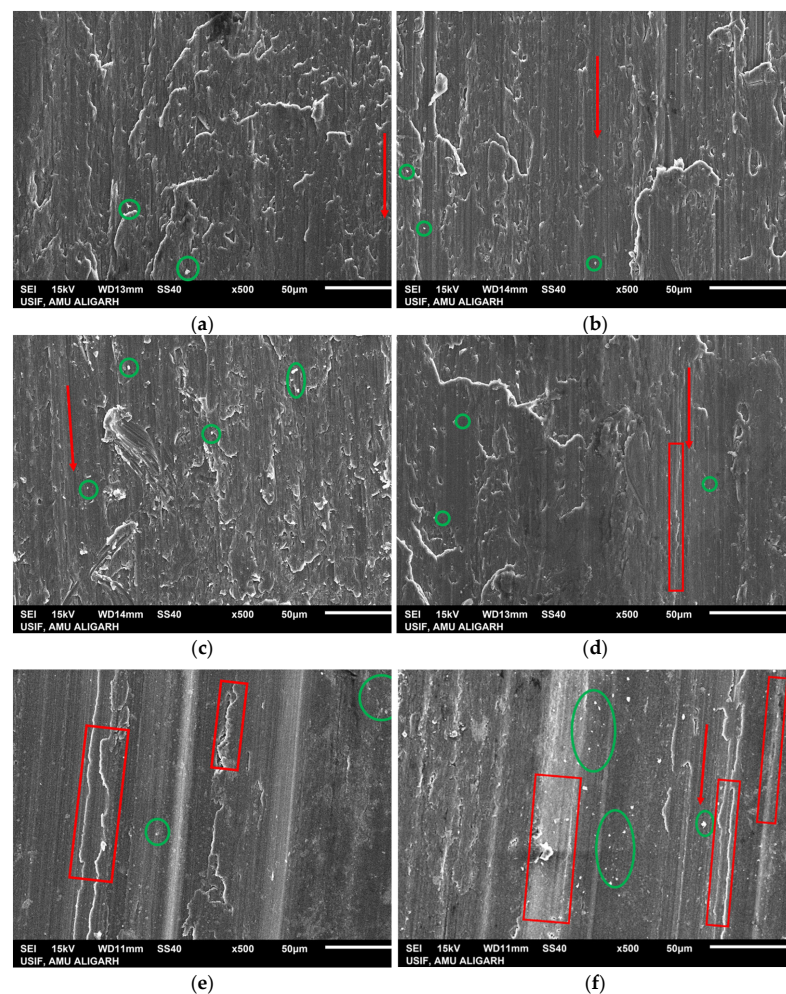
- In the samples with higher reinforcement, larger amounts of SiC inclusions could be observed, while the wear mechanisms were mostly adhesive and abrasive, with almost negligible delamination and wear debris. It can further be stated that strong adhesion and intermetallic bonding between the ductile AA8011 and hard SiC particles were responsible for bearing the load and enhancing hardness, restricting the wear of the surface of the composite [35].
- As the amount of reinforcement increased, the wear behaviour can possibly transition from slightly adhesive to abrasive, due to the presence of more abrasive SiC particles in the metal matrix. Overall, the predominant wear mechanism was abrasive in the HMMCs. Similar observations have been reported by other researchers in the literature [36–38].
- Another important observation could be the presence of small, almost circular bright spots scattered on the worn surface (marked with green circles), identified thanks



to closer inspection. These could possibly be explained as wear debris, which was also reported by Mazahari and Shabani. They reported that this debris is found due to the propagation of cracks perpendicular to the sliding direction in the Al alloys with lower reinforcement [39]. The presence of larger amounts of this debris in the samples with less and no reinforcement indicates that, most probably, the material was being removed at a rate which did not allow for oxide film to be formed on the softer metal matrix surface [19]. In summary, the addition of RHA and SiC to AA8011 metal composites improved the wear resistance by affecting multiple wear mechanisms, including abrasion and adhesion.

The hard and abrasive particles of SiC act as reinforcements and increase the hardness and wear resistance of the composite material. Further, the abrasive particles could form a protective layer on the surface of the composite material, which can prevent adhesion between the mating surfaces and reduce the amount of wear.

- d. In the cases of 0SiC/2.5RHA and 0SiC/0RHA, plastic deformation and subsequent delamination were observed due to the comparatively softer nature of the materials. It was most severe in the latter case, as it was unreinforced and more prone to delamination at higher temperatures. This indicated that there was a minor effect of RHA on the reduction in delamination wear. This observation strongly supported the results obtained regarding the increased loss of weight in the unreinforced alloy [11].



**Figure 11.** SEM micrographs of worn pin surfaces at 200 °C. (a) 8SiC/2.5RHA, (b) 6SiC/2.5RHA, (c) 4SiC/2.5RHA, (d) 2SiC/2.5RHA, (e) 0SiC/2.5RHA, and (f) 0SiC/0RHA, at  $\times 500$  magnification. Sliding direction, debris, and delamination are highlighted by red arrows, green circles, and red rectangles, respectively.

The results of the wear behaviour under different operating temperatures were also consistent with the previous research conducted in this field. However, at any fixed temperature, the best performance was exhibited by the composite with the highest SiC/RHA reinforcement. Therefore, the reinforced composites could be a better option for use in high-temperature applications. As the use of AA8011 alloys is prevalent in the automotive sector, the use of reinforced AA8011 alloys will provide a suitable option for parts for which relative movement is involved.

#### 4. Conclusions

Hybrid metal matrix composites of AA8011 Al alloys were manufactured using the stir casting technique, with a fixed content (2.5 wt.%) of RHA and varying percentage of SiC (0, 2, 4, 6, 8 wt.%). From the microstructural, hardness, and wear testing results, the following conclusions can be drawn:

RHA particles showed bimodal particle distribution with smaller particles (7  $\mu\text{m}$  to 15.6  $\mu\text{m}$ ) and larger particles (46.8  $\mu\text{m}$  to 156.2  $\mu\text{m}$ ) with amorphous structures. The RHA particles completely dissolved in the metal matrix to form a second phase of Si in Al, which was spread homogeneously throughout the matrix. From the surface morphology of the cast composites, a distribution of large groups or pockets of SiC particles throughout the AA8011 matrix was evident.

The effect of Al alloy reinforcement with SiC+RHA on the hardness and wear behaviour was consistent, with different weight percentages used in the synthesis of HMMCs. Although there was no significant difference in hardness observed after adding RHA in the aluminium alloy, SiC addition significantly increased the composite hardness. The tribological performance increased almost linearly with the level of reinforcement. The best performance was exhibited by the 8 wt.% SiC+RHA composite, while the lowest resistance to wear was exhibited by the pure cast Al alloy. The results of the wear behaviour under different operating temperatures were also consistent, while the wear properties degraded with increasing temperatures, irrespective of the amount of reinforcement. However, at a fixed temperature, the best performance was exhibited by the composite with the highest SiC/RHA reinforcement. It should also be noted that COF increased with the addition of both RHA and SiC particles, although the latter displayed a significant effect. Therefore, hybrid composites could be a better option for use as alternatives to cast iron in automotive engine components.

**Author Contributions:** Conceptualisation, S.M., J.H., P.M. and Q.M.; methodology, S.M. and J.H.; validation, J.H.; formal analysis, S.M., J.H., P.M. and Q.M.; investigation, S.M. and Q.M.; resources, Q.M.; data curation, S.M.; writing—original draft preparation, S.M.; writing—review and editing, S.M., J.H., P.M. and Q.M.; visualisation, S.M.; supervision, P.M. and J.H.; project administration, Q.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available within the article.

**Acknowledgments:** The authors sincerely acknowledge the facility provided by Delhi Technological University (D.T.U.), New Delhi, India, to prepare the stir-cast samples. The authors also acknowledge the facilities provided by Aligarh Muslim University (AMU) for the characterisation of the samples. We would like to thank Unais Sait (Free University of Bozen-Bolzano, Italy) for helping us to structure this article.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Ngo, T.-D. (Ed.). Introduction to Composite Materials. In *Composite and Nanocomposite Materials*; IntechOpen: Rijeka, Croatia, 2020.
2. Verma, A.; Chauhan, S.S.; Dwivedi, S.P. Review Paper on Thermal Expansion and Tribological Behavior of Composite Materials. *Mater. Today Proc.* **2023**, *79*, 235–246. [\[CrossRef\]](#)
3. Mavhungu, S.; Akinlabi, E.; Onitiri, M.; Varachia, F. Aluminum Matrix Composites for Industrial Use: Advances and Trends. *Procedia Manuf.* **2017**, *7*, 178–182. [\[CrossRef\]](#)
4. Miteva, A.; Petrova, A.; Stefanov, G. Surface Oxidation of Al-Si Alloys at Elevated Temperatures. *Appl. Eng. Lett. J. Eng. Appl. Sci.* **2021**, *6*, 105–110. [\[CrossRef\]](#)
5. Reddy, K.S.K.; Lekha, B.C.; Sakshi, K.U.; Chouhan, M.S.; Karthikeyan, R.; Aparna, S. Effect of Different Reinforcements on Aluminium Composite Properties—A Review. *Mater. Today Proc.* **2022**, *62*, 3963–3967. [\[CrossRef\]](#)
6. Yigezu, B.S.; Mahapatra, M.M.; Jha, P.K. Influence of Reinforcement Type on Microstructure, Hardness, and Tensile Properties of an Aluminum Alloy Metal Matrix Composite. *J. Miner. Mater. Charact. Eng.* **2013**, *1*, 33948.
7. Singh, K.M.; Chauhan, A.K. Fabrication, Characterization, and Impact of Heat Treatment on Sliding Wear Behaviour of Aluminium Metal Matrix Composites Reinforced with B4C. *Adv. Mater. Sci. Eng.* **2021**, *2021*, 1–9. [\[CrossRef\]](#)
8. Singh, J. Fabrication Characteristics and Tribological Behavior of Al/SiC/Gr Hybrid Aluminum Matrix Composites: A Review. *Friction* **2016**, *4*, 191–207. [\[CrossRef\]](#)
9. Chaudhari, A.D.; Danej, A.A.; Nirbhavane, P.S.; Shinde, S.S.; Pawar, S.Y. Wear Behaviour Analysis of Silicon Carbide Based Aluminium Metal Matrix Composites. *Int. Res. J. Eng. Technol.* **2020**, *7*, 5754–5759.
10. Daniel, A.A.; Murugesan, S.; Sukkasamy, S. Others Dry Sliding Wear Behaviour of Aluminium 5059/SiC/MoS<sub>2</sub> Hybrid Metal Matrix Composites. *Mater. Res.* **2017**, *20*, 1697–1706. [\[CrossRef\]](#)
11. Subramaniam, B.; Natarajan, B.; Kaliyaperumal, B.; Chelladurai, S.J.S. Wear Behaviour of Aluminium 7075—Boron Carbide-Coconut Shell Fly Ash Reinforced Hybrid Metal Matrix Composites. *Mater. Res. Express* **2019**, *6*, 1065d3. [\[CrossRef\]](#)
12. David Raja Selvam, J.; Dinaharan, I.; Mashinini, P. High Temperature Sliding Wear Behavior of AA6061/Fly Ash Aluminum Matrix Composites Prepared Using Compocasting Process. *Tribol.-Mater. Surf. Interfaces* **2017**, *11*, 39–46. [\[CrossRef\]](#)
13. Aghaie-Khafri, M. Formability of AA8011 Aluminum Alloy Sheet in Homogenized and Unhomogenized Conditions. *J. Mater. Sci.* **2004**, *39*, 6467–6472. [\[CrossRef\]](#)
14. Fayomi, J.; Popoola, A.; Popoola, O.; Fayomi, O.; Ajenifuja, E. Response Evaluation of AA8011 with Nano ZrB<sub>2</sub> Inclusion for Multifunctional Applications: Considering Its Thermal, Electrical, and Corrosion Properties. *J. Alloys Compd.* **2021**, *853*, 157197. [\[CrossRef\]](#)
15. Karthikeyan, A.; Nallusamy, S. Experimental Analysis on Sliding Wear Behaviour of Aluminium-6063 with SiC Particulate Composites. *Int. J. Eng. Res. Afr.* **2017**, *31*, 36–43. [\[CrossRef\]](#)
16. Benal, M.M.; Shivanand, H. Effects of Reinforcements Content and Ageing Durations on Wear Characteristics of Al (6061) Based Hybrid Composites. *Wear* **2007**, *262*, 759–763. [\[CrossRef\]](#)
17. Hillary, J.J.M.; Ramamoorthi, R.; Chelladurai, S.J.S. Dry Sliding Wear Behaviour of Al6061–5% SiC—TiB<sub>2</sub> Hybrid Metal Matrix Composites Synthesized by Stir Casting Process. *Mater. Res. Express* **2020**, *7*, 126519. [\[CrossRef\]](#)
18. Karthikkumar, C.; Baranirajan, R.; Premnauth, I.; Manimaran, P. Investigations on Mechanical Properties of Al 8011 Reinforced with Micro B4C/Red Mud by Stir Casting Method. *Int. J. Eng. Res. Gen. Sci.* **2016**, *4*, 405.
19. Pramanik, A. Effects of Reinforcement on Wear Resistance of Aluminum Matrix Composites. *Trans. Nonferrous Met. Soc. China* **2016**, *26*, 348–358. [\[CrossRef\]](#)
20. Zhu, H.; Jar, C.; Song, J.; Zhao, J.; Li, J.; Xie, Z. High Temperature Dry Sliding Friction and Wear Behavior of Aluminum Matrix Composites (Al<sub>3</sub>Zr +  $\alpha$ -Al<sub>2</sub>O<sub>3</sub>)/Al. *Tribol. Int.* **2012**, *48*, 78–86. [\[CrossRef\]](#)
21. Al-Salihi, H.A.; Mahmood, A.A.; Alalkawi, H.J. Mechanical and Wear Behavior of AA7075 Aluminum Matrix Composites Reinforced by Al<sub>2</sub>O<sub>3</sub> Nanoparticles. *Nanocomposites* **2019**, *5*, 67–73. [\[CrossRef\]](#)
22. Hossain, S.S.; Mathur, L.; Roy, P. Rice Husk/Rice Husk Ash as an Alternative Source of Silica in Ceramics: A Review. *J. Asian Ceram. Soc.* **2018**, *6*, 299–313. [\[CrossRef\]](#)
23. Bulei, C.; Stojanovic, B.; Utu, D. Developments of Discontinuously Reinforced Aluminium Matrix Composites: Solving the Needs for the Matrix. *Proc. J. Phys. Conf. Ser.* **2022**, *2212*, 012029. [\[CrossRef\]](#)
24. Bhushan, B. *Introduction to Tribology*; John Wiley & Sons: Hoboken, NJ, USA, 2013.
25. ASTM G99-17 2017; Standard Test Method for Wear Testing with a Pin-on-Disk Apparatus. ASTM International: West Conshohocken, PA, USA, 2017.
26. García-Miranda, J.S.; Aguilera-Camacho, L.D.; Hernández-Sierra, M.T.; Moreno, K.J. A Comparative Analysis of the Lubricating Performance of an Eco-Friendly Lubricant vs Mineral Oil in a Metallic System. *Coatings* **2023**, *13*, 1314. [\[CrossRef\]](#)
27. Karthikeyan, A.; Jinu, G. Investigation on Mechanical and Corrosion Behaviour of AA8011 Reinforced with TiC and Graphite Hybrid Composites. *Mater. Res. Express* **2019**, *6*, 1065b5. [\[CrossRef\]](#)
28. Liou, T.-H.; Chang, F.-W.; Lo, J.-J. Pyrolysis Kinetics of Acid-Leached Rice Husk. *Ind. Eng. Chem. Res.* **1997**, *36*, 568–573. [\[CrossRef\]](#)
29. Azizi, S.N.; Yousefpour, M. Spectroscopic Studies of Different Kind of Rice Husk Samples Grown in North of Iran and the Extracted Silica by Using XRD, XRF, IR, AA and NMR Techniques. *Eurasian J. Anal. Chem.* **2008**, *3*.

30. Arif, S.; Jamil, B.; Shaikh, M.B.N.; Aziz, T.; Ansari, A.H.; Khan, M. Characterization of Surface Morphology, Wear Performance and Modelling of Graphite Reinforced Aluminium Hybrid Composites. *Eng. Sci. Technol. Int. J.* **2020**, *23*, 674–690. [[CrossRef](#)]
31. Mistry, J.M.; Gohil, P.P. An Overview of Diversified Reinforcement on Aluminum Metal Matrix Composites: Tribological Aspects. *Proc. Inst. Mech. Eng. Part J J. Eng. Tribol.* **2017**, *231*, 399–421. [[CrossRef](#)]
32. Padmavathi, K.; Ramakrishnan, R. Tribological Behaviour of Aluminium Hybrid Metal Matrix Composite. *Procedia Eng.* **2014**, *97*, 660–667. [[CrossRef](#)]
33. Lakshmikanthan, A.; Angadi, S.; Malik, V.; Saxena, K.K.; Prakash, C.; Dixit, S.; Mohammed, K.A. Mechanical and Tribological Properties of Aluminum-Based Metal-Matrix Composites. *Materials* **2022**, *15*, 6111. [[CrossRef](#)]
34. Naik, M.H.; Manjunath, L.; Koti, V.; Lakshmikanthan, A.; Koppad, P.G.; Kumaran, S.P. Al/Graphene/CNT Hybrid Composites: Hardness and Sliding Wear Studies. *FME Trans.* **2021**, *49*, 414–421. [[CrossRef](#)]
35. Fayomi, J.; Popoola, A.P.I.; Popoola, O.M.; Oladijo, O.P.; Fayomi, O.S.I. Tribological and Microstructural Investigation of Hybrid AA8011/ZrB<sub>2</sub>-Si<sub>3</sub>N<sub>4</sub> Nanomaterials for Service Life Improvement. *Results Phys.* **2019**, *14*, 102469. [[CrossRef](#)]
36. Sharma, A.; Belokar, R.; Kumar, S. Dry Sliding Wear Characterization of Red Mud Reinforced Aluminium Composite. *J. Braz. Soc. Mech. Sci. Eng.* **2018**, *40*, 1–12. [[CrossRef](#)]
37. Thirumalai Kumaran, S.; Uthayakumar, M. Investigation on the Dry Sliding Friction and Wear Behavior of AA6351-SiC-B<sub>4</sub>C Hybrid Metal Matrix Composites. *Proc. Inst. Mech. Eng. Part J J. Eng. Tribol.* **2014**, *228*, 332–338. [[CrossRef](#)]
38. Sahin, Y.; Murphy, S. The Effect of Sliding Speed and Microstructure on the Dry Wear Properties of Metal-Matrix Composites. *Wear* **1998**, *214*, 98–106. [[CrossRef](#)]
39. Mazahery, A.; Shabani, M.O. Ascending Order of Enhancement in Sliding Wear Behavior and Tensile Strength of the Compocast Aluminum Matrix Composites. *Trans. Indian Inst. Met.* **2013**, *66*, 171–176. [[CrossRef](#)]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

# Topological phase transition and tunable surface states in YBi

Ramesh Kumar<sup>1</sup>, Mukhtiyar Singh<sup>1\*</sup>

<sup>1</sup>Department of Applied Physics, Delhi Technological University, New Delhi-110042, India

\*[mukhtiyarsingh@dtu.ac.in](mailto:mukhtiyarsingh@dtu.ac.in); [msphysik09@gmail.com](mailto:msphysik09@gmail.com)

A unique co-existence of extremely large magnetoresistance (XMR) and topological characteristics in non-magnetic rare-earth monpnictides stimulating intensive research on these materials. Yttrium monobismuthide (YBi) has been reported to exhibit XMR up to 10<sup>5</sup>% but its Topological properties still need clarification. Here we use the hybrid density functional theory to probe the structural, electronic and topological properties of YBi in detail. We observe that YBi is topologically trivial semimetal at ambient pressure which is in accordance with reported experimental results. The topological phase transitions i.e., trivial to non-trivial are obtained with volumetric pressure of 6.5 GPa and 3% of epitaxial strain. This topological phase transitions are well within the structural phase transition of YBi (24.5 GPa). The topological non-trivial state is characterized by band inversions among *Y-d* band and *Bi-p* band near  $\Gamma$ - and *X-point* in the Brillouin zone. This is further verified with the help of surface band structure along (001) plane. The  $Z_2$  topological invariants are calculated with the help of product of parities and evolution of Wannier charge centers. The occurrence of non-trivial phase in YBi with a relatively small epitaxial strain, which a thin film geometry can naturally has, might make it an ideal candidate to probe inter-relationship between XMR and non-trivial topology.

## I. INTRODUCTION

Topological Insulators (TIs) [1–3] with unique metallic surface states and insulating bulk band gap have grabbed high attention in condensed matter physics. These topological surface states are protected by time reversal symmetry (TRS) which separates them from conventional insulators [4–6]. These surface states are spin-momentum locked and robust against local perturbations [4,7]. Some semimetals are also reported to exhibit non-trivial topological states and can be divided into various categories e.g., Weyl, Dirac and nodal-line semimetals, etc. [8–10]. Breaking of TRS or spatial symmetry transforms Dirac semimetals into Weyl semimetals and a Dirac point splits into a pair of Weyl points [11]. The surface states in Dirac and Weyl semimetals can be described by Fermi arc [12–14], unlike the Dirac cone in case of Tis, which is due to the overlaps between the surface and the bulk states. These topological semimetals can be characterized by a band inversion in the bulk band structure and the  $Z_2$  topological invariants which

can be calculated with the help of product of parities at time reversal invariant momenta (TRIM) points as well as using Wilson loop [2,15,16]. Several rare-earth pnictides such as  $\text{LnPn}$  (Ln: rare earth element; Pn: As, Sb, Bi) have reported to exhibit topological states at ambient conditions [17–19]. Spin-orbit coupling (SOC) is a viable tool to tune a trivial topological material to non-trivial one. The SOC strength can be enhanced using pressure, strain, chemical doping, and alloying [19–22] etc. Amongst these, external pressure and strain are most suitable owing to their non-disruptive nature. The effect of volumetric pressure or epitaxial strain on any material reduces its bond length in the respective directions, bandwidth, and energy differences across the bands without affecting charge neutrality or stoichiometry. Various semimetals such as LaAs [20], LaSb [21], TmSb [23], TaAs [24] and YbAs [25] have been shown to become topologically non-trivial with pressure. Topological phase in LaSb [26], and SnTe [27] has also been observed under epitaxial strain and the same has been confirmed experimentally by angle-resolved photoemission spectroscopy (ARPES) in SnTe [27].

The YBi has been reported to be a perfectly compensated semimetal with XMR up to 10<sup>5</sup>% with equal hole and electron carrier concentration [28,29]. *First-principles* calculations within Perdew-Burke-Ernzerhof (PBE) functional have predicted YBi as a topologically non-trivial semimetal with band inversion near  $\Gamma$ -point [30]. Recent study using ARPES and *first-principles* calculations with mBJ functional has indicated that YBi is a topologically trivial semimetal [29]. This disparity raises a debate on true topological nature of YBi and deserves a thorough analysis as it may be useful to settle down the debate concerning whether XMR is caused by nontrivial topological aspects or complete electron-hole compensation.

This motivates us to systematically explore structural, electronic, and topological properties of YBi using density functional theory (DFT) with relatively accurate hybrid functional Heyd, Scuseria and Ernzerhof (HSE06). This functional had predicted accurate electronic states of other similar rare earth monopnictides and gave carrier densities in good agreements with experimental results [20,21,23,25,26]. We study the topological properties under external volumetric pressure and epitaxial strain and analysed the quantum phase transitions in detail. The topological states are observed with the help of band inversion in bulk band structure and surface Dirac cone projected on (001) plane. The  $\mathbb{Z}_2$  topological invariants are calculated using parities of wavefunctions TRIM points and evolution of Wannier charge centers (WCCs).



## II. COMPUTATIONAL DETAILS

All structural and electronic calculations of YBi with applied volumetric pressure and epitaxial strain was carried out in the framework of DFT [31–33] based *first-principles* approach with projector augmented wave (PAW) [33] technique as implemented in VASP code [34]. The PBE [35] functional followed by screened hybrid functional HSE06 [36,37] was used for more accurate results. The long range and short-range parts of HSE06 was employed with screening parameters as  $\omega=0.201 \text{ \AA}^{-1}$ . The PBE functional used for long range part but a mixing of 25% Fock exchange was carried out in short range part of HSE06 functional. The PAW potentials used for Y and Bi were having eleven valance electrons (i.e.,  $4s^2 4p^5 5s^1 4d^2$ ) and fifteen valance electrons (i.e.,  $5d^{10} 6s^2 6p^3$ ), respectively. An optimized Monkhorst-Pack type k-mesh of  $7 \times 7 \times 7$  and kinetic energy cutoff of 340 eV were used to calculate the plane wave basis set. Gaussian smearing method was set at width of 0.001 eV for Fermi level broadening and all atomic positions were fully relaxed. To apply the volumetric pressure and epitaxial stain, we were used eight atom cubic unit cell. Out of which, two atom primitive unit cell was extracted for band structure calculations. This extraction of primitive unit cell was helpful in avoiding band folding. The effect of SOC was included in band structure calculation. Dynamical stability of YBi under the effect of hydrostatic pressure and epitaxial strain was verified with phonon dispersion calculations using Phonopy code [38]. Product of parities at TRIM points, in the presence of TRS and inversion symmetry (IS), was used to calculate  $Z_2$  topological invariants. To parametrized the tight-binding (TB) Hamiltonian, we were obtained the maximally localised wannier functions (MLWF) using wannier90 code [39]. The surface band structure and WCCs were calculated with surface Green's function methods using WannierTools code [40].

## III. RESULTS AND DISCUSSIONS

### A. Structural and stability analysis, and electronic structure at ambient condition

Alike many other rare-earth mononpnictides, YBi exists in a stable *rocksalt type* (*NaCl-type*) crystal structure having space group  $Fm\bar{3}m$  (#225) with Y (0.5, 0.5, 0.5) and Bi (0, 0, 0) atoms as shown in Fig. 1(a). Its optimized lattice parameter is  $a = 6.338 \text{ \AA}$ , which is in good agreement with previous theoretical and experimental reports as mentioned in Table I.

TABLE I: Lattice parameter and Structural phase transition (SPT) of YBi.

YBi	Previous experimental study	Previous theoretical study	Present study
Lattice parameter, $a$ (Å)	6.2597 [41]	6.3378 [42]; 6.345 [43]; 6.252 [44] ; 6.29 [45]	6.338
SPT in GPa	-----	28.1 [42]; 23 [42]; 24 [44]; 23.4 [45]	24.5

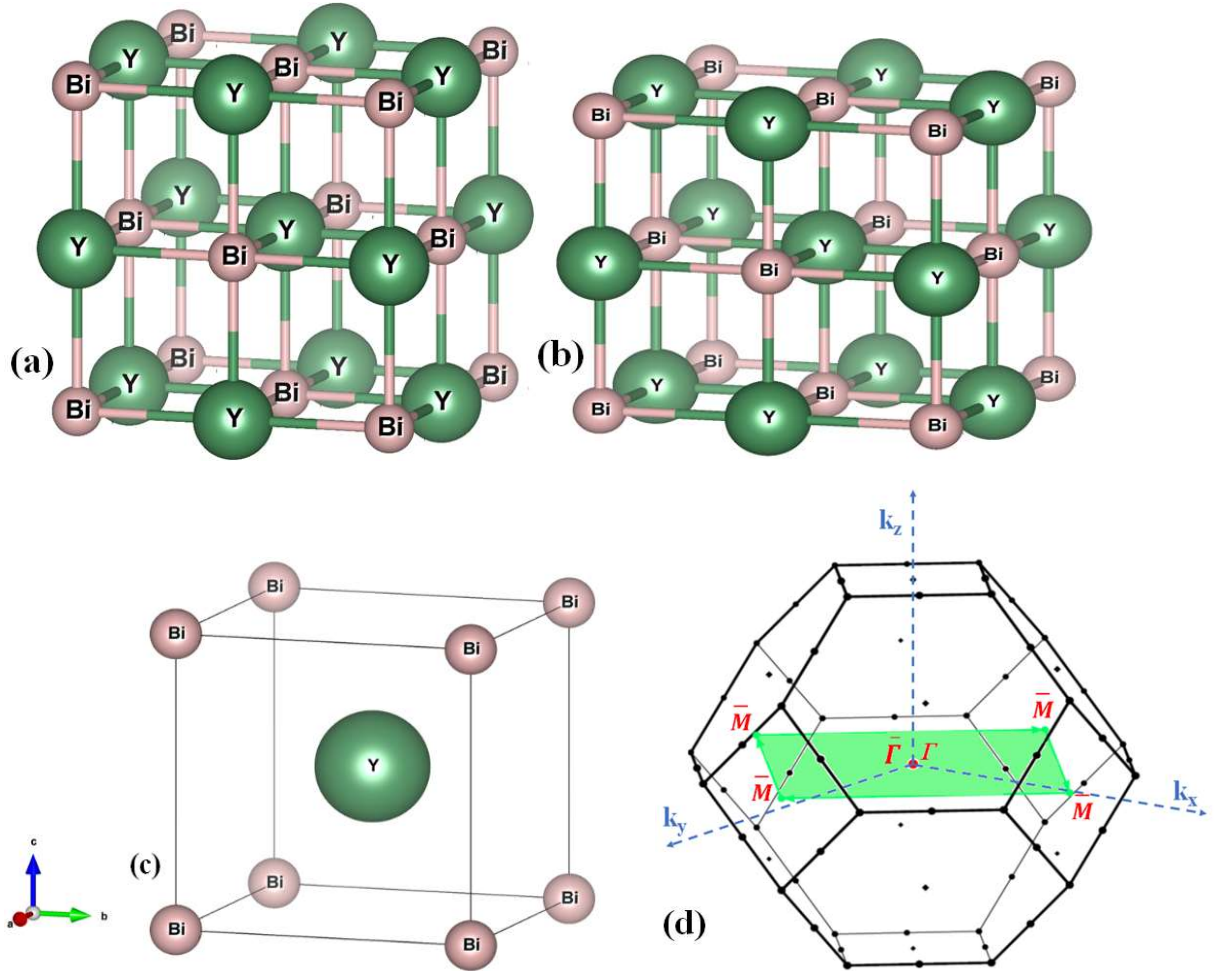


FIG. 1: Crystal structure of YBi in (a) FCC (*NaCl-type*), (b) tetragonal, (c) BCC (*CsCl-type*) structure, (d) The Brillouin zone (BZ) of YBi. The shaded area (green colour) is representing the projection of the bulk Brillouin zone on the (001) surface Brillouin zone (SBZ), with symmetry points in the SBZ displayed (red colour). Here, center of BZ ( $\Gamma$ ) and its projection in SBZ ( $\bar{\Gamma}$ ) are coinciding.

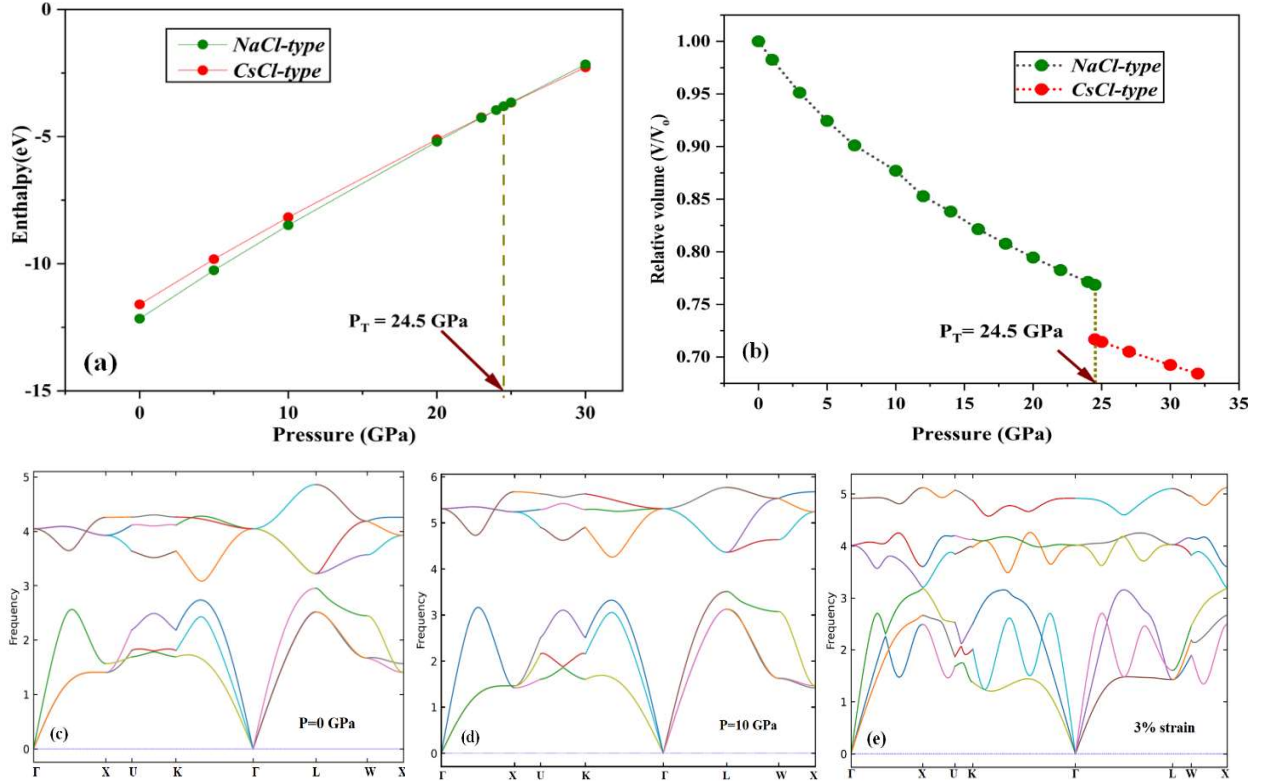
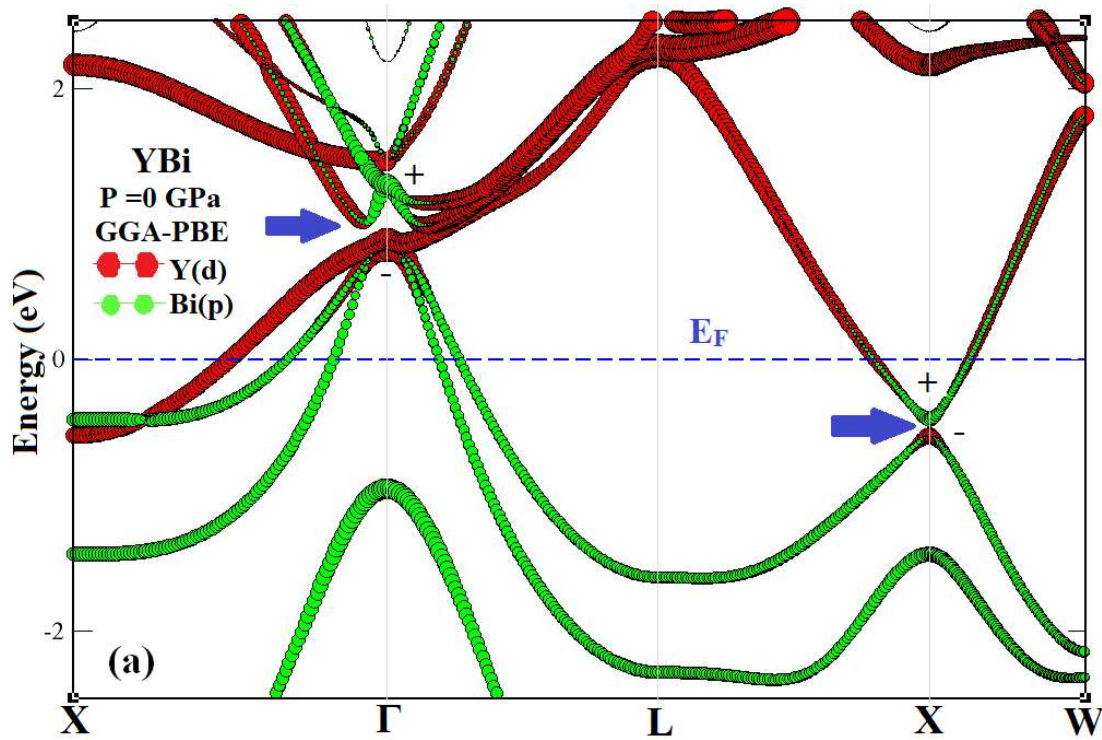


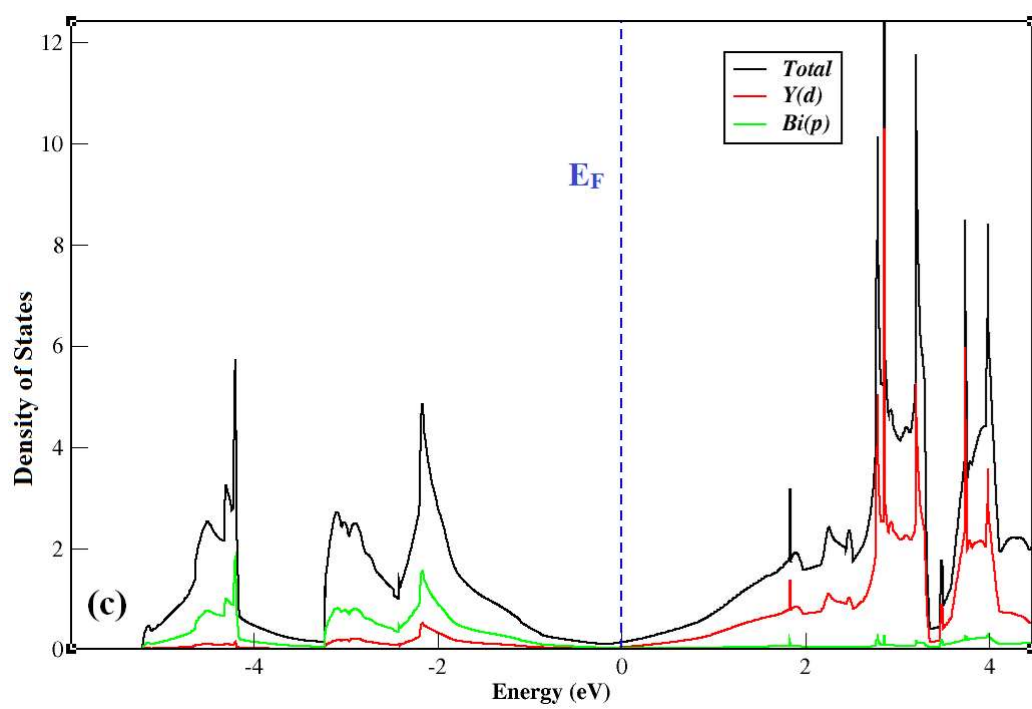
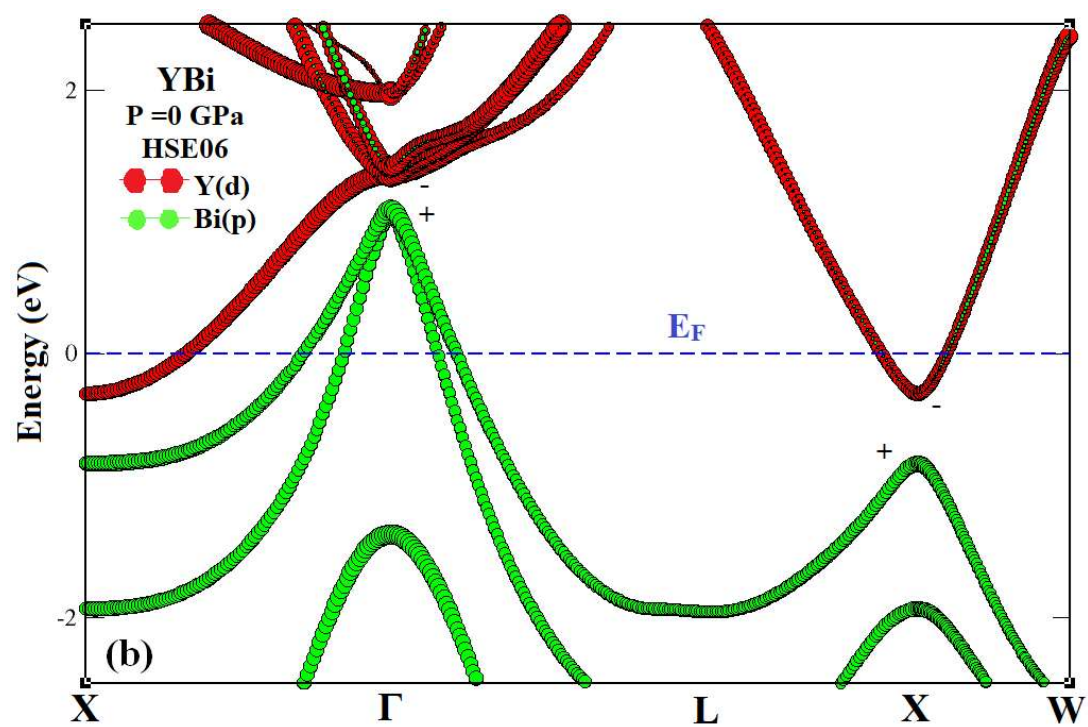
FIG. 2: (a) Enthalpy of YBi as function of pressure for *NaCl*-type to *CsCl*-type structure. (b) Variation in relative volume of YBi as a function of pressure. The phonon dispersion of YBi at (c) 0 GPa, (d) 10 GPa pressure and (e) 3% strain.

The YBi shows structural phase transition (SPT), with applied volumetric pressure, and converts to a *CsCl*-type structure (Fig. 1(c)) from the *NaCl*-type structure as shown in Fig. 2 (a). The stability of a structure at a given pressure can be accessed through Enthalpy which is defined as  $H = E + PV$ , where  $E$  is total energy,  $V$  is volume and  $P$  is the external pressure on unit cell. We found the SPT at around 24.5 GPa which is in good agreement with previous reports as listed in Table I. The variation in the relative volume of YBi with applied volumetric pressure is shown in Fig. 2(b). The sudden change in volume at SPT signifies a first-order phase transition resulting in change of the crystal symmetry. The *rocksalt* crystal structures have truncated octahedron Brillion zone (BZ) under equilibrium conditions, with  $\Gamma$  as a center (Fig. 1(d)). The (001) plane (green colour) containing center of BZ ( $\Gamma$ -point) and  $\bar{M}$ -points at center of squares are shown in Fig.1(d). When we apply volumetric pressure, the changes in the BZ are the same in all directions and it holds its truncated octahedron shape. On the other hand, under epitaxial strain, three  $X$ -points along momentum axis are divided into two in-plane and one out-of-plane  $Z$ -point, and a distorted BZ with preserved inversion symmetry is observed. The dynamical stability of YBi under applied volumetric pressure and epitaxial strain is also analysed. As shown in Fig. 2(c-e), the phonon

dispersion spectrums have no negative frequency which confirms that YBi is dynamically stable and can be realized experimentally under studied pressure and strain conditions.

To establish the true nature of YBi at ambient pressure, we have plotted the band structures with SOC using two functionals i.e., GGA-PBE and HSE06. With former, we found that YBi is topologically trivial semimetal with even (two) number of band inversions between Y-*d* band and Bi-*p* band at  $\Gamma$ - and  $X$ -points as shown in Fig. 3(a). A previous study has identified it as topological semimetal with a single Dirac cone at  $\Gamma$ -point [30]. Further, it has been reported experimentally using ARPES that YBi is topologically trivial having no Dirac cone [29]. To accurately predict the true nature of this material, we have used more accurate hybrid functional HSE06 as shown in Fig. 3(b). It can be seen the partially occupied bands at  $\Gamma$ -point (hole pockets) and  $X$ -point (electron pockets) which are acquired mostly by 6*p*-orbitals of Bi and 4*d*-orbital of Y, respectively. A small overlap in energy between the Y-*d* band and Bi-*p* band at Fermi level confirms the semimetallic nature of YBi with no band inversion in line with the experimental report [29] which can also be verified with projected density of state (PDOS) (Fig. 3(c)). This topological trivial nature of YBi is also established by absence of Dirac cone in surface states as shown in Fig. 3(d).







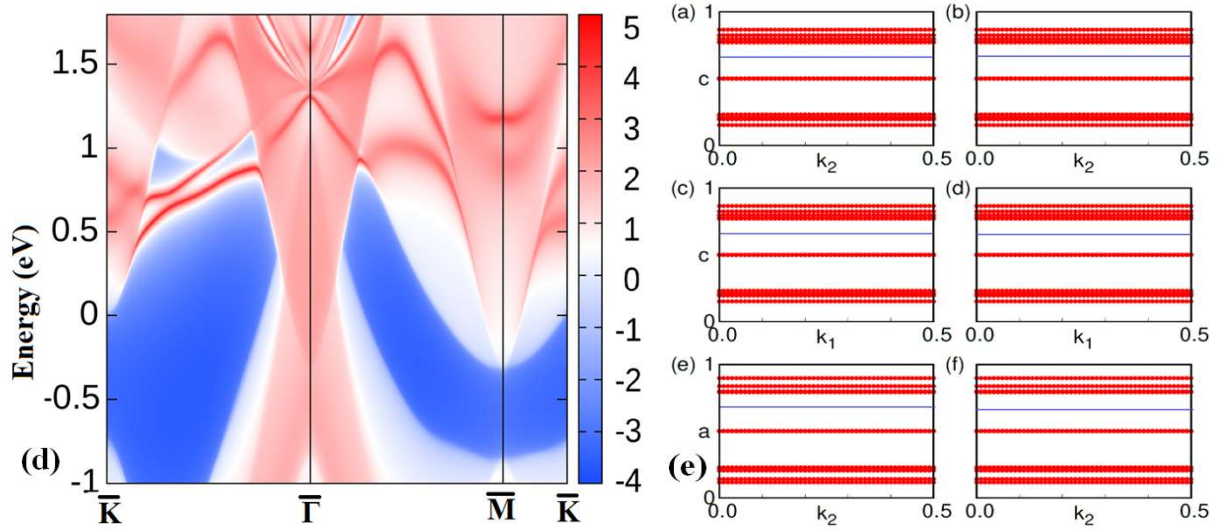


FIG. 3: The band structures of YBi with inclusion of SOC effect using (a) GGA-PBE, (b) HSE06; (c) Projected density of states; (d) The surface state and (e) Wannier charge centers (WCCs) of YBi along (001) plane.

TABLE II: The Parities of all the occupied bands at all the TRIM points in BZ of YBi at ambient pressure.

Band No.	L	L	L	L	$\Gamma$	X	X	X	Total
1	-	-	-	-	-	-	-	-	+
3	-	-	-	-	-	-	-	-	+
5	-	-	-	-	-	-	-	-	+
7	-	-	-	-	+	+	+	+	+
9	+	+	+	+	-	-	-	-	+
11	+	+	+	+	-	-	-	-	+
13	+	+	+	+	-	-	-	-	+
Total	+	+	+	+	+	+	+	+	+

Further, the calculation of the  $Z_2$  topological invariant is performed using product of parities of bands to verify the topological nature of YBi. For three-dimensional materials, having both inversion symmetry as well as TRS, four  $Z_2$  topological invariants can be calculated from the product of parities of occupied bands at TRIM points as suggested by Kane and Mele [15]. These four  $Z_2$  topological invariants i.e.,  $v_0$ ;  $v_1$ ,  $v_2$ ,  $v_3$  can be identified using relations;



$$(-1)^{v_0} = \prod_{m_j=0,1} \delta_{m_1 m_2 m_3} \quad (1)$$

$$(-1)^{v_{i=1,2,3}} = \prod_{m_{j \neq i}=0,1 \& m_i=1} \delta_{m_1 m_2 m_3} \quad (2)$$

where  $\delta$  shows the product of parities of all occupied bands at selected TRIM points,  $v_0$  identifies the topological phase and  $v_1, v_2, v_3$  are used to identify the weak topological nature of YBi. Parities of the all filled energy states at ambient condition are represented in Table II. The first  $Z_2$  invariant ( $v_0$ ) is zero (from equation (1)) which signifies topological trivial nature of YBi.

The  $Z_2$  topological invariants for a bulk material can also be obtained using Wilson loop method also [16]. In this analysis, the  $Z_2$  topological invariants can be calculated with the help of evolution of WCCs [16,40,46] along six TRIM planes i.e.,  $k_x=0, \pi$ ;  $k_y=0, \pi$  and  $k_z=0, \pi$ . The appearance of WCCs has been analysed for YBi using the planes that are spanned by TRIM points. Since the system exhibits TRS, one can place a random reference line across the x-axis, which corresponds to the pumping direction over half of the BZ, to figure out whether it is topologically trivial or non-trivial. The number of crossings of reference line with the evolution lines of WCCs with SOC, provide information about topological nature of YBi. If even number of crossings between reference line and WCCs takes place than it represents the trivial nature and  $Z_2$  topological index have value 0. On the other hand, odd number of reference line and WCCs crossings indicates non-trivial nature with non-zero  $Z_2$  topological index. The non-zero and zero values of  $Z_2$  topological index in planes having  $k_x, k_y, k_z=0$  and  $k_x, k_y, k_z=0.5$ , respectively, represents the strong TI with  $Z_2 = (1;000)$ . It can be observed in Fig. 3(e) that WCCs evolution lines have no crossing with the reference line (blue) in  $k_x, k_y, k_z=0$  and  $k_x, k_y, k_z=0.5$  planes and hence the  $Z_2$  topological invariants are (0;000), which confirms the topological trivial nature of YBi at ambient pressure condition.

## B. Volumetric Pressure

After understating the electronic structure and topological properties of YBi at ambient pressure, we now include the effect of volumetric pressure. We have examined the band structures of the YBi *rocksalt* structure using the HSE06 functional across a pressure range of 1 to 24.5 GPa. We have found no band inversion till 6.4 GPa of volumetric pressure. At 6.5 GPa, a clear band inversion has detected at  $\Gamma$ -point where *d-orbital* of Y and *p-orbital* of Bi gets inverted (Fig. 4(a)). Therefore, we can say that YBi undergoes a topological phase transition at 6.5 GPa. Further, the non-trivial topological nature of YBi can be observed with the help of (001) surface band structure

as shown in Fig. 4(b). It is found that the surface states show single Dirac cone (Fig. 4(b)) at  $\bar{\Gamma}$ -point which corresponds to band inversion at  $\Gamma$ -point in bulk band structure projected SBZ. Further increase in the pressure up to 10 GPa results in another band inversion at  $X$ -point which can be seen in Fig. 5(a). Now, we have an even numbers of band inversions i.e., one at  $\Gamma$ - and other at  $X$ -point, which makes YBi a topologically trivial or weak in nature again. Bulk band structure projection on SBZ also confirms the existence of two Dirac cone at  $\bar{\Gamma}$ - and  $\bar{M}$ -points in surface band structure as shown in Fig. 5(b).

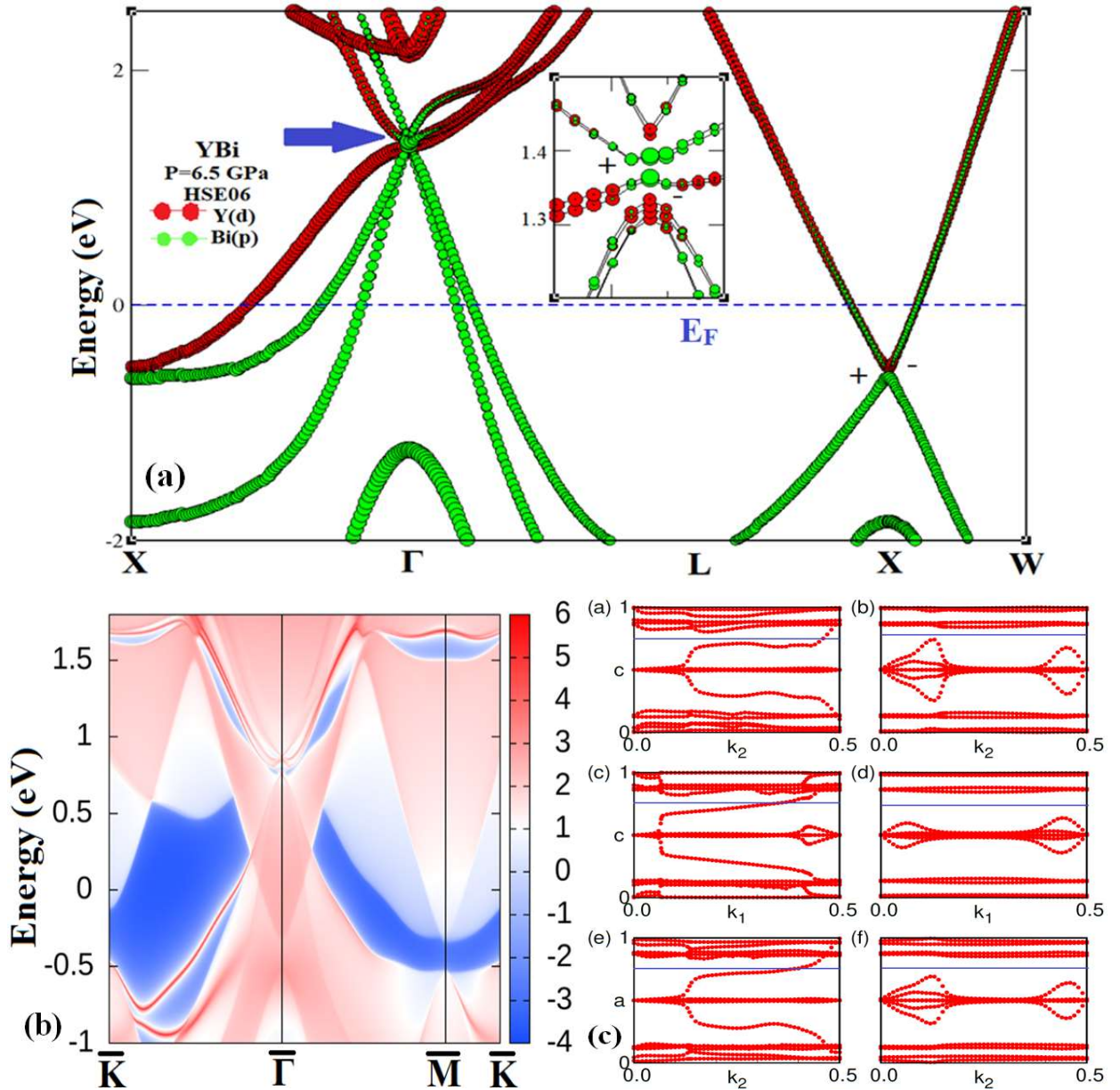


FIG. 4: (a) The band structures of YBi with inclusion of SOC effect using HSE06 functional at 6.5 GPa. (b) The surface state and (c) Wannier charge centers (WCCs) of YBi along (001) plane at 6.5 GPa.

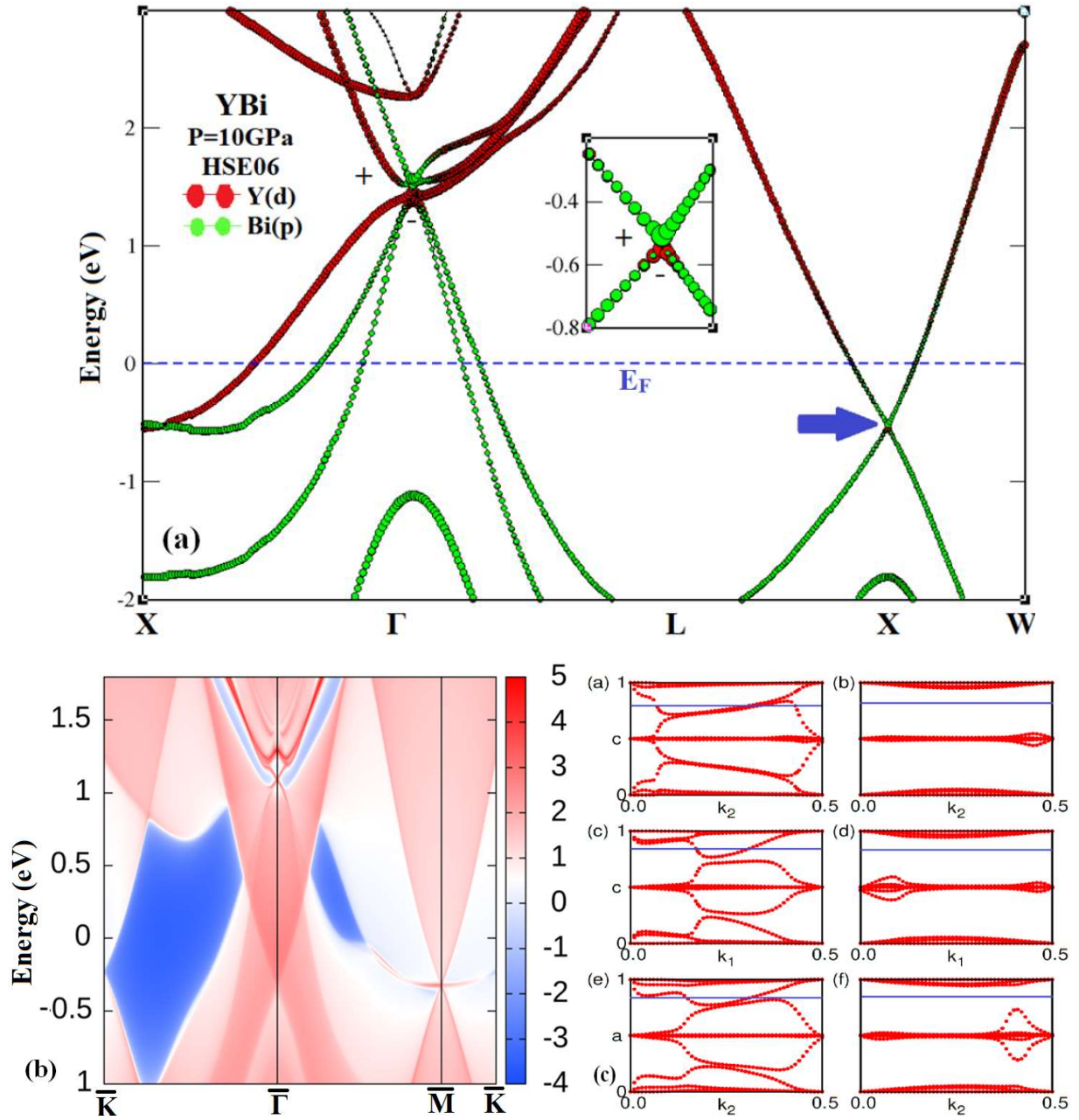


FIG. 5: (a) The band structures of YBi with inclusion of SOC effect using HSE06 functional at 10 GPa. (b) The surface state and (e) wannier charge centers (WCCs) of YBi along (001) plane at 10 GPa.

To understand the band inversion and change in parity under volumetric pressure, we have analysed the band structure evolution at  $\Gamma$ - and  $X$ -points starting from atomic energy levels and then introducing (i) octahedral field (ii) crystal field (iii) spin-orbit interaction (SOI), and (iv) pressure (Fig. 6(a-b)). We have used  $pd$  model for analysis of crystal field splitting in YBi [47-48]. Under the effect of applied volumetric pressure,  $Y-d_{z^2}$  orbital shifts down and  $Bi-p_{x,y}$  orbital shifts up, as expected. At critical values of the pressure 6.5 GPa and 10 GPa, respectively, band

inversions take place at  $\Gamma$ - and  $X$ -points due to these shifts in orbitals as shown in Fig. 6(a-b). Therefore, YBi changes from a normal semimetal to a topological one. Fig. 6(c) depicts the phase diagram with respect to the different exchange-correlation functionals with SOC and pressure. The GGA-PBE and HSE06 shows band inversions at ambient pressure and 6.5 GPa, respectively.

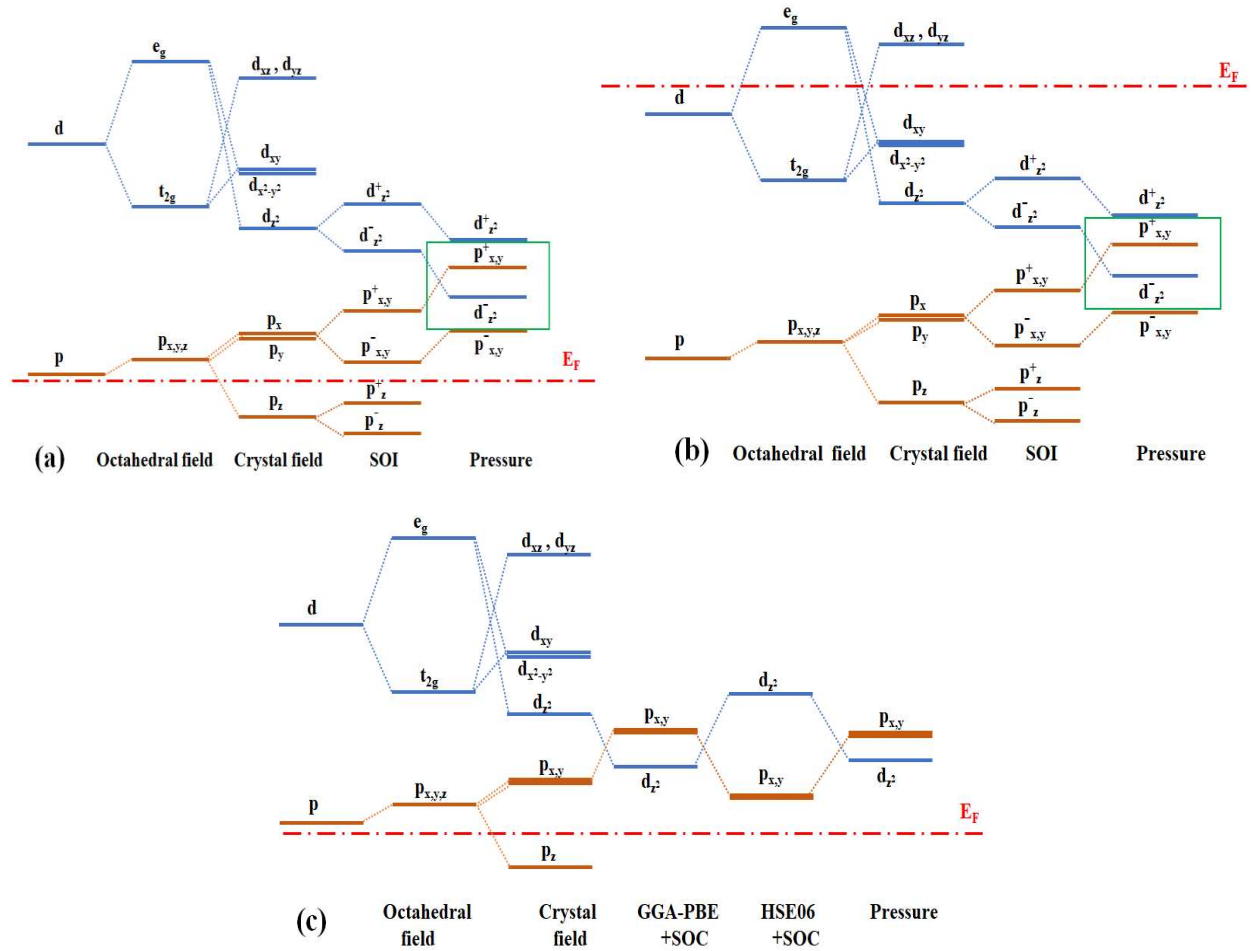


FIG. 6 The band structure evolution of YBi starting from atomic orbitals—octahedral field—crystal field splitting—SOI—applied pressure (a) at  $\Gamma$ -point (b) at  $X$ -point (c) with GGA-PBE and HSE06 functionals at  $\Gamma$ -point.

Parities of all the filled energy states at 6.5 GPa and 10 GPa are depicted in Tables III and IV, respectively. At first band inversion i.e., at 6.5 GPa, the parity of the highest occupied band at  $\Gamma$  is changed to positive, which switched the overall parity of YBi to negative as shown in Table III.

TABLE III: The Parities of all the occupied bands at all the TRIM points in BZ of YBi at 6.5 GPa.

Band No.	L	L	L	L	$\Gamma$	X	X	X	Total
1	-	-	-	-	-	-	-	-	+
3	-	-	-	-	-	-	-	-	+
5	-	-	-	-	-	-	-	-	+
7	-	-	-	-	+	+	+	+	+
9	+	+	+	+	-	-	-	-	+
11	+	+	+	+	-	-	-	-	+
13	+	+	+	+	+	-	-	-	-
Total	+	+	+	+	-	+	+	+	-

TABLE IV: The Parities of all the occupied bands at all the TRIM points in BZ of YBi at 10 GPa.

Band No.	L	L	L	L	$\Gamma$	X	X	X	Total
1	-	-	-	-	-	-	-	-	+
3	-	-	-	-	-	-	-	-	+
5	-	-	-	-	-	-	-	-	+
7	-	-	-	-	+	+	+	+	+
9	+	+	+	+	-	-	-	-	+
11	+	+	+	+	-	-	-	-	+
13	+	+	+	+	+	+	+	+	+
Total	+	+	+	+	-	-	-	-	+

Now, the first  $Z_2$  topological invariant ( $v_0$ ) becomes 1 using equation (1), which verifies the non-trivial nature of YBi. At second inversion (10 GPa), the parity of three  $X$ -points switches from positive to negative and first  $Z_2$  topological invariant ( $v_0$ ) changes 0 from 1 (equation (1)). Now, the YBi becomes either a weak topological insulator or topologically trivial insulator. To verify this, we have calculated the other three topological invariants ( $v_1, v_2, v_3$ ) using equation (2). Table III shows that parities at three  $X$ -points and four  $L$ -points are the same, which indicates that the other three topological invariants are (0, 0, 0). So, it can be concluded that at 10 GPa, YBi shows an even number of band inversions and is topologically trivial in nature.

At 6.5 GPa volumetric pressure, the WCCs evolution lines in Fig. 4(c) cuts odd numbers of time to the reference line (blue) in  $k_x, k_y, k_z=0$  and  $k_x, k_y, k_z=0.5$  planes which confirms the topological

non-trivial nature and  $Z_2$  indices are (1;000). Whereas, in Fig. 5(c), even number of crossings between WCCs evolution lines and reference line (blue) can be seen for 10 GPa pressure, which again confirms the transition from non-trivial to trivial nature with  $Z_2$  topological indices (0;000).

### C. Epitaxial Strain

The coherently strained films on lattice mismatched substrates can influence the electronic structure of materials by means of epitaxial strain. The implementation of molecular beam epitaxy method had successfully shown the presence of epitaxial strain induced during the growth process of rare-earth pnictides on III-V semiconductors [49]. The III-V semiconductors [26,50] have attained compressive epitaxial strain of up to 3%, and a similar behaviour can be expected from rocksalt rare-earth monpnictides e.g., for LaSb and SnTe, respectively, 1.6% epitaxial and 1.1% out-of-plane tensile strain have been reported previously [26,27]. This induced strain may influence the charge transfer at the interface which can further affect the carrier compensation [26].

Now, in the following section, we will discuss about the topological phase transition in YBi when it is subjected to epitaxial strain. The space group symmetry of YBi is changes from  $Fm\bar{3}m$  to  $I4/mmm$  (Fig.1(b)) with epitaxial strain but the inversion symmetry remains preserved. Here we have demonstrated that the epitaxial strain pushes the band structure of YBi from topological trivial to non-trivial nature and thus creating an inevitable Dirac node at  $\Gamma$ -point.

The electronic band structure of YBi under compressive epitaxial strain is obtained along  $X$ - $\Gamma$ - $L$ - $X$ - $W$  k-path as shown in Fig. 7(a). With epitaxial strain, the Y- $d$  band shifted towards the Bi- $p$  bands at  $\Gamma$  and  $X$  points which results in reduction of the total volume of the cell. At 3% strain, we find a band inversion at  $\Gamma$ -point; but still at  $X$  point, the Y- $d$  and Bi- $p$  bands continues to avoid band crossing. The band inversion at  $\Gamma$ -point can be seen in Fig. 7(a) and the inverted contribution of  $d$ -orbital of Y and  $p$ -orbital of Bi is shown in Fig. 7(a) (inset). To further verify the topological nontrivial nature of YBi under epitaxial strain, we computed the surface band structure along (001) plane. Since epitaxial strain causes the bulk band inversion in YBi only at the  $\Gamma$ -point, we found a single Dirac cone to emerge at the  $\bar{\Gamma}$ -point on (001) plane. The surface band structure along  $M$ - $\Gamma$ - $M$  path of (001) plane is shown in Fig. 7(b). Unlike volumetric pressure, no Dirac cone is observed at  $X$ -point for epitaxial strain.



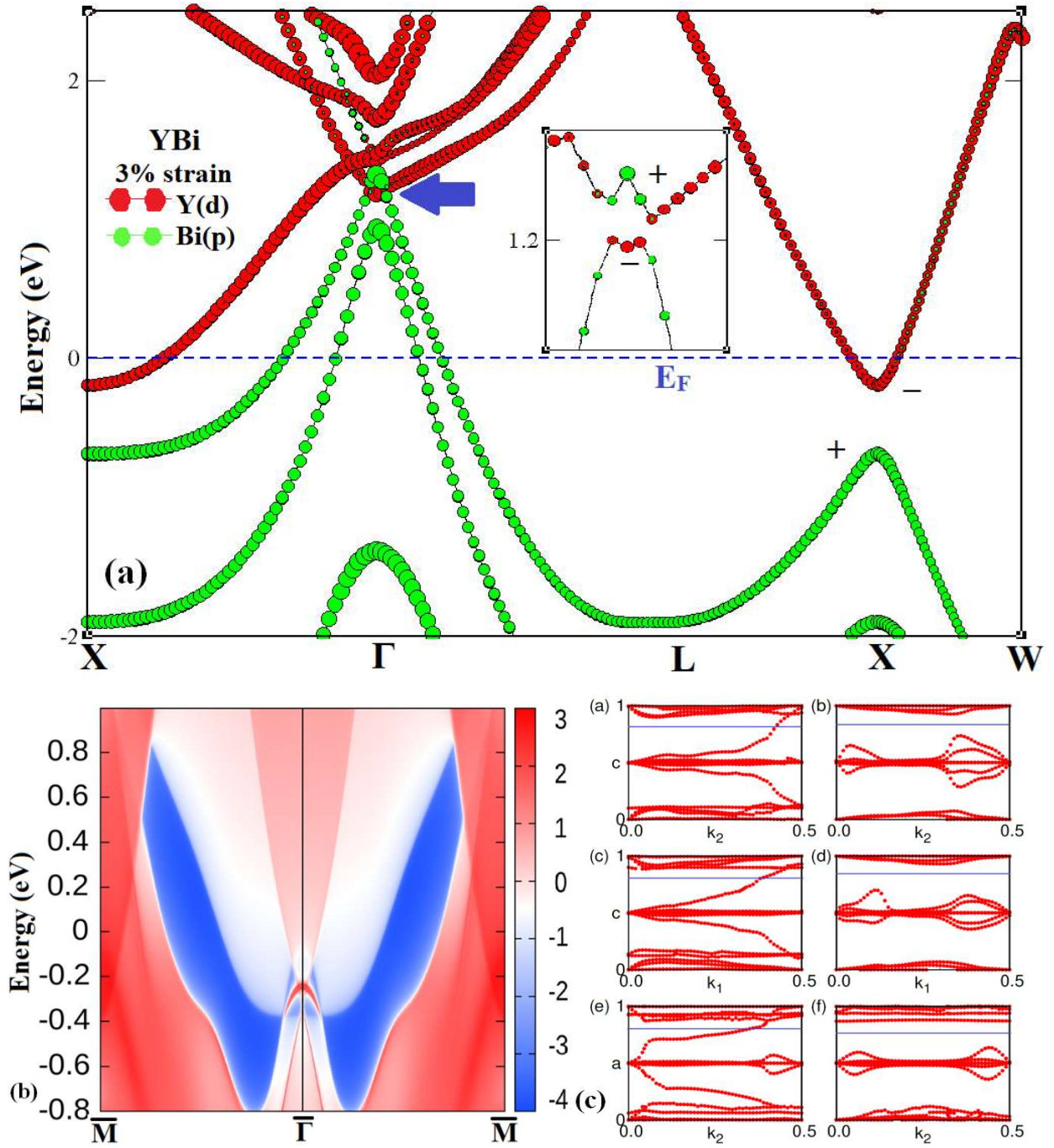


FIG. 7: (a) The band structures of YBi with inclusion of SOC effect using HSE06 functional at 3% epitaxial strain. (b) The surface state (SS) and (e) Wannier charge centers (WCCs) of YBi along (001) plane at 3% epitaxial strain.

In order to establish the occurrence of bulk band inversion under epitaxial strain and its connection with nontrivial topology in YBi, we determined the  $Z_2$  topological invariant. Table V contains the parities of various occupied bands at TRIM points at an epitaxial strain of 3%. It can be observed that the parity at the  $\Gamma$ -point undergoes exchanged, whereas the parity at the  $X$ -points stays unaltered with respect to the ambient conditions (Table II). The  $Z_2$  invariant change from 0 to 1 as

a result of a change in parity at the  $\Gamma$ -point under epitaxial strain which is evidence of the topological non-trivial character in YBi. Moreover, the single crossing in WCCs evolution lines and reference line (blue) (Fig. 7(c)) also verifies the strong topological phase in YBi with  $Z_2=(1;000)$ .

TABLE V: The Parities of all the occupied bands at all the TRIM points in BZ of YBi at 3% epitaxial strain.

Band No.	L	L	L	L	$\Gamma$	X	X	X	Total
1	-	-	-	-	-	-	-	-	+
3	-	-	-	-	-	-	-	-	+
5	-	-	-	-	-	-	-	-	+
7	-	-	-	-	+	+	+	+	+
9	+	+	+	+	-	-	-	-	+
11	+	+	+	+	-	-	-	-	+
13	+	+	+	+	+	-	-	-	-
Total	+	+	+	+	-	+	+	+	-

We have shown that YBi show topological phase transition under volumetric pressure as well as epitaxial strain. In the volumetric pressure range of 6.5 GPa to 10 GPa, YBi has non-trivial topological character and an epitaxial strain of 3% transform it from trivial to non-trivial. We have calculated the  $Z_2$  topological invariant  $v_0$  at different values of applied volumetric pressure as well as epitaxial strain. Fig.8(a-b) is an illustration that how the value of the first  $Z_2$  topological index varies as a function of volumetric pressure and epitaxial strain.

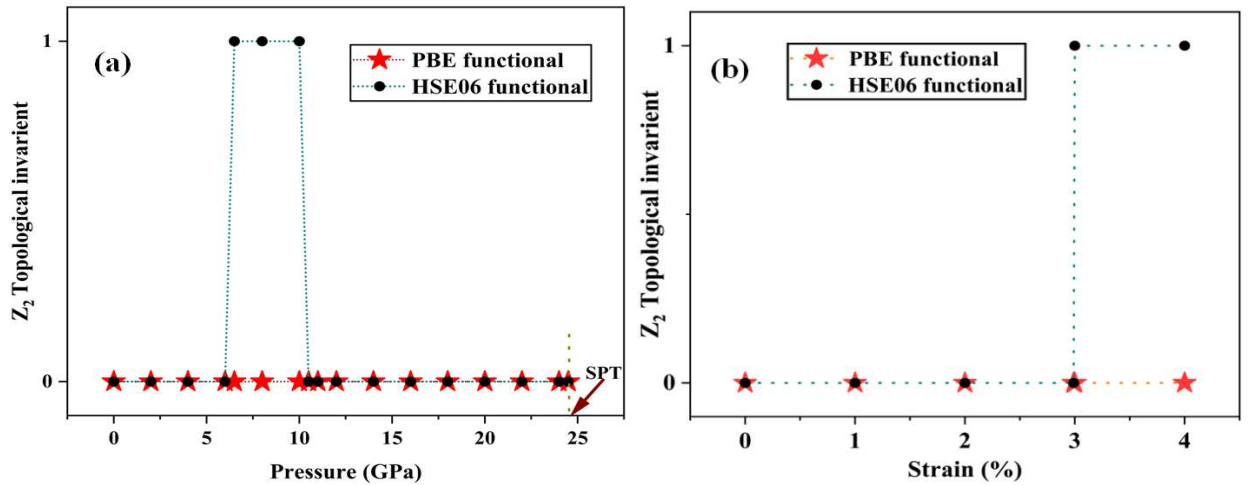


FIG. 8: The variation of First  $Z_2$  topological index ( $v_0$ ) with (a) applied volumetric pressure (b) applied epitaxial strain.

It is important to note that the rise in pressure leads to an increase in the overlap between the valence band and the conduction band of YBi, which in turn results in an increase in the carrier concentration. Since XMR counts on the carrier concentration as well as the mobility of the charge carriers, the enhancement of mobility with pressure is an effective way to analyse evolution of XMR as a function of pressure. Our results regarding topological phase transitions in YBi with 6.5 GPa of volumetric pressure can be a vital platform explore the relation between XMR and pressure. The occurrence of non-trivial phase in YBi with a relatively small epitaxial strain, which a thin film geometry can naturally has, might make it ideal candidate to probe inter-relationship between XMR and non-trivial topology.

### III. SUMMARY

We have used hybrid density functional theory to investigate the structural, electronic, and topological properties of XMR material YBi at ambient and elevated volumetric pressure and epitaxial strain. The structural and dynamical stabilities of the system have been ascertained and a structural phase transition has been predicted at 24.5 GPa. The GGA-PBE functional has overestimated the bands overlap near the Fermi level and an even number of band inversions have been observed. The hybrid functional HSE06 has accurately predicted the topologically trivial semimetallic nature of YBi which agrees with existing experimental report. The YBi has undergone a topological phase transition at 6.5 GPa of volumetric pressure and 3% of epitaxial strain. The non-zero values of  $Z_2$  topological index, calculated with the help of product of parities of all the occupied bands at TRIM points and evolution of WCCs, have confirmed these topological phase transitions. The *d-orbital* of Y and *p-orbital* of Bi have mainly contributed near the Fermi level and take part in topological band inversion of YBi. The existence of single Dirac cone on plane (001) has confirmed the non-trivial nature of YBi. A rise in volumetric pressure (10 GPa) has make it trivial again which has been verified with even number of Dirac cone on (001) plane. The  $Z_2$  topological index has switched from 1 to 0 at 10 GPa, which has also observed in evolution of WCCs. A small epitaxial strain of 3%, which can arise due to lattice-substrate mismatch during coherent growth of thin films of YBi, can be an opportunity to interrelate non-trivial topology, electron-hole compensation and XMR in rare-earth mononictides.

### ACKNOWLEDGEMENTS

One of the authors (Ramesh Kumar) would like to thank Council of Scientific and Industrial Research (CSIR), Delhi, for financial support. All the authors acknowledge the National

Supercomputing Mission (NSM) for providing computing resources of 'PARAM SEVA' at IIT, Hyderabad, which is implemented by C-DAC and supported by the Ministry of Electronics and Information Technology (MeitY) and Department of Science and Technology (DST), Government of India.

## References

- [1] C. L. Kane and E. J. Mele, Phys. Rev. Lett. **95**, 146802 (2005).
- [2] L. Fu, C. L. Kane, and E. J. Mele, Phys. Rev. Lett. **98**, 106803 (2007).
- [3] J. E. Moore, Nature **464**, 194 (2010).
- [4] M. Z. Hasan and C. L. Kane, Rev. Mod. Phys. **82**, 3045 (2010).
- [5] M. Z. Hasan and J. E. Moore, Annu. Rev. Condens. Matter Phys. **2**, 55 (2011).
- [6] M. Z. Hasan, S.-Y. Xu, M. Neupane, S. Roche, and S. O. Valenzuela in Topological Insulators: Fundamentals and Perspectives, edited by F. Ortmann, S. Roche, and S. O. Valenzuela, (Wiley-VCH Verlag GmbH & Co. KGaA, 2015).
- [7] X. L. Qi and S. C. Zhang, Rev. Mod. Phys. **83**, 1057 (2011).
- [8] X. Wan, A. M. Turner, A. Vishwanath, and S. Y. Savrasov, Phys. Rev. B **83**, 205101 (2011).
- [9] H. Weng, C. Fang, Z. Fang, B. Andrei Bernevig, and X. Dai, Phys. Rev. X **5**, 011029 (2015).
- [10] Z. K. Liu, L. X. Yang, Y. Sun, T. Zhang, H. Peng, H. F. Yang, C. Chen, Y. Zhang, Y. F. Guo, D. Prabhakaran, M. Schmidt, Z. Hussain, S.-K. Mo, C. Felser, B. Yan and Y. L. Chen, Nat. Mater. **15**, 27 (2016).
- [11] M. Hirayama, R. Okugawa, and S. Murakami, J. Phys. Soc. Jpn. **87**, 041002 (2018).
- [12] Z. K. Liu, B. Zhou, Y. Zhang, Z. J. Wang, H. M. Weng, D. Prabhakaran, S.-K. Mo, Z. X. Shen, Z. Fang, X. Dai, Z. Hussain, Y. L. Chen, Science **343**, 864 (2014).
- [13] M. Zobel, R. B. Neder, and S. A. J. Kimber, Science **347**, 292 (2015).
- [14] S.-Y. Xu, I. Belopolski, N. Alidoust, M. Neupane, G. Bian, C. Zhang, R. Sankar, G. Chang, Z. Yuan, C.-C. Lee, S.-M. Huang, H. Zheng, J. Ma, D. S. Sanchez, B. K. Wang, A. Bansil, F. Chou, P. P. Shibayev, H. Lin, S. Jia, M. Z. Hasan, Science **349**, 613 (2015).
- [15] L. Fu and C. L. Kane, Phys. Rev. B **76**, 045302 (2007).
- [16] R. Yu, X. L. Qi, A. Bernevig, Z. Fang, and X. Dai, Phys. Rev. B **84**, 075119 (2011).
- [17] J. Nayak, S.-C. Wu, N. Kumar, C. Shekhar, S. Singh, J. Fink, E. E.D. Rienks, G. H. Fecher, S. S.P. Parkin, B. Yan and C. Felser, Nat. Commun. **8**, 13942 (2017).
- [18] X. H. Niu, D. F. Xu, Y. H. Bai, Q. Song, X. P. Shen, B. P. Xie, Z. Sun, Y. B. Huang, D. C. Peets, and D. L. Feng, Phys. Rev. B **94**, 165163 (2016).

- [19] X. Duan, F. Wu, J. Chen, P. Zhang, Y. Liu, H. Yuan, and C. Cao, *Commun. Phys.* **1**, 71 (2018).
- [20] S. Khalid, F. P. Sabino, and A. Janotti, *Phys. Rev. B* **98**, 220102 (2018).
- [21] P. J. Guo, H. C. Yang, K. Liu, and Z. Y. Lu, *Phys. Rev. B* **96**, 081112(2017).
- [22] P. Barone, T. Rauch, D. Di Sante, J. Henk, I. Mertig, and S. Picozzi, *Phys. Rev. B* **88**, 045207 (2013).
- [23] P. Wadhwa, S. Kumar, A. Shukla, and R. Kumar, *J. Phys.: Condens. Matter* **31**, 335401 (2019).
- [24] YZhou, P. Lu, Y. Du, X. Zhu, G. Zhang, R. Zhang, D. Shao, X. Chen, X. Wang, M. Tian, J. Sun, X. Wan, Z. Yang, W. Yang, Y. Zhang, and D. Xing, *Phys. Rev. Lett.* **117**, 146402 (2016).
- [25] M. Singh, R. Kumar, and R. K. Bibiyan, *Eur. Phys. J. Plus* **137**, 633 (2022).
- [26] S. Khalid and A. Janotti, *Phys. Rev. B* **102**, 035151 (2020).
- [27] S. Fragkos, R. Sant, C. Alvarez, E. Golias, J. M.-Velasco, P. Tsipas, D. Tsoutsou, S. A.-Giamini, E. Xenogiannopoulou, H. Okuno, G. Renaud, O. Rader, and A. Dimoulas, *Phys. Rev. Mater.* **3**, 104201 (2019).
- [28] O. Pavlosiuk, P. Swatek, D. Kaczorowski, and P. Wiśniewski, *Phys. Rev. B* **97**, 235132 (2018).
- [29] S. Xiao, Y. Li, Y. Li, X. Yang, S. Zhang, W. Liu, X. Wu, B. Li, M. Arita, K. Shimada, Y. Shi, and S. He, *Phys. Rev. B* **103**, 115119 (2021).
- [30] C. Q. Xu, B. Li, M. R. van Delft, W. H. Jiao, W. Zhou, B. Qian, N. D. Zhigadlo, D. Qian, R. Sankar, N. E. Hussey, and X. Xu, *Phys. Rev. B* **99**, 024110 (2019).
- [31] P. Hohenberg and W. Kohn, *Phys. Rev.* **136**, B864 (1964).
- [32] W. Kohn and L. J. Sham, *Phys. Rev.* **140**, A1133 (1965).
- [33] G. Kresse and D. Joubert, *Phys. Rev. B* **59**, 1758 (1999).
- [34] G. Kresse and J. Furthmuller, *Phys. Rev. B* **54**, 11169 (1996).
- [35] J. P. Perdew, K. Burke, and M. Ernzerhof, *Phys. Rev. Lett.* **77**, 3865 (1996).
- [36] J. Heyd, G. E. Scuseria, and M. Ernzerhof, *J. Chem. Phys.*, 8207 (2003).
- [37] J. Heyd, G. E. Scuseria, and M. Ernzerhof, *J. Chem. Phys.* **124**, 219906 (2006).
- [38] A. Togo and I. Tanaka, *Scr. Mater.* **108**, 1 (2015).
- [39] N. Marzari, A. A. Mostofi, J. R. Yates, I. Souza, and D. Vanderbilt, *Rev Mod Phys* **84**, 1419 (2012).
- [40] Q. Wu, S. Zhang, H. F. Song, M. Troyer, and A. A. Soluyanov, *Comput. Phys. Commun.* **224**, 405 (2018).
- [41] K. A. Gschneidner, J. C. G. Bünzli, and V. K. Pecharsky, *Handbook on the Physics and Chemistry of Rare Earths (Elsevier, 2006)*, Vol. 36.

- [42] M. Narimani and Z. Nourbakhsh, J. Phys. Chem. Solids **145**, 109537 (2020).
- [43] S. Azzi, A. Zaoui, and M. Ferhat, Phys. Scr. **88**, 055601 (2013).
- [44] A. K. Ahirwar, M. Aynyas, Y. S. Panwar, and S. P. Sanyal, Adv. Mat. Res. **1141**, 39 (2016).
- [45] N. Acharya and S. P. Sanyal, Solid State Commun. **266**, 39 (2017).
- [46] A. A. Soluyanov and D. Vanderbilt, Phys Rev B **83**, 235401 (2011).
- [47] D. Pasquier and O. V. Yazyev, 2d Mater. **6**, (2019).
- [48] Y. Sun, Q. Z. Wang, S. C. Wu, C. Felser, C. X. Liu, and B. Yan, Phys Rev B **93**, 205303 (2016).
- [49] S. Chatterjee, S. Khalid, H. S. Inbar, T. Guo, Y.-H. Chang, E. Young, A. V. Fedorov, D. Read, A. Janotti, and C. J. Palmstrøm, Sci. Adv. **7**, eabe8971 (2020).
- [50] C. J. K. Richardson and M. L. Lee, MRS Bulletin **41**, 193 (2016).



# Traffic Prediction Model Using Machine Learning in Intelligent Transportation Systems

<sup>1</sup>Abhilasha Sharma

Department of Software Engineering  
Delhi Technological University  
Delhi, India  
abhi16.sharma@gmail.com

<sup>2</sup>Prabhat Ranjan

Department of Software Engineering  
Delhi Technological University  
Delhi, India  
prabhat\_2k22dsc10@dtu.ac.in

**Abstract** — The term Traffic Environment refers to everything that has the potential to disrupt the flow of vehicular traffic on the road, but not particulates to accidents, traffic signals, rallies and even for the maintenance of roads, which may lead to backups. If an individual possesses prior knowledge that is similar to the above-mentioned factors, along with an understanding of other everyday variables that may influence traffic, they can make informed decisions as a driver or passenger. Advanced transport systems, tourist database systems, and traffic management systems now have much better traffic predictions because of the adoption of intelligent transportation systems. Therefore, the continuous improvement and implementation of these systems will continue to enhance traffic prediction accuracy. The goal of this study is to use modern communication technologies and to propose a methodology based on machine learning for improving transportation safety, mobility and productivity. According to the research, the suggested method (SVR) performs better than the competing approaches in terms of performance on all assessment metrics, regardless of the dataset.

**Keywords:** Intelligent Transportation Systems (ITS); Support Vector Machine (SVR); Traffic Management & Prediction; Vehicle Detection.

## I. INTRODUCTION

Precise and timely traffic flow data is essential for many businesses, governments, and individuals. Riders and drivers are aided in making more informed decisions about their routes, which helps decrease delays, boost the effectiveness of traffic operations, and cut down on emissions. Intelligent Transportation Systems (ITSs) provide more precise forecasting of traffic flows as they are developed and implemented. It is addressed since it is so important to the functioning of high-tech transport management systems, public transport networks, and tourist info networks [1]. ITS is a promising emerging technology with the potential to enhance transportation security, flow management, and traveller convenience soon. The idea behind such a system is that cars would be permanently connected to an Internet of Vehicles (IoV), giving drivers a comprehensive picture of traffic conditions [2]. Vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications are the backbones of the ITS, as illustrated in Figure 1 [3].

Improvements in hardware, software, and communication channels within Information and Communication Technologies (ICT) have made it more feasible to create robust, smart transportation networks. Sustainable, integrated, safe, and

responsive ITS are the foundation for a more pleasant and secure journey, and they can only be realized if ICT is completely included in the transportation infrastructure.

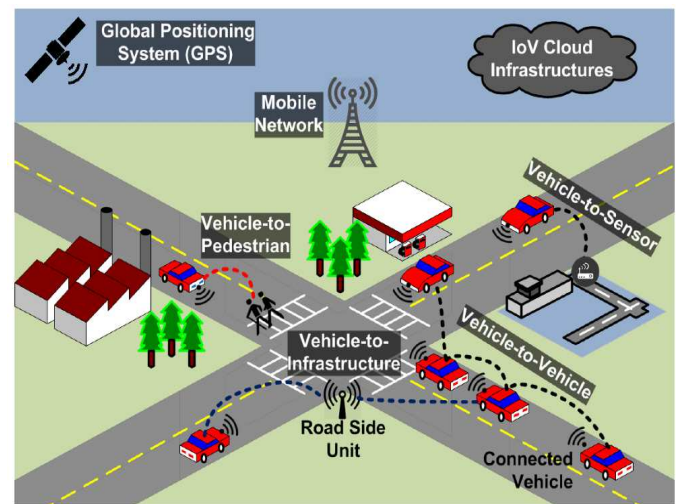


Fig. 1. Intelligent Transportation System

The goals of ITS, which include improved accessibility and mobility, more environmentally friendly operations, and a boost to the economy, will be significantly aided by the implementation of these guiding principles [4]. Accessing, collecting, and processing reliable data from the environment is crucial to the effectiveness of ITS. Two major groups may be identified among sensing platforms. Intra-vehicle sensing platforms are the first kind and are used to gather information about a vehicle's internal state. Urban sensing systems, the second kind, are put to use to gather data on traffic patterns. When cars talk to cars, or cars talk to infrastructure, sensors play a crucial role in gathering information for analysis. Transportation management systems get this information and use it to make choices and take action. Fuel pricing, carbon dioxide emissions, traffic congestion, and road quality are just a few of the problems that may be alleviated by more sophisticated and smart ITS systems [5].

**Application of ITS:** Recent developments in ICT, as well as the ever-expanding use of AI in a variety of contexts, have prepared the way for the development of ITS. The goal of this development is to cater the cutting-edge solutions for managing traffic, congestion, and other issues associated with roadways and transportation systems [6]. It primarily includes

AI-powered wireless, sensor, and computing technologies. Application areas where it is employed are categorized in Figure 2 [8]. Some of the most important functions of a smart city are handled by ITS, such as traffic management and citizen security [7]. These services help keep traffic flowing smoothly by doing things like alerting drivers to impending danger, automatically enforcing speed limits, preventing accidents, advising drivers on where to park, reporting the weather, and reporting when bridges need to be moved or removed [8]. Human resource management and resource allocation are two areas where it falls short. To combat the ever-increasing traffic congestion issue, cities must prioritize the scheduling and reallocating their traffic police force following changes in traffic density.

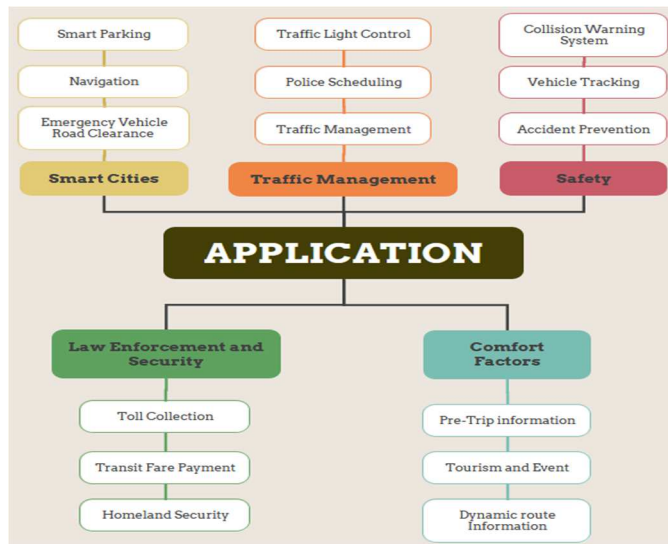


Fig. 2. ITS Applications

### A. Problem Formulation

Traffic flow data pretreatment is critical for missing values, noise removal, and data reduction. Vehicle detection counts vehicles in each image. Choose traffic flow data features using the CNN algorithm model. After separating unique traffic data for a train set and test set, it can begin to analyze the data. Find the sweet spot for the SVR's settings. Use the training data to educate a support vector machine (SVM) prediction model and then get the model's prediction. Check if the present parameters meet maximum accuracy standards. When conditions are met, parameter optimization might terminate. Because the current settings are optimal, the SVR model ran successfully. When testing, the warning level is anticipated and then matched with the training data, the fulfilled, and the final output to predict the final output.

### B. Research Objectives

1. To increase transportation safety, mobility, and productivity through new communications technology using a machine learning approach.
2. To achieve lower-error support, a vector regression machine was employed to accomplish the precise classification error rate.

3. Improve Mobility by boosting system efficiency and elevating individual mobility for all members of society, from novice to seasoned motorists.
4. Reduce transportation-related environmental impacts (including their contributions to climate change and air pollution) through improved traffic management.

## II. RELATED WORK

The section contains a literature survey related to the area of traffic prediction using machine learning in ITS.

**Kuok et al., (2021) [9]** resolve the two issues by offering three solutions: (i) a computer program that simulates traffic conditions in real-time, (ii) A neural network based on pheromones for predicting traffic and rerouting it, and (iii) Weighted Missing Data Imputation is an approach to missing data that makes use of the weighted historical data technique (WEMDI). Google Maps' rerouting system served as a reference point for the advancement of the traffic simulation model. The rerouting mechanism gets compared either by using WEMDI or not for evaluation under various levels of missing data and checking its effectiveness. The findings demonstrated a strong connection among traffic-simulation methods, Google Maps and its allied traffic parameters improved by 38%-44% thanks to the WEMDI's integrated system, when missing data amounts to 50%, the performance of the augmented rerouting system improved by around 19.39% as compared to the standard rerouting system. The WEMDI system was equally resilient in its routing to additional sites, demonstrating the same level of excellent performance.

**Cong and Pei, (2021). [10]** create an enhanced Support Vector Machine (SVR) model for predicting traffic flows soon by optimizing Support Vector Machine (SVM) parameters for use in short-term forecasting. Data show that SVR has the lowest classification error rate (3.22%). Results show that predicting morning and evening peak times reduces MAPE of SVR by 19.93% and 42.87 %, respectively, while reducing RMSE by 29.71 % and 47.22 %, also depending on the time of day. The enhanced method has been shown to increase the reliability of traffic flow predictions and speed up the time it takes to find the best possible parameters for SVR. The counting accuracy has been vastly enhanced by the goal-tracking pedestrian counting approach suggested in this study. After implementing a more effective framework for counting pedestrians and expanding how HOG characteristics are calculated (for example, by allowing users to choose which neighbourhoods to use), this research provides a step toward improving pedestrian safety.

**Chen et al., (2020) [11]** proposed a method for traffic flow detection and monitoring using an automobile identification algorithm based on the YOLOv3 model, which was trained on large amounts of traffic data. They optimized the model for peripheral devices and improved the DeepSORT algorithm's performance for tracking multiple vehicles. They also introduced a real-time vehicle monitoring counter to detect traffic flow. The authors demonstrated the accuracy and efficiency of their approach and deployed it on the Jetson TX2

edge device with a processing speed of 37.9 FPS. Overall, their method shows promise for effective traffic flow detection and monitoring.

**Wang et al., (2020) [12]** introduced a revolutionary Rear-end Collision Prediction Mechanism (RCPM), which uses deep learning to construct a Convolutional Neural Network (CNN) model. For reference to the issue of a class discrepancy, RCPM smooths and expands the dataset following genetic theory. When the author educates the author's convolutional neural network model, it uses the training and testing sets that were created from the preprocessed dataset. The experimental findings demonstrate that RCPM considerably enhances working in forecasting rear-end crashes compared to the Berkeley, Honda, and multi-layer perceptron neural network-based methods.

**Zhou et al., (2020) [13]** provide a novel Bayesian framework for traffic forecasting based on neural networks, termed Variational Graph Recurrent Attention (VGRAN). Dynamic graph convolution algorithms can be utilized to capture a time series of road-sensor data and learn latent variables related to sensor representations and traffic sequences. This probabilistic approach is more versatile in generating models as it considers the unpredictability of sensor features and temporal traffic correlations. Additionally, it allows for precise modelling of implied backlogs in traffic information, which are often unstructured, geographically connected, and dependent on a variety of time scales. After conducting thorough tests on two real traffic datasets, it was found that the VGRAN proposed in this study performed comparably to the state-of-the-art approach while accurately representing the inherent uncertainty of the predicted outcomes.

**Neetesh et al., (2020) [14]** suggested that the proposed Dynamic and Intelligent Traffic Light Control System (DITLCS) aims to improve traffic flow by utilizing real-time traffic data to adjust the duration of green and red lights. The system operates in three modes: Fair Mode (FM), Priority Mode (PM), and Emergency Mode (EM). The FM treats vehicles equally, the PM treats vehicles differently based on category, and the EM gives the highest priority to emergency vehicles. The system uses a deep reinforcement learning-based model that alternates between three states for the traffic lights (Red, Yellow, and Green) and a fuzzy INS selects from three operational modes (FM, PM, and EM) based on the available data. The Simulation of Urban Mobility (SUMO) was used by the author to simulate the street layout of Gwalior, India, and he discovered that DITLCS outscored other cutting-edge algorithms on a variety of performance criteria.

**Shengnan Guo et al., [15]** presented the AST-GCN (Attention-Based Spatial-Temporal Graph Convolutional Networks) model for traffic flow forecasting. By utilizing graph convolutional networks and attention mechanisms, the model effectively captures the spatial and temporal dependencies in traffic data. The GCN component captures the relationships among traffic sensors represented as a graph, enabling spatial

feature learning. The attention mechanisms are applied to capture the temporal dependencies, allowing the model to focus on important patterns and long-term trends in the data. Experimental evaluations on real-world traffic datasets demonstrate that the AST-GCN model outperforms existing methods, achieving higher accuracy and better prediction quality. The proposed model's ability to incorporate spatial and temporal information offers promising potential for improving traffic flow forecasting, which can greatly benefit transportation management and planning applications. The findings of this study contribute to the advancement of traffic prediction techniques and provide valuable insights for future research in this field.

**Li et al., (2019) [16]** create a model that uses a combination of different data types to forecast the average speed in space. To begin, stacked autoencoders will extract and train on the chronological and spatial characteristics, which are the natural input. The data's most salient traits are retrieved next, and then combined with them. Now that authors know what the relationships are, they can build prediction models. As a result, the prediction model may take into account both geographical and temporal correlation as well as the relationship between different kinds of information. It seems that the context of the text is missing, but based on the provided sentence, it appears that the authors have applied several machine learning models to real-world data and compared their performance. The models include artificial neural networks, support vector regression trees, and k-nearest neighbours. Additionally, the authors have suggested a deep feature-level fusion technique and compared its performance against the standard data-level fusion approach. They have found that applying both a deep feature fusion model and a support vector regression method together may provide the best results. The findings suggest that the suggested deep feature fusion model has the potential to improve upon existing methods.

**Ferdowsi et al., (2019) [17]** addressed the latency and dependability issues plaguing ITSs, and a new edge analytics architecture is presented. This process worked on data at the vehicle level or with the help of roadside smart sensors. It uses deep-learning strategies for mobile sensing, which is made possible by the increased capabilities of passengers' mobile devices and intra-vehicular processors. Different types of data, autonomous control, vehicle platooning, and cyber-physical security are all explored as they apply to the difficulties of implementing ITS mobile edge analytics. Then, deep-learning solutions are shown. Deep-learning techniques provide ITS devices with sophisticated computer vision and signal processing features for edge analytics. As shown by these early findings, the new edge analytics architecture, when combined with the strength of deep-learning algorithms, creates a transportation system that is dependable, secure, and smart. Table 1 gives the comparison of literature review for several researchers who have employed various methods and published their findings.

TABLE I. FINDINGS OF THE REVIEWED SOURCES

Author	Model	Datasets	Outcome	Implication
<b>Kuok et al., (2021) [9]</b>	WEMDI	PEMSD4	A high association between Google Maps' traffic simulation model and the WEMDI integrated system, which improved traffic factors by 38% to 44% compared to re-routing systems. It shows a 19.39% improvement compared to the original one, with missing data levels around 50%.	The WEMDI integrated system can be used to improve traffic factors by 38% to 44% compared to re-routing systems. This suggests that the WEMDI system can be a valuable tool for improving traffic management.
<b>Cong and Pei, (2021) [10]</b>	SVM	PEMSD8	Results show that the suggested technique significantly outperforms the SVR approach, with the final $err_{all}$ error and final $mse_{all}$ error of 0.7898 times and 0.6519 times, respectively, as compared to the SVR error.	The SVM approach can significantly outperform the SVR approach in terms of accuracy and other metrics. This suggests that the SVM approach may be a better choice for traffic forecasting.
<b>Chen et al., (2020) [11]</b>	YOLOv3	UCAS-Avenue	Results show their model accurately identifies traffic rate via an edge device having processing speed of around 37.9 FPS & 92.0% accuracy.	The YOLOv3 model can accurately identify traffic rates via an edge device. This suggests that the YOLOv3 model can be used to develop edge-based traffic monitoring systems.
<b>Wang et al., (2020) [12]</b>	RCPM	NGSIM	Predictions of rear-end crashes using RCPM are shown to be much more accurate than those using the Berkeley, Honda, or multi-layer perceptron neural network-based techniques.	The RCPM approach can predict rear-end crashes more accurately than other approaches. This suggests that the RCPM approach can be used to develop more effective crash prediction systems.
<b>Zhou et al., (2020) [13]</b>	VGRAN	PEMSD4	The results demonstrate the deterministic spatiotemporal modelling used in existing sensor network-based traffic forecasting systems, vector node representation uncertainty improves the capability of the model to represent dynamic, time-varying sensor properties and intricate topological topologies of sensor networks.	The VGRAN model can improve the capability of existing sensor network-based traffic forecasting systems. This suggests that the VGRAN model can be used to develop more accurate and reliable traffic forecasting systems.
<b>Neetesh et al., (2020) [14]</b>	DITLCS	PEMSD4	To demonstrate the efficacy and high efficiency of DITLCS, its results were further compared to fuzzy neural networks and some priority-based approaches.	The DITLCS approach can be used to improve the efficacy and high efficiency of traffic management systems. This suggests that the DITLCS approach can be used to develop more efficient and effective traffic management systems.
<b>Shengnan Guo et al. (2019) [15]</b>	ASTGCN (Attention based spatial-temporal graph convolutional network)	PEMS04, PEMS08	The ASTGCN model was evaluated on two real-world traffic datasets. The results showed that the ASTGCN model was able to outperform the state-of-the-art methods in terms of accuracy and other metrics.	The ASTGCN model can outperform the state-of-the-art methods in terms of accuracy and other metrics. This suggests that the ASTGCN model may be a better choice for traffic forecasting than other deep learning approaches.
<b>Li et al., (2019) [16]</b>	Artificial Neural Networks (ANN)	PEMSD8	The results show that when the whole upstream and downstream context is considered, all of the methods such as ANN, SVM, Regression-tree and KNN (k-nearest neighbour) perform better than when just data from the examined segment is used.	When the whole upstream and downstream context is considered, all of the methods such as ANN, SVM, Regression-tree and KNN perform better than when just data from the examined segment is used. This suggests that it is important to consider the upstream and downstream context when developing traffic forecasting systems.
<b>Ferdowsi et al., (2019) [17]</b>	LSTM, RBM, RNN, and CNN	PEMSD8	It provides edge analytics architecture along with some deep-learning approaches and introduces a safe and secure TIS.	The edge analytics architecture, along with some deep-learning approaches, can introduce a safe and secure TIS. This suggests that edge analytics and deep learning can be used to develop more secure and reliable traffic information systems.

### III. METHODOLOGY

The problem of regulating traffic has persisted since the beginning of time. To function in today's environment, one must have access to current technology as it becomes trendy for its automatic features. Intelligent Traffic Systems, also called Intelligent Transportation Systems, use data and communication networks to address problems with traffic management. ITS refers to software that uses sensors and wireless connections to increase transportation efficiency. This ITS employs cutting-edge methods of traffic management to

address issues like congestion and a lack of safety. Information, control, and electronic technologies that use wireless and wireline connectivity are used to enhance ITS. The main issue in the modern traffic management system is addressing the problem of excessive speed. Doppler's Phenomenon is a tool for determining velocity [18].

#### A. Proposed Methodology

In this section, we proposed a traffic prediction model using machine learning in detail.



## 1. Machine Learning (ML)

The author A. Samuel [19] defines machine learning as the study of how computers may learn to improve themselves without being given a set of instructions and developed a popular checkers-playing software. When applied to computers, machine learning (ML) helps them learn how to process data more effectively. Even after looking at the data, it can't always deduce what's going on. Therefore, use machine learning. The abundance of datasets has sparked widespread interest in machine learning, which is being utilized across various industries to extract valuable insights. The primary goal of ML is to learn from data. To tackle the challenge of managing massive datasets, mathematicians and computer scientists have developed various methods. When addressing data-related problems, machine learning utilizes numerous techniques. It is worth noting that data science experts emphasize that there is no one-size-fits-all approach to problem-solving, as factors such as the number of variables involved and the type of model that is most effective can significantly influence the method used. Figure 3 provides the categorization of techniques of machine learning [20].

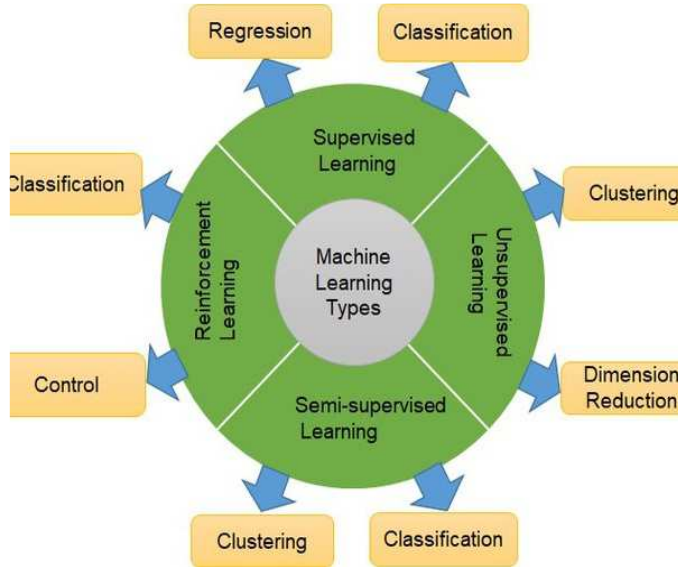


Fig. 3. Categories of Machine Learning Techniques

### i. Support Vector Regression (SVR)

The SVR model aims to find a function  $f(x)$  that approximates the relationship between input variables  $x$  and corresponding output variables  $y$  [22]. The general form of the SVR model can be represented as:

$$f(x) = w \cdot \phi(x) + b \quad (1)$$

Where:

- $f(x)$  is the predicted output for input  $x$
- $w$  represents the weight vector
- $\phi(x)$  denotes the feature mapping of input  $x$
- $b$  is the biased term

The linear kernel function is one of the commonly used kernel functions in SVR. It can be expressed as the inner product of the input vectors.

$$K(x, x') = xx' \quad (2)$$

The polynomial kernel function allows for non-linear mappings by introducing polynomial terms.

$$K(x, x') = (\gamma(xx') + r)^d \quad (3)$$

## 2. The proposed Model: Intelligent Transportation Model (ITM)

This section defined a proposed methodology based on Traffic Prediction Using Machine Learning in Intelligent Transportation Systems. Figure 4 describes the proposed approach.

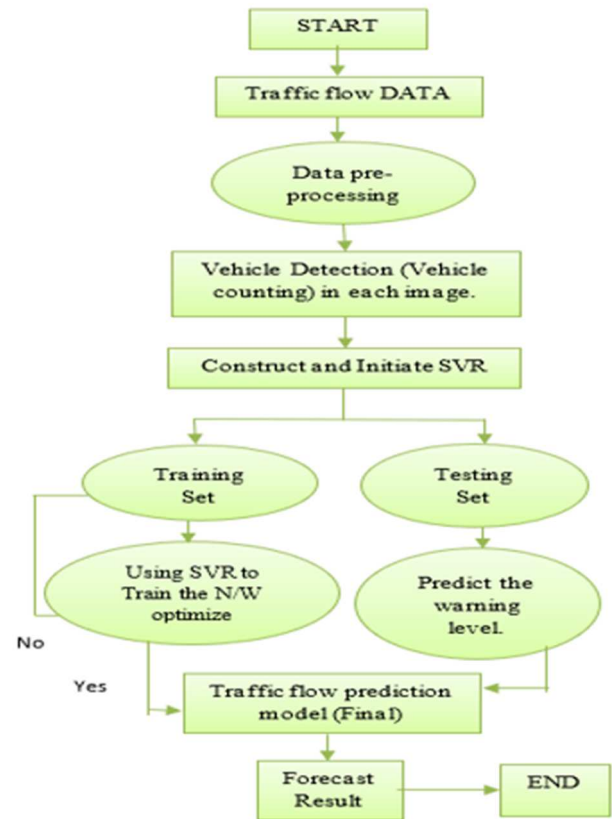


Fig. 4. Block Diagram of Proposed Intelligent Transportation Model (ITM)

TABLE II: PROCESS FOR IMPLEMENTATION OF ITM MODELS

<i>Procedure For ITM Model:</i>	
<b>Step 1:</b>	The preprocessing of the data on the traffic flow primarily includes the processing of missing values, denoising, and data reduction.
<b>Step 2:</b>	Detect Vehicle Detection (Vehicle counting) in each image.

<b>Step 3:</b>	Select the data features to be used for the traffic flow using the SVR algorithm model.
<b>Step 4:</b>	The removal of redundant and unnecessary traffic data results in the formation of two distinct datasets, one for training and one for testing.
<b>Step 5:</b>	Determine the acceptable value range for the SVR parameters.
<b>Step 6:</b>	Then training data is fed into the SVR model. Further, the process is initiated to generate predictions for the desired outcome.
<b>Step 7:</b>	Determine whether the currently used combination of parameters fulfils the standards for maximum accuracy. If the conditions are met, the parameter optimization process can be terminated. The SVR model has been successfully run with the present parameter combination because it is the best possible parameter combination.
<b>Step 8:</b>	Similarly, for testing, the warning level is predicted and then passed on to be matched with the output of the training data to be matched with the output of the final model to forecast the final output that occurred.

#### B. Datasets

To validate the proposed model, two distinct datasets of California highway traffic, namely PeMSD4 and PeMSD8, were utilized. Caltrans' Performance Measuring System (PeMS) [23] acquires the data in real-time with a 30-second interval. The original traffic data is aggregated into 5-minute intervals. Over 39,000 detectors have been placed along main thoroughfares in California's major cities as part of this system. All of the sensor stations' locations are captured in the data sets. In proposed models, it takes into account three metrics related to traffic: flow rates, average occupancy and average speed.

**PeMSD4:** - PeMSD4 is a collection of road traffic data collected from 3848 detectors installed along 29 roadways in the Bay Area of San Francisco. The dataset comprises information from January 2018 to February 2018, for model training, the initial 50 days of data are used and for testing, the rest of the data are used.

**PeMSD8:** - PeMSD8 contains traffic data collected from 1979 detectors located along eight highways in San Bernardino over the months of July and August 2016. The dataset is split into two sets, with the first 50 days of data used for training and the final 12 days reserved for testing.

## IV. EXPERIMENTAL ANALYSIS

To test the proposed model's efficacy, conduct comparison tests on two true-world highway traffic datasets.

#### A. Evaluation Parameters

In this section, RMSE (Root Mean Square), MAE (Mean Absolute Error) is taken as evaluation parameters for the proposed model. These metrics can be defined as follows:

##### 1. Root Mean Square Error (RMSE)

In the fields of meteorology, quality of air, and climate research, the RMSE is a common statistical tool for gauging model performance. Despite its widespread use throughout the years, experts still disagree on which measure of model mistakes is most accurate. It is common practice in geosciences to provide the RMSE as a measure of model error, while other researchers prefer to report just the mean-absolute error (MAE), arguing that the RMSE is too subjective [24]. An indicator of the distance between known and estimated places. RMSE to compare model results. A total of  $n$  journeys have been taken. Whereas  $y_i$  is the present time,  $\hat{y}_i$  indicates the forecasted value. The following equation is used to calculate root-mean-squared error [25]:

$$RSME = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4)$$

##### 2. Mean Absolute Error (MAE)

To evaluate the regression model based on vector-to-vector approaches, MAE is statistically applied; it was originally developed as a measure of average error [26]. Which represent, respectively, the maximum and median dissimilarity between observed and predicted traffic flows and the results obtained from the proposed framework can be assessed using the MAE formula. Here is the formula for the MAE:

$$MAE = \frac{1}{n} \sum_{i=0}^n |\hat{l}_i - l_i| \quad (5)$$

In this formula,  $\hat{l}_i$  indicates the observed traffic situation,  $l_i$  indicates the expected traffic condition, and  $n$  indicates the total number of forecasted points. A more precise forecast result, indicated by a lower value of MAE, corresponds to a smaller difference between the actual traffic rate and the predicted outcome [27].

#### B. Baseline Models

- LSTM (Hochreiter and Schmidhuber 1997): LSTM is a unique RNN model [29].
- A spatial-temporal graph convolution model based on the spatial approach is called STGCN (Li et al. 2018) [30].
- A multi-level attention-based recurrent neural network model called GeoMAN has been presented for the geo-sensory time series prediction problem (Liang et al. 2018) [31].
- ASTGCN (Shengnan Guo, 2019): To apply weights to nodes or time steps based on relevance, ASTGCN uses attention methods [15].
- MSTGCN (Shengnan Guo, 2019) : It incorporates attention mechanisms to assign importance weights to nodes or time steps [15].



## V. RESULTS AND DISCUSSION

Among the six models evaluated (LSTM, STGCN, GeoMAN, MSTGCN, ASTGCN, SVR) on PeMSD4 and PeMSD8 datasets for traffic flow prediction, SVR emerged as the best performing model. The results summarized in Table 3 present the root mean square error (RSME) and mean absolute error (MAE) over the next hour.

Our findings indicate that the SVR model consistently outperformed the other models across both datasets in terms of all evaluation metrics. In the comparison of deep learning-based models, SVR demonstrated better prediction results than LSTM, STGCN, GeoMAN, MSTGCN, and ASTGCN. The SVR model's performance highlights its strength in capturing the underlying patterns and correlations in traffic data, leading to accurate predictions.

It is worth noting that while the deep learning models considered temporal and spatial correlations, SVR, as a machine learning-based method, leverages support vector regression to capture the complex relationships in the data. This approach proved effective in achieving high prediction accuracy, outperforming the deep learning models in this study. Its superior performance, as demonstrated by lower RSME and MAE values, showcases its efficacy in accurately forecasting traffic patterns.

TABLE III. OVERALL PERFORMANCE OF SEVERAL METHODS ON THE PEMSD4 AND PEMSD8.

Models	PeMSD4		PeMSD8	
	MAE	RSME	MAE	RSME
<b>LSTM</b>	29.45	45.82	23.18	36.96
<b>STGCN</b>	25.15	38.29	18.88	27.87
<b>GeoMAN</b>	23.64	37.84	17.84	28.91
<b>MSTGCN</b>	22.73	35.64	17.47	26.47
<b>ASTGCN</b>	21.80	32.82	16.63	25.27
<b>SVR</b>	<b>19.21</b>	<b>30.57</b>	<b>13.87</b>	<b>22.76</b>

Figure 5, 6, 7, 8 illustrates the performance variations of different prediction methods as the prediction interval expands. Generally, as the prediction interval lengthens, the prediction task becomes more challenging, leading to an increase in prediction errors. Notably, the accuracy of predictions declines notably as the prediction interval widens.

Deep learning methods exhibit relatively slower error increments as the prediction interval increases, demonstrating their overall favourable performance. Among these methods, our SVR model consistently achieves the highest prediction accuracy across the entire range. This is particularly evident in long-term predictions, where the disparities between SVR and other baseline models become more pronounced. This

highlights the effectiveness of our approach, which combines attention mechanisms and graph convolution to effectively capture dynamic spatial-temporal patterns in traffic data.

It is worth noting that the SVR model's superior performance indicates its capability to uncover and leverage intricate spatio-temporal patterns. This showcases the potential of our proposed methodology for accurate traffic flow forecasting.

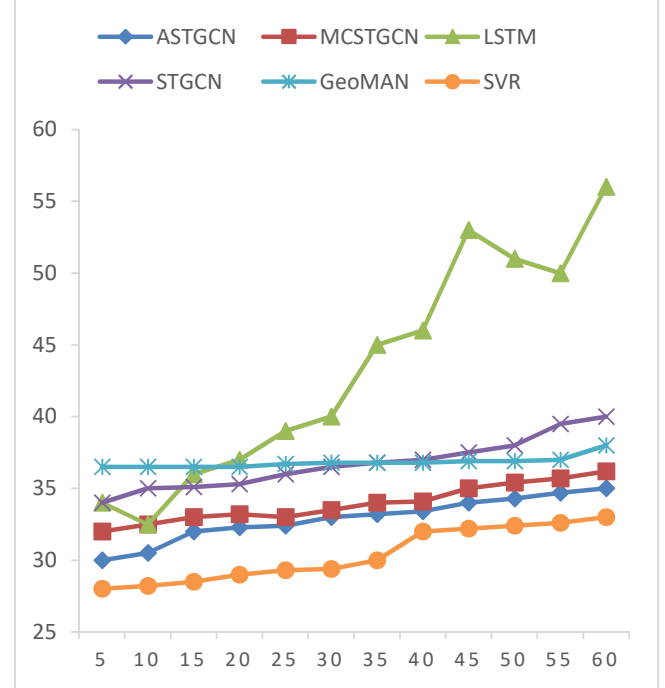


Fig. 5. Graph of RMSE of various approaches on PeMSD4.

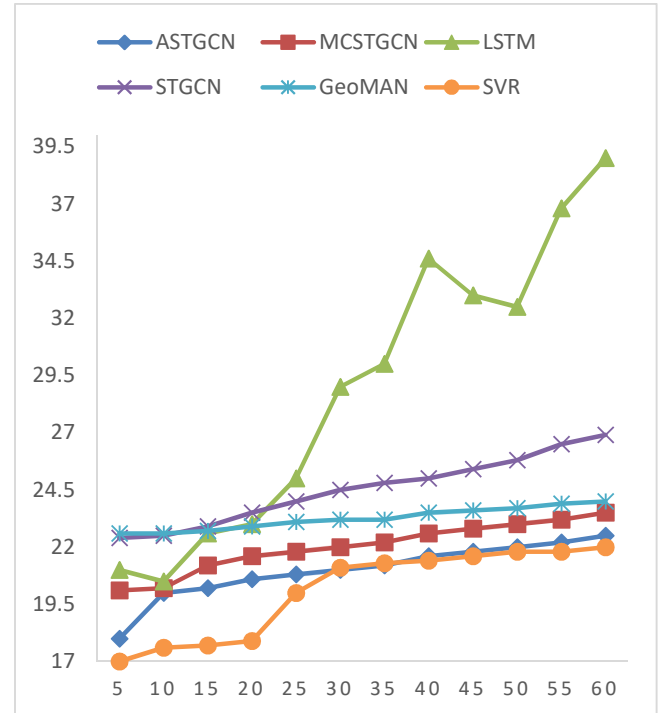


Fig. 6. Graph of MAE of various approaches on PeMSD4.

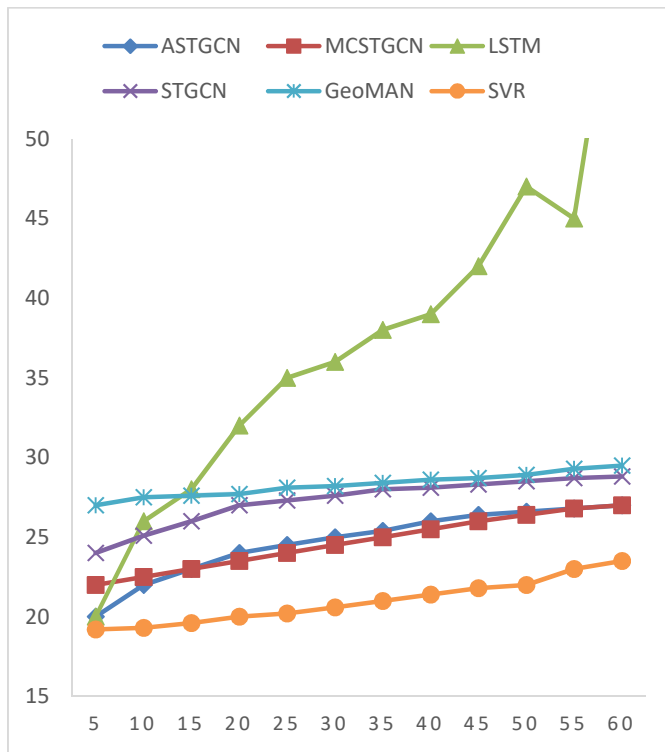


Fig. 7. Graph of RMSE of various approaches on PeMSD8.

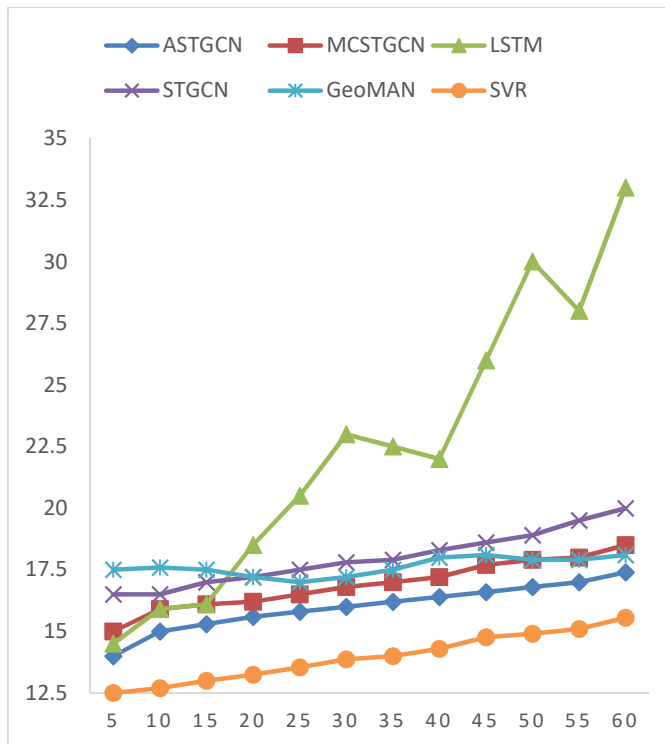


Fig. 8. Graph of MAE of various approaches on PeMSD8.

## VI. CONCLUSION

Many organizations and individuals could benefit from current and accurate information regarding traffic patterns. Helping riders and drivers make better route decisions improves traffic flow, increases operational efficiency, and decreases

emissions. Intelligent Transportation Systems (ITSs) provide more precise traffic forecasts as they are refined and implemented. High-tech public transportation networks, tourist information networks, and traffic management systems all rely on it, so it is being fixed. Compared proposed models to the two previous approaches on PeMSD4 and PeMSD8. The outcomes of the traffic flow forecast presentation are displayed in table 2 below. Information for the next hour can be found in this table. As can be shown, SVR achieves the best results across all evaluation measures in both datasets. Both sets of data have this property. The proposed machine-learning-based approach typically achieves better prediction outcomes than competing approaches. When compared to competing models like MSTGCN and ASTGCN, the proposed method's solution stands head and shoulders above the rest.

## REFERENCES

- [1]. Meena, Gaurav, Deepanjali Sharma, and Mehul Mahrishi. "Traffic prediction for intelligent transportation system using machine learning." In *2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)*, pp. 145-148. IEEE, 2020.
- [2]. Javed, Muhammad Awais, Duy Trong Ngo, and Jamil Yusuf Khan. "Distributed spatial reuse distance control for basic safety messages in SDMA-based VANETs." *Vehicular Communications* 2, no. 1 (2015): 27-35.
- [3]. Javed, Muhammad Awais, Elyes Ben Hamida, and Wassim Znaidi. "Security in intelligent transport systems for smart cities: From theory to practice." *Sensors* 16, no. 6 (2016): 879.
- [4]. Guerrero-Ibanez, Juan Antonio, Sherali Zeadally, and Juan Contreras-Castillo. "Integration challenges of intelligent transportation systems with the connected vehicle, cloud computing, and internet of things technologies." *IEEE Wireless Communications* 22, no. 6 (2015): 122-128.
- [5]. Contreras-Castillo, Juan, Sherali Zeadally, and Juan Antonio Guerrero-Ibanez. "Internet of vehicles: architecture, protocols, and security." *IEEE internet of things Journal* 5, no. 5 (2017): 3701-3709.
- [6]. Gao, Honghao, Wanqiu Huang, and Xiaoxian Yang. "Applying Probabilistic Model Checking to Path Planning in an Intelligent Transportation System Using Mobility Trajectories and Their Statistical Data." *Intelligent Automation & Soft Computing* 25, no. 3 (2019).
- [7]. Tanwar, Sudeep, Sudhanshu Tyagi, and Sachin Kumar. "The role of the internet of things and smart grid for the development of a smart city." In *Intelligent communication and computational technologies*, pp. 23-33. Springer, Singapore, 2018.
- [8]. The intelligent transportation system. <https://bit.ly/3e8ZtGz>. Accessed: 24-12-2020; 2019.
- [9]. Chan, Robin Kuok Cheong, Joanne Mun-Yee Lim, and Rajendran Parthiban. "A neural network approach for traffic prediction and routing with missing data imputation for the intelligent transportation system." *Expert Systems with Applications* 171 (2021): 114573.
- [10]. Li, Cong, and Pei Xu. "Application on traffic flow prediction of machine learning in intelligent transportation." *Neural Computing and Applications* 33, no. 2 (2021): 613-624.
- [11]. Chen, Chen, Bin Liu, Shaohua Wan, Peng Qiao, and Qingqi Pei. "An edge traffic flow detection scheme based on deep learning in an intelligent transportation system." *IEEE Transactions on Intelligent Transportation Systems* 22, no. 3 (2020): 1840-1852.
- [12]. Wang, Xin, Jing Liu, Tie Qiu, Chaoxu Mu, Chen Chen, and Pan Zhou. "A real-time collision prediction mechanism with deep learning for the

- intelligent transportation system." *IEEE transactions on vehicular technology* 69, no. 9 (2020): 9497-9508.
- [13]. Zhou, Fan, Qing Yang, Ting Zhong, Dajiang Chen, and Ning Zhang. "Variational graph neural networks for road traffic prediction in intelligent transportation systems." *IEEE Transactions on Industrial Informatics* 17, no. 4 (2020): 2802-2812.
- [14]. Kumar, Neetesh, Syed Shameerur Rahman, and Navin Dhakad. "Fuzzy inference enabled deep reinforcement learning-based traffic light control for the intelligent transportation system." *IEEE Transactions on Intelligent Transportation Systems* 22, no. 8 (2020): 4919-4928.
- [15]. Guo, Shengnan, et al. "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. No. 01. 2019
- [16]. Li, Linchao, Xu Qu, Jian Zhang, Yonggang Wang, and Bin Ran. "Traffic speed prediction for intelligent transportation system based on a deep feature fusion model." *Journal of Intelligent Transportation Systems* 23, no. 6 (2019): 605-616.
- [17]. Ferdowsi, Aidin, Ursula Challita, and Walid Saad. "Deep learning for reliable mobile edge analytics in intelligent transportation systems: An overview." *IEEE vehicular technology magazine* 14, no. 1 (2019): 62-70.
- [18]. Meena, Gaurav, Deepanjali Sharma, and Mehul Mahrishi. "Traffic prediction for intelligent transportation system using machine learning." In *2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)*, pp. 145-148. IEEE, 2020.
- [19]. A.L. Samuel, some studies in machine learning using the game of checkers, IBM J. Res. Develop. 3 (3) (1959) 210-229
- [20]. Mahesh, Batta. "Machine learning algorithms-a review." *International Journal of Science and Research (IJSR)*. [Internet] 9 (2020): 381-386.
- [21]. Swana, Fezeka & Doorsamy, Wesley. (2021). An Unsupervised Learning Approach to Condition Assessment on a Wound-Rotor Induction Generator. *Energies*. 14. 602. 10.3390/en14030602.
- [22]. Cong, Z., & Pei, J. (2021). Support vector regression for traffic speed forecasting. *IEEE Access*, 9, 123649-123660.
- [23]. Guo, Shengnan, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. "Attention-based spatial-temporal graph convolutional networks for traffic flow forecasting." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, pp. 922-929. 2019.
- [24]. Chai, Tianfeng, and Roland R. Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)." *Geoscientific Model Development Discussions* 7, no. 1 (2014): 1525-1534.
- [25]. Lee, Yong-Ju, and Okgee Min. "Comparative analysis of machine learning algorithms to urban traffic prediction." In *2017 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 1034-1036. IEEE, 2017
- [26]. Qi, Jun, Jun Du, Sabato Marco Siniscalchi, Xiaoli Ma, and Chin-Hui Lee. "On mean absolute error for deep neural network-based vector-to-vector regression." *IEEE Signal Processing Letters* 27 (2020): 1485-1489.
- [27]. Zhao, Wentian, Yanyun Gao, Tingxiang Ji, Xili Wan, Feng Ye, and Guangwei Bai. "Deep temporal convolutional networks for short-term traffic flow forecasting." *IEEE Access* 7 (2019): 114496-114507
- [28]. A.L. Samuel, some studies in machine learning using the game of checkers, IBM J. Res. Develop. 3 (3) (1959) 210-229.
- [29]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.
- [30]. Li, Z., Chen, Y., & Wang, L. (2018). Spatial-temporal graph convolutional networks for traffic forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 19(11), 3144-3155.
- [31]. Liang, Z., Wang, S., & Liu, Y. (2018). GeoMAN: A multi-level attention-based recurrent neural network for geo-sensory time series prediction. *IEEE Transactions on Knowledge and Data Engineering*, 30(12), 2954-2966.



# TransNet: a comparative study on breast carcinoma diagnosis with classical machine learning and transfer learning paradigm

Gunjan Chugh<sup>1</sup> · Shailender Kumar<sup>1</sup> · Nanhay Singh<sup>2</sup>

Received: 20 February 2023 / Revised: 7 July 2023 / Accepted: 11 September 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

## Abstract

Breast Carcinoma is a deadly disease; therefore, timely diagnosis is one of the most critical concerns that must be addressed globally since it can significantly enhance overall survival rates. Currently, Medical Imaging relies on Machine Learning (ML) and Deep Learning (DL) for accurate and early identification of diseases. In this article, a framework is proposed for diagnosing & classifying breast tumors using deep learning approaches. We have performed two experiments on the CBIS-DDSM (Curated Breast Imaging Subset of Digital Database for Screening Mammography) dataset. In the first approach, i.e., Deep feature extraction with ML classifier head, Deep Convolutional Neural Network (DCNN) models such as VGG16, VGG19, Res Net 50, and Res Net 152 are deployed as feature extractors, and the obtained features are utilized for training conventional machine learning classifiers. The second approach, called Deep Learning feature extraction with a neural network classifier, exploits Mobile Net, VGG16, VGG19, ResNet50, Res Net 152, and, Dense Net 169 for feature extraction and categorization. The results show that in the first case, Random Forest (RF) and XG Boost (XGB) Classifier perform best with 100% accuracy on the training set, whereas Support Vector Machine (SVM) and XGB exhibit 95%(+5%) on the Test dataset for all the models. In the second approach, Mobile Net, ResNet50, and Dense Net 169 outperform the other models with an accuracy of 97%(+2%) for both the Training and Test sets. The evaluated results have shown that the second approach depicts an increase in accuracy by 4%.

**Keywords** Breast Carcinoma · Computer-Aided diagnosis · Deep Convolutional Neural Network · Deep-learning

---

✉ Gunjan Chugh  
chugh.gunjan8917@gmail.com

<sup>1</sup> Department of Computer Science and Engineering, Delhi Technological University, Delhi, India

<sup>2</sup> Department of Computer Science and Engineering, Netaji Subhas University of Technology (East Campus), Delhi, India

## 1 Introduction

Breast Carcinoma is the most prominent tumor that impacts patients' physical and emotional health [1]. Breast Cancer grows and expands in breast tissues. More than 99 percent of all occurrences of breast carcinoma are diagnosed in women [2]. Because of its prevalence in low- and middle-income countries, female breast carcinoma has surpassed lung carcinoma as one of the most often diagnosed malignancies. Figure 1 shows the statistics for all types of cancers in 2020. Breast Carcinoma accounts for one-fourth of all cancers diagnosed in women worldwide [3].

In women, invasive ductal carcinoma (IDC) is the leading breast tumor [4]. As per the records of the World Health Organization(WHO), 2.3 million women were diagnosed with breast carcinoma in 2020, with 685000 fatalities reported worldwide [5]. In 2022, women in the US were expected to be diagnosed with 287,850 and 51,400;

### CANCER STATISTICS-YEAR 2020

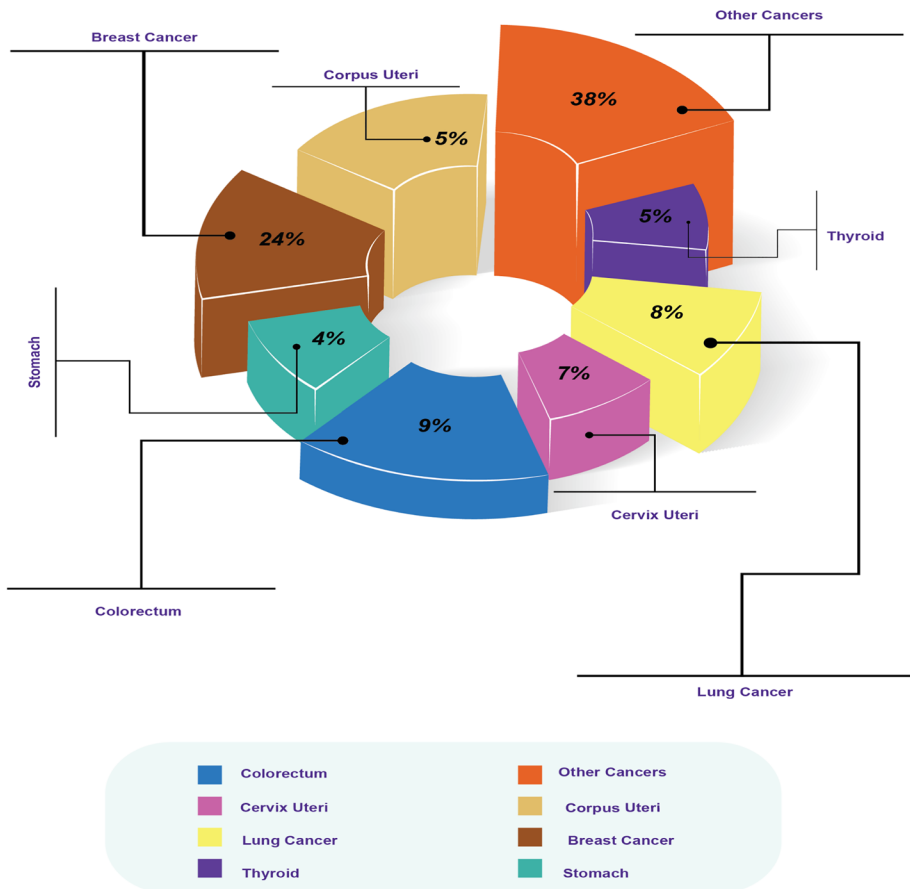


Fig. 1 Cancer Statistics (Female)- 2020 [3]

instances of invasive and non-invasive breast tumors, respectively, whereas 2,710 new incidences are likely to be detected in males [6]. By 2030, there might be 1.1 million new instances of breast cancer, and the difference between developed and developing countries may expand, in accordance with International Agency for Research on Cancer (IARC) of the WHO [7].

Breast carcinoma could be fatal if not detected in the preliminary stages. A malignant tumor grows aggressively and speedily spreads to other regions of the body [4]. Therefore, early and accurate diagnosis can reduce mortality and save the lives of people suffering from this deadly disease. Usually, an oncologist examines the clinical images and writes a summary based on a pre-defined documentary structure, such as Breast Imaging Reporting and Data System (BI-RADS). The ability to analyze medical images before the 1990s was constrained by processing power and diagnostic imaging procedures that focused at the composition of histology images [7]. But the accuracy and efficiency of diagnosis are affected due to huge number of patients and the scarcity of competent radiologists, which in turn, causes radiologists' mental attention to diminish [2].

Thus, Computer-aided detection/diagnosis(CAD) based medical image analysis has emerged as a useful tool to categorize and diagnose the tumor in the initial stages [8]. It serves as a second opinion and intimate the patients about breast anomalies [9]. ML and DL, the sub-branches of Artificial Intelligence, have developed various techniques/ algorithms for diagnosing and classifying breast cancer early. These algorithms were earlier adopted for chest radiographs and mammograms and are now used in alternative modalities like Computed Tomography(CT), ultrasound, etc. [10]. These algorithms provide timely diagnosis and thus improve the health of patients. As a result, the Breast Carcinoma fatality rate was reduced by 40% between the 1980s and 2020. Therefore, 2.5 million deaths from breast carcinoma could be avoided between 2020-2040 if the yearly fatality rate drops by 2.5% worldwide [5].

Some pre-processing operations are to be followed before feeding the image to ML/ DL models. First, noise is removed from the image, and then unwanted things, such as tags, and pectoral muscles, are eliminated. The approach to attain the best results in categorizing malignant and benign tumors is breast mass segmentation with CAD systems [11]. The Region of interest (ROI) is then located and provided to the model for further classification and diagnosis [12]. Figure 2 shows the CAD system pipeline. Convolutional neural networks (CNNs) have proven efficient in solving various medical pattern recognition challenges. CNNs with deeper architectures have been possible because of developments in GPUs and the availability of large-scale data sets. DCNN has shown promising results in interpreting biomedical images including segmentation, detection, classification, etc. [13].

In this study, we have designed a CAD model, i.e., TransNet for diagnosing breast carcinoma using Deep Neural Networks. The research has the following significant contributions:



**Fig. 2** CAD Pipeline



- The proposed framework i.e. TransNet presents a dual approach for breast carcinoma diagnosis. It involves deep feature extraction using Deep Convolution Neural Network (DCNN) Models and then their classification using Machine learning classifiers and Neural Net classifiers respectively.
- Pre-processing and augmentation of the CBIS-DDSM dataset are performed to optimize the proposed strategy and to reduce overfitting.
- The study shows the comparison between several pre-trained models.
- Several measures, including Accuracy, AUC, Precision, Recall, and Loss, are utilized and plotted on graphs to estimate the models' efficiency.
- In the TransNet framework the best results were attained by Mobile Net Model with Accuracy, AUC, Precision, and Sensitivity values of 98%, 0.98, 97%, and 96% respectively on the Testing Dataset.
- The research also explores the Comparison of the proposed architecture with other cutting-edge approaches to evaluate the effectiveness of the proposed methodology.

The article is structured in the following manner: Section 2 analyses the Related Work in the field of Breast Carcinoma diagnosis. Section 3 summarizes Deep Learning Approaches and Convolutional Neural Networks. The Proposed Architecture is discussed in Section 4. Section 5 presents the detailed results of the research along with a discussion summary. Section 6 provides a comparison of the proposed Architecture with state-of-the-art techniques and finally, Section 7 concludes the paper with future avenues.

## 2 Related work

Breast Cancer particularly affects younger people, and there is a crucial need to diagnose it early. In recent years, much research has been performed for diagnosing breast carcinoma using deep learning techniques. This section explores the analysis of several experts in this field.

In [14], the authors presented an improved and effective neural network model called Dense Net II for diagnosing and classifying breast carcinoma. The model yields 94.55% accuracy and an AUC of 0.91. In another research, Bevilacqua et al. [15] compared two approaches, i.e. Artificial Neural Network(ANN) and CNN for the classification of breast tomosynthesis images. The research concluded that the CNN-based approach gives higher accuracy and AUC value than other approaches. The authors in [16] utilized the Break His Dataset and performed breast cancer classification on histopathology images using VGG16, VGG19, Mobile Net, and ResNet 50. VGG16 outperforms the other models with the greatest accuracy of 94.67%.

[17] presents a systematic literature survey on ML and DL approaches for diagnosing breast carcinoma using mammograms. The study discusses the imaging modalities, datasets, and, techniques used, for the breast cancer CAD system. Performance measures, potential limitations, and future challenges are also outlined. In [13], the authors utilized 2282 mammograms and 324 tomosynthesis images for mass detection using the transfer learning approach. In this model, the researchers used the model trained on mammography images for further training of tomosynthesis images and achieved a gain in the AUC value from 0.81 to 0.90. By applying pruning, the number of neurons and parameters were lowered by 87.2% and 34.4%, respectively.

In another research by Sharma & Mehra [18], two approaches were proposed for categorizing breast carcinoma histopathology images on the BreakHis Dataset. One technique focuses on extracting handcrafted features, while the other uses models such as VGG16, VGG19, and ResNet 50. VGG16 with linear SVM gives the highest accuracy for classifying histopathology images. In [19], a feature fusion technique is discussed for diagnosing breast carcinoma using mammography, Ultrasound, and MRI. The strategy integrates handcrafted radiomic features with the deep features extracted from pre-trained networks. The authors concluded that the fusion classifier outperforms CNN and conventional CAD classifiers.

[20] provides an in-depth study on analyzing medical images using ML and DL techniques. The authors observed that DL outperforms ML models when it comes to evaluating enormous quantities of data. This study focuses on the detection & diagnosis of various medical illnesses such as Breast tumors, Brain disease, Diabetes, etc.

The researchers in [21] presented a multi-activation deep neural network for diagnosis on the WBCD dataset. Four types of Activation functions are used in separate hidden layers: Sigmoid, RELU, Leaky RELU, and Swish. The authors concluded that the multi-activation proposed deep neural network model performs better than other single-activation networks. In the future, researchers can use different combinations of activation functions for deep neural networks.

Senan et al. [22] used Alex Net on the Breast cancer digital repository (BCDR) dataset to categorize malignancy on histopathology images. The proposed model gave superior results to previous models, with an accuracy of 95% at the magnification factor of 40x and 400x. [23, 24] proposed a CAD framework and utilized the “You Look Only Once (YOLO)” model for detecting, segmenting, and classifying breast abnormalities on the DDSM dataset. Results demonstrated that YOLO provides tremendous results for detecting masses over pectoral muscle and dense tissues.

### 3 Deep Learning and Deep Convolutional Neural Networks(DCNN)

Artificial Intelligence(AI) has emerged as the most promising field for various types of research in the current industries. Deep Learning and Machine learning, the subfields of AI, are giving tremendous results in each & every sector. We also use these applications in our daily life, like scrolling the search engines, taking to digital assistants, playing innovative games, and using social media apps. Etc. In recent years DL and ML are also been widely used in the medical sector. These advanced technologies are helping doctors in the treatment, reducing the diagnosis time and thus saving the life of patients.

In Cancer Detection and Diagnosis, Convolutional Neural Network(CNN) has performed remarkably. These networks consist of multilayer neurons capable of recognizing valuable features and thus aiding detection and classification. ‘Deep CNN i.e. DCNN’ refers to the layers in the network [10]. Initial layers learn generalized features from the images, and the deeper layers learn more specific features. A generalized model for CNN is shown in Fig. 3.

**CNN can be implemented through the following mechanisms:**

1. **Training from scratch:** When training a CNN in this approach, a significant amount of training sample is fed to the network so that model could learn the features right

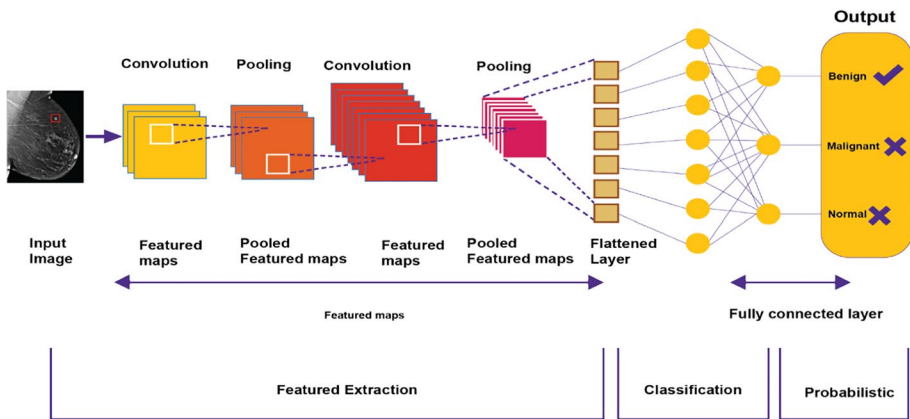


Fig. 3 CNN Architecture [25]

from the beginning. This process requires selecting suitable layers, hyper-parameters, optimizers etc. It is a time-consuming process and involves processing on powerful GPUs.

2. **Transfer Learning:** This strategy of training a CNN is used when there are fewer training instances in the target class. Various pre-trained models such as Mobile Net, Google Net, Res Net, Dense Net, and VGG could be utilized for training the network. These networks are already trained on massive datasets; thus, the network has learned the generic features. Therefore, the knowledge learned from the base domain is transferred or used for training the destination domain where training samples are less. As a result, this approach requires less time and thus could be used on CPUs.

Figure 4(a) and (b) visualize CNN Implementation Techniques

The Transfer Learning approach can be further implemented in the following three ways [8]:

1. **As Baseline Model:** In this category, the complete model is trained from beginning to end, and only the structure of the pre-trained model is exploited (Fig 5(a)).
2. **Fine-tuning:** This includes transferring weights from a pre-trained model to the destined model and could be accomplished in two methods: Layer by Layer and partial fine-tuning of the model. Layer-level tuning initiates with the outermost layer. Then additional layers are trained in chronological order whereas, in partial training, the weights of the initial layers are left unchanged, and the upper layer's weights are modified to train the unfamiliar dataset. (Fig. 5(b))
3. **Feature extractor:** This approach utilizes a pre-trained network's convolutional base in its original form, with no changes to its specified weights. Traditional classifiers substitute the dense layers of the pre-trained model. The convolutional base outcomes are passed directly to the train classifiers. (Fig. 5(c))

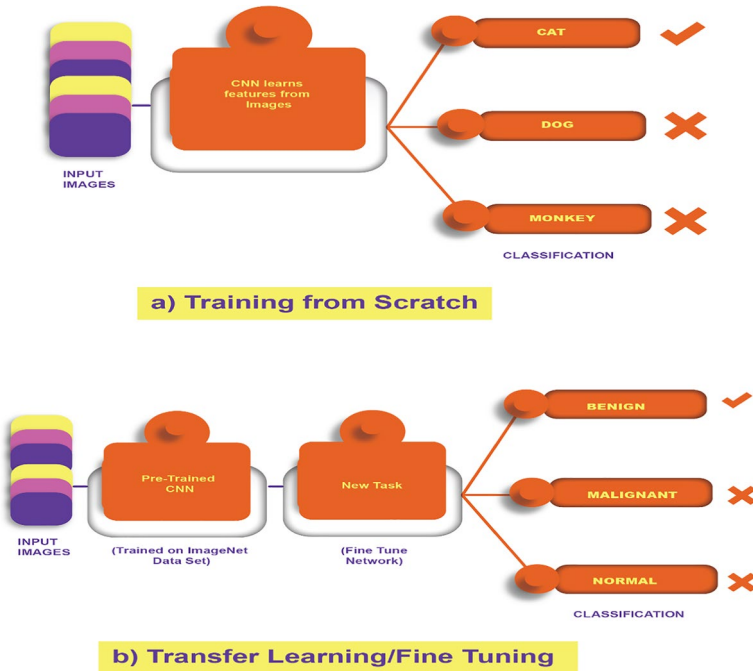


Fig. 4 CNN Implementation [26]

## 4 Proposed architecture

This research presents a TransNet model for diagnosing breast carcinoma using deep learning techniques. This section is outlined as follows: Sub-Section 4.1 outlines the dataset utilized. Sub-Section 4.2 discuss the Preprocessing and Data augmentation techniques used in the study. The proposed Framework is described in sub-section 4.3 followed by DCNN models in 4.4. Sub-Section 4.5 provides performance measures for evaluating a deep learning model.

### 4.1 Dataset

Breast Carcinoma could be diagnosed using different imaging mechanisms such as Mammography, Magnetic Resonance Imaging(MRI), Ultrasound, Digital Breast Tomosynthesis(DBT), Computed Tomography(CT), Histopathology Images etc. The details, along with the pros and cons for each modality, could be referred to in [8]. Breast cancer diagnosis via mammography is still recognized as a key element, as it detects tumors at initial stages and is associated with low radiation compared to other modalities. But the main drawback of mammography is that it fails to detect cancer in dense breasts. Digital Breast Tomosynthesis, which is often called 3D mammography, although associated with high radiation exposure, overcomes the limitations of mammography and works well in dense breasts too.

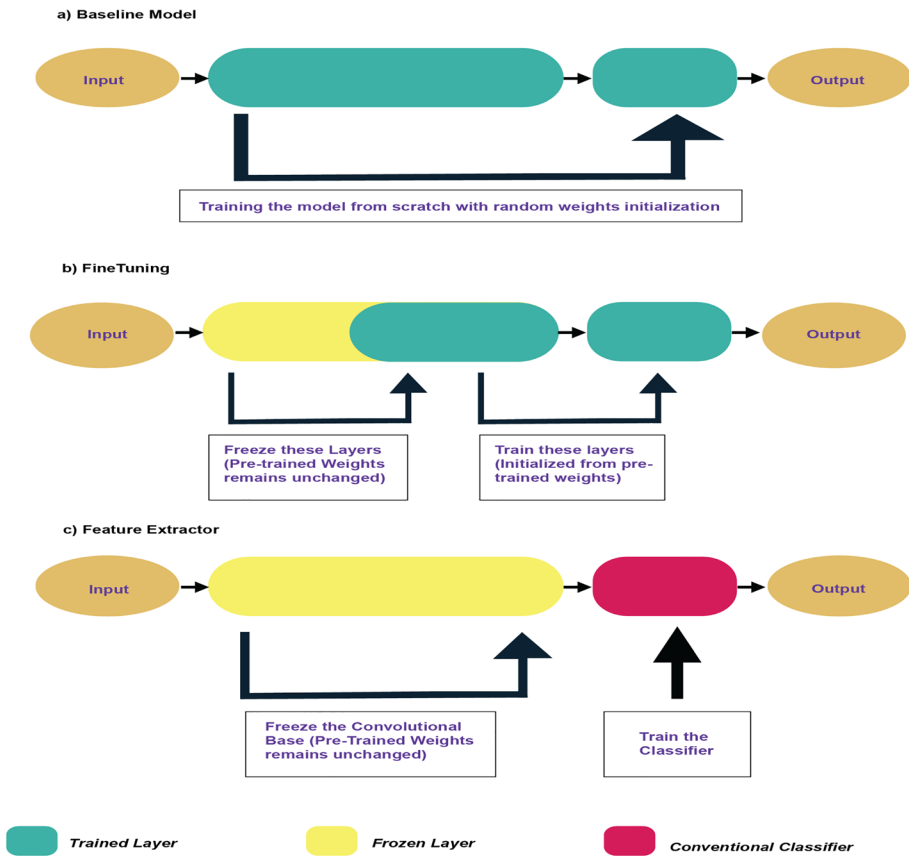


Fig. 5 Strategies for Transfer Learning Implementation [27]

In this research, we have used a mammography dataset i.e., CBIS-DDSM [28], which contains 9648 mammography images for 2412 patients. 695 cases are Normal, 867 are malignant(Positive), and 850 are benign(Negative). After applying data augmentation (Discussed in Section 4.2), the dataset comprises 55,885 images which include: 16,103 Normal, 20,088 Malignant(Positive), and 19,694 Benign(Negative) samples. The dataset includes three types of images- full images, ROI masks and cropped images. We have explored cropped images and utilized the data stored as a tfrecords file for Tensor-Flow (Fig. 6).

## 4.2 Pre-processing & data augmentation

### 4.2.1 Pre-processing

An important factor affecting Deep Convolutional Neural Networks' performance is Pre-Processing. We have utilized Cropped images from the DDSM dataset [28]. Steps followed for pre-processing and augmentation are shown in Fig. 7

Fig. 6 DDSM Data Set [28]

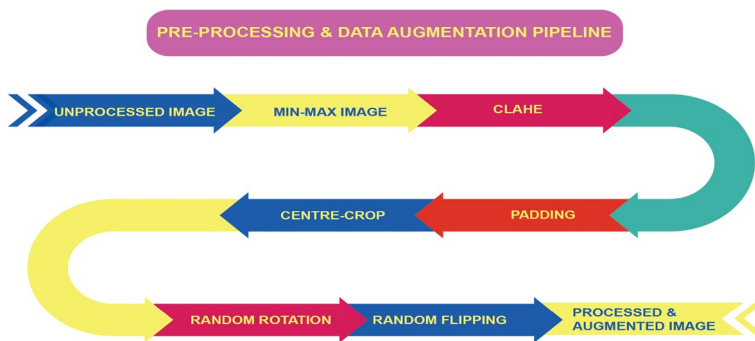
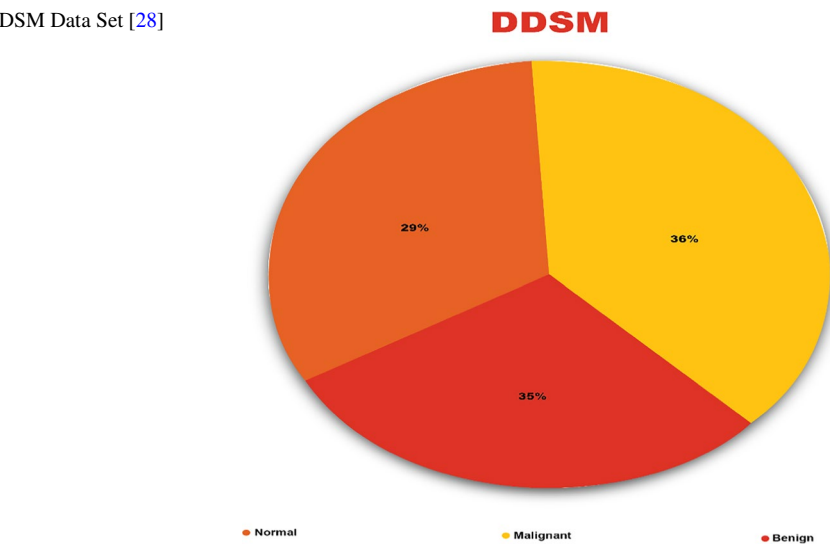


Fig. 7 Pre-Processing Pipeline

**Min-max normalize** Normalization is scaling your data within a specified range. In this technique, all values are scaled in the range  $[0,1]$ . It also reduces the unwanted noise in the image.

**CLAHE** To boost the contrast of greyscale pictures, Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied. It strengthens the model's ability to learn minute details, textures, and characteristics from the mammogram [29]

**Padding** The majority of computer vision tasks accept square images as input. In this step, images are padded into squares to feed them into the model.

**Centre-crop** Every ROI in the dataset was randomly cropped into  $598 \times 598$  images and then scaled down to  $299 \times 299$  pixels.



### 4.2.2 Data augmentation

Deep Neural Networks require vast amounts of data for training. Medical Images are not available in abundance, and thus problems such as overfitting might occur. Augmentation involves increasing the dataset size, i.e., the creation of additional data from the existing through several operations such as translation, rotation, scaling, etc. [30]

In our proposed approach, we applied the following operations on the DDSM dataset:

1. *Random Flipping*

The images are randomly flipped both horizontally and vertically.

2. *Random Rotation*

In this case, images are randomly rotated at different angles.

After applying data augmentation, the dataset comprises 55,885 images which are then split into three ratios i.e., 80:10:10 for the training, validation, and test sets.

## 4.3 Proposed methodology

### 4.3.1 First approach: Deep feature extraction with ML classifier head

This strategy employs Deep Neural Network Models i.e. VGG16, VGG19, ResNet50, and ResNet152 (Discussed in Section 4.4) as feature extractors. Extracted features are utilized for training the Machine Learning classifiers. This strategy deploys the following classifiers: KNN (i.e., k-Nearest Neighbors) with a k value of 8, SVM with RBF (Radial Basis Function) kernel, Random Forest (RF), Ada Boost, and XGB.

### 4.3.2 Second approach: Deep feature extraction with neural network classifier

Here, the pre-trained networks are deployed to extract features as well as for their further categorization. We utilized Mobile Net, VGG16, VGG19, Res-Net 50, Res-Net 152, and Dense-Net 169 (Discussed in Section 4.4).

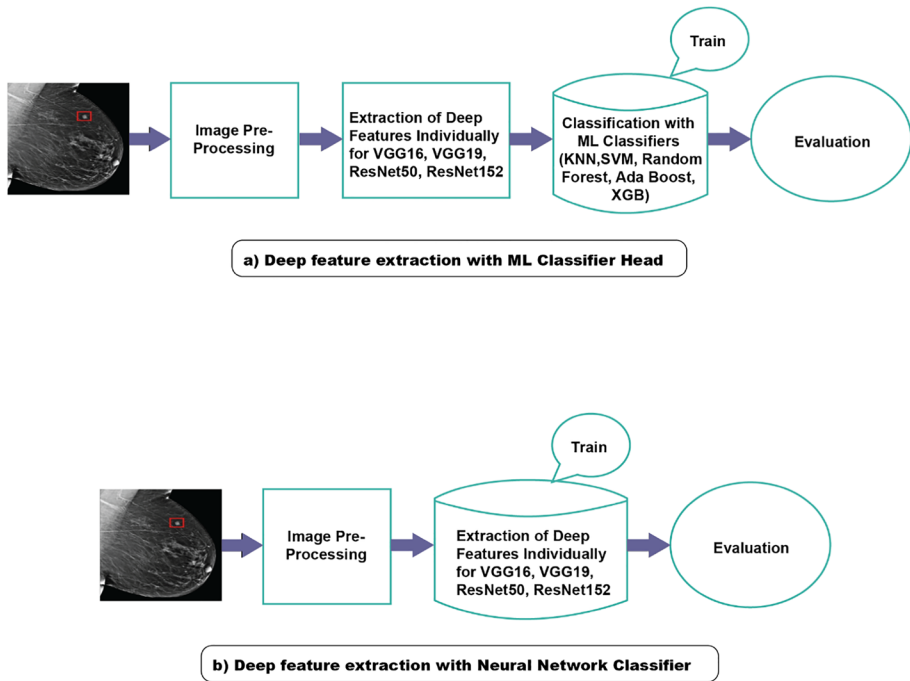
The results obtained for both Approaches are addressed in Section 5.1 proceeded by the discussion in Section 5.2

Figure 8 shows the Proposed Methodology.

## 4.4 Deep Convolutional Neural Network (DCNN) models

In this study, we proposed a framework, i.e., TransNet, for diagnosing breast carcinoma diagnosis by feature extraction strategies using DCNN. The CBIS-DDSM dataset was utilized for performing experiments. The framework proposed a dual approach for breast carcinoma diagnosis i.e. using ML classifier head and using Neural Net Classifier. Feature extraction is performed by training the models on the CBIS-DDSM dataset with a classifier head. The subsequent layers of the model contained various features extracted from the data.

The implementation was completed on the NVIDIA Tesla GPU system with 32 GB RAM. The model was trained for 20 epochs with early stopping criteria, set to minimize



**Fig. 8** Proposed Methodology

the validation loss with an interval of 5 epochs. The best weights during the training were restored at the end. The model underwent its initial training with a learning rate and batch size of .01 & 64, but the results were unsatisfactory. Thus, some hyper-parameters were optimized, and finally, the learning rate of .0001 and batch size of 128 was adopted. Both approaches utilize deep neural network models such as VGG16, VGG19, ResNet50, ResNet152, DenseNet169, and Mobile Net v2. In the second approach, Fine-tuning was performed for each model whereas in the first approach (i.e. ML Classifier head) No fine-tuning was performed, and the raw outputs without any activation functions were captured as the features. The initial default input image size of all the models is 224 X 224. But the input layer is resized to 100x100x3 as per the image size of the dataset. The specific details of these architectures are discussed as follows:

1. **VGG Net** [31]: VGG Net is a very popular and efficient DCNN that has shown promising results in image classification. Out of the various configurations, the most popular are VGG16 and VGG19 with 16 layers and 19 layers' depth. In VGG Net, 11 x 11 and 5 x 5 filters are substituted with a stack of 3 x 3 filters. The simultaneous deployment of a 3 x 3 filter might induce the impact of a large filter size (7 x 7, 5 x 5). VGG Net places 1x 1 convolution in between the convolution layers. Due to the large number of parameters, VGG Net is computationally expensive and requires longer training time on the system with less computational power.
2. **Res Net** [32]: This is a renowned deep-learning pre-trained network. As layers increase in a network, a problem known as a vanishing gradient usually occurs. To overcome

this, Res Net introduced the concept of residual learning. Here, we use Skip Connections, i.e., it bypasses a few stages of training and connects immediately to the output. Res Net comes in different versions with 50/101/152 layers' depth with an input image size of 224 x 224. ResNet50 has 24,649,953 trainable parameters whereas ResNet152 has 59,303,265 trainable parameters.

3. **Mobile Net** [30]: This model is developed specifically for mobile applications. Mobile Net reduces the computational complexity, i.e., the number of parameters, by using depth-wise separable convolutions. It has 3521569 trainable parameters. These low-powered small models are suitable for applications where resources are limited.
4. **Dense Net** [33]: Also called Dense Convolutional Network. Dense Nets provide numerous benefits, such as solving vanishing-gradient problems, enhancing feature propagation, reusing features, and reducing parameters count. Dense Net models are easy to train because they provide enhanced information flow across the network. Dense Nets usually have hundreds of layers and provide no optimization problems. These networks have several versions- DenseNet121, Dense Net 169, Dense Net 201, etc. In this research, we have utilized the Dense Net 169 model which possesses large number of trainable parameters i.e. 12,994,913, and thus could be deployed as feature extractors in various computer vision tasks.

#### 4.5 Performance measures

Table 1 outlines various measures for evaluating the efficiency of a deep learning model. Confusion matrix plots predicted and actual values. In Table 1,

- TP defines True Positive, i.e., the model correctly identified a breast Cancer patient.
- TN stands for True Negative, i.e., the model correctly identified non-breast Cancer patient.
- FP defines False Positive, i.e., the model detected a non-breast tumor person as a breast tumor patient.
- FN refers to False Negative, i.e., the model failed to identify a breast Cancer patient.

**Table 1** Performance Estimators

S No.	Performance Metric		
1.	$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$		
2.	$Sensitivity(Recall) = \frac{TP}{TP+FN}$		
3.	$Specificity = \frac{TN}{TN+FP}$		
4.	$Precision = \frac{TP}{TP+FP}$		
5.	$F - Measure = \frac{2 \times Precision \times Recall}{Precision+Recall}$		
6.	Confusion Matrix =		
		Actual results	
		Positive	Negative
	Positive	TP	FP
	Predicted Results		
	Negative	FN	TN

## 5 Results & discussion

### 5.1 Results

#### 5.1.1 First approach: Deep feature extraction with ML classifier head

Tables 2, 3, 4 depicts the Accuracy, Precision, and Recall Metrics on Training, Validation, and Test Datasets. The best outcomes shown by the DCNNs and the corresponding classifiers are highlighted in the tables. These results are discussed in Section 5.2.

Figures 9, 10, and 11 shows Accuracy, Precision, and Recall (Sensitivity) plots of several pre-trained models with their specific classifiers on Training, Validation, and Test Datasets.

#### 5.1.2 Second approach: Deep feature extraction with neural network classifier

Tables 5, 6, 7 depicts the Accuracy, AUC, Precision and Recall on the Training, Validation and Test dataset of DDSM for Mobile Net, ResNet50, VGG16, VGG19, DenseNet169, and ResNet152. The best results exhibited by the models are highlighted in the tables and discussed in Section 5.2.

Figures 12, 13 and 14 show the Accuracy, AUC, and Loss plots of Mobile Net, VGG16, VGG19, ResNet 50, ResNet 152, and Dense Net 169 models respectively.

**Table 2** Performance metrics on the DDSM Training dataset (Approach 1)

SNo.	DCNN Model	Classifier	Accuracy	Precision	Recall (Sensitivity)
1	VGG16	KNN	0.9	0.9	0.9
		SVM	0.94	0.94	0.94
		<b>RF</b>	<b>1</b>	<b>1</b>	<b>1</b>
		Ada Boost	0.9	0.88	0.89
		<b>XGB</b>	<b>1</b>	<b>1</b>	<b>1</b>
2	VGG19	KNN	0.89	0.9	0.89
		SVM	0.93	0.93	0.93
		<b>RF</b>	<b>1</b>	<b>1</b>	<b>1</b>
		Ada Boost	0.9	0.89	0.9
		<b>XGB</b>	<b>1</b>	<b>1</b>	<b>1</b>
3	ResNet50	KNN	0.91	0.9	0.91
		SVM	0.95	0.95	0.95
		<b>RF</b>	<b>1</b>	<b>1</b>	<b>1</b>
		Ada Boost	0.91	0.9	0.91
		<b>XGB</b>	<b>1</b>	<b>1</b>	<b>1</b>
4	ResNet152	KNN	0.91	0.9	0.91
		SVM	0.98	0.97	0.97
		<b>RF</b>	<b>1</b>	<b>1</b>	<b>1</b>
		Ada Boost	0.92	0.94	0.93
		<b>XGB</b>	<b>0.98</b>	<b>0.98</b>	<b>0.98</b>

**Table 3** Performance metrics on the DDSM Validation dataset (Approach 1)

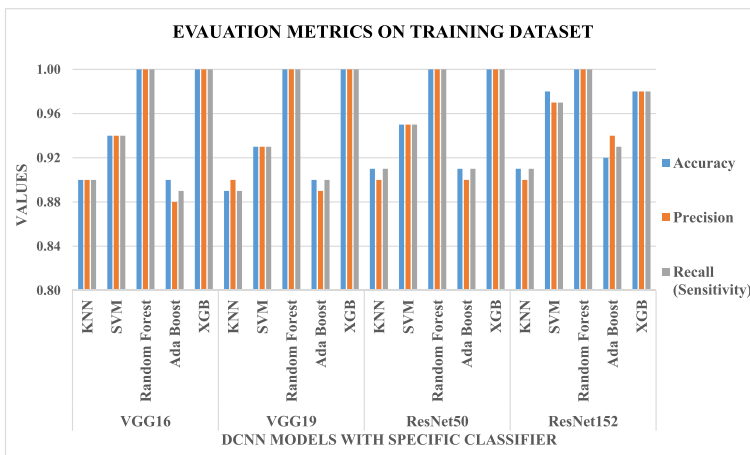
SNo.	DCNN Model	Classifier	Accuracy	Precision	Recall (Sensitivity)
1	VGG16	KNN	0.88	0.86	0.88
		SVM	0.91	0.91	0.91
		RF	0.89	0.88	0.89
		Ada Boost	0.89	0.87	0.89
		XGB	0.91	0.9	0.91
2	VGG19	KNN	0.88	0.87	0.88
		SVM	0.89	0.89	0.89
		RF	0.88	0.88	0.88
		Ada Boost	0.88	0.86	0.88
		XGB	0.9	0.89	0.9
3	ResNet50	KNN	0.89	0.87	0.89
		SVM	0.93	0.92	0.93
		RF	0.9	0.89	0.89
		Ada Boost	0.9	0.89	0.9
		XGB	0.92	0.92	0.91
4	ResNet152	KNN	0.88	0.86	0.88
		SVM	0.93	0.92	0.91
		<b>RF</b>	<b>0.98</b>	<b>0.97</b>	<b>0.98</b>
		Ada Boost	0.9	0.9	0.91
		<b>XGB</b>	<b>0.95</b>	<b>0.94</b>	<b>0.94</b>

## 5.2 Discussion

We compare the two proposed methods by evaluating the dataset on an 80:10:10 stratified split ratio for training, validation, and testing. 5 fold stratified cross-validation was performed and the average of all observations across the folds was taken. The first approach, i.e., Deep Learning feature extraction with ML classifier Head uses pre-trained models with ImageNet weights. No training or fine-tuning was performed on these models, and the raw outputs without any activation functions were captured as the features. The extracted features were supplied to the ML classifiers for binary categorization. The ML classifiers varied from simplistic models like k-Nearest-Neighbour to sophisticated gradient-boosting models like XG-Boost. As evident from Table 2, during the training phase, Random Forest and XG-Boost exhibited 100% accuracy, precision, and recall rate on all pre-trained models, whereas simpler models like KNN could achieve an average of 90% (+/- 2%) accuracy, precision and recall with a margin of 10% lower than other ML Classifiers. Support Vector Machine Classifier achieved intermediate results of 95% (+/- 2%) accuracy, precision, and recall rate. These results closely matched the testing data, as evident from Table 4. It is deduced from Table 4 that the deeper pre-trained models like ResNet152 outperform shallower models like ResNet50, VGG19, and VGG16 by a margin of 6% increase in accuracy, precision, and recall rates. Additionally, Random Forest and SVM outperformed the other models by a margin of 5% in all observable metrics. This approach is less computationally complex as the pre-trained networks were not retrained.

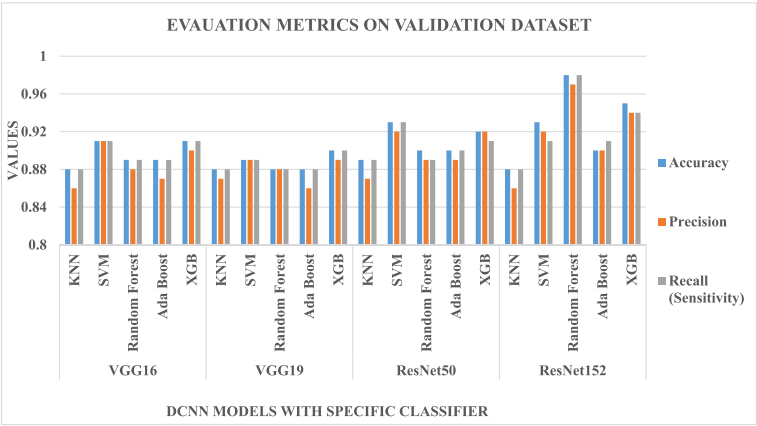
**Table 4** Performance metrics on the DDSM Test dataset (Approach 1)

SNo.	DCNN Model	Classifier	Accuracy	Precision	Recall (Sensitivity)
1	VGG16	KNN	0.88	0.87	0.88
		SVM	0.91	0.9	0.91
		RF	0.89	0.88	0.89
		Ada Boost	0.88	0.87	0.88
		XGB	0.91	0.91	0.9
2	VGG19	KNN	0.88	0.86	0.88
		SVM	0.9	0.89	0.9
		RF	0.88	0.87	0.88
		Ada Boost	0.89	0.87	0.89
		XGB	0.9	0.89	0.9
3	ResNet50	KNN	0.88	0.86	0.88
		SVM	0.93	0.93	0.93
		RF	0.9	0.89	0.89
		Ada Boost	0.9	0.89	0.9
		XGB	0.92	0.92	0.92
4	ResNet152	KNN	0.88	0.86	0.88
		SVM	0.93	0.92	0.91
		<b>RF</b>	<b>0.98</b>	<b>0.97</b>	<b>0.98</b>
		Ada Boost	0.9	0.9	0.91
		<b>XGB</b>	<b>0.95</b>	<b>0.94</b>	<b>0.94</b>

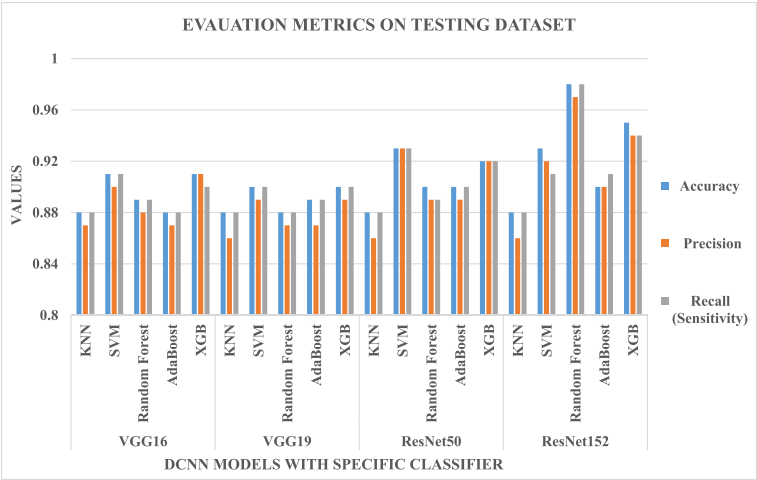
**Fig. 9** Evaluation Metrics Plots of DCNN models with various ML classifiers on the DDSM Training Dataset

The second approach, i.e., Deep Learning feature extraction with Neural Network Classifier, uses these models to execute feature extraction and classification. As a result, the computational complexity increased several folds. Dedicated GPU rigs were required to fine-tune the models. For the training phase, Mobile Net, ResNet50, and DenseNet169





**Fig. 10** Evaluation Metrics Plots of DCNN models with different ML classifiers on the DDSM Validation Dataset



**Fig. 11** Evaluation Metrics Plots of DCNN models with different ML classifiers on the DDSM Testing Dataset

**Table 5** Training Metrics

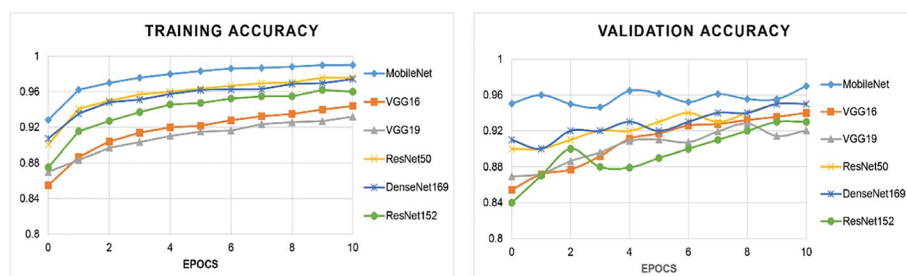
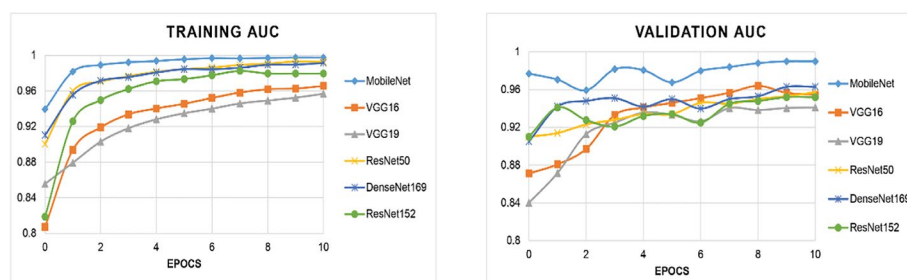
S No.	DCNN Model	Accuracy	AUC	Precision	Recall	Loss
1	Mobile Net	0.99	0.99	0.96	0.95	0.03
2	ResNet50	0.98	0.99	0.94	0.92	0.06
3	VGG16	0.94	0.96	0.92	0.91	0.14
4	VGG19	0.93	0.95	0.92	0.93	0.16
5	DenseNet169	0.97	0.99	0.94	0.92	0.07
6	ResNet152	0.96	0.98	0.92	0.90	0.08

**Table 6** Validation Metrics

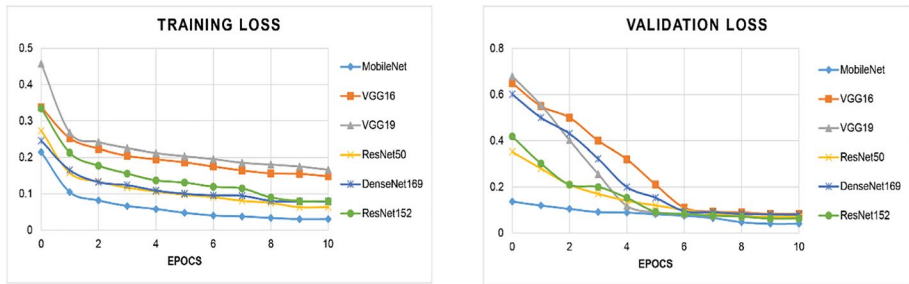
S No.	DCNN Model	Accuracy	AUC	Precision	Recall	Loss
<b>1</b>	<b>Mobile Net</b>	<b>0.97</b>	<b>0.98</b>	<b>0.94</b>	<b>0.95</b>	<b>0.04</b>
2	ResNet50	0.95	0.96	0.94	0.95	0.07
3	VGG16	0.92	0.93	0.90	0.91	0.08
4	VGG19	0.90	0.92	0.89	0.90	0.06
5	DenseNet169	0.92	0.94	0.92	0.91	0.06
6	ResNet152	0.93	0.93	0.93	0.92	0.08

**Table 7** Test Metrics

S No.	DCNN Model	Accuracy	AUC	Precision	Recall	Loss
<b>1</b>	<b>Mobile Net</b>	<b>0.98</b>	<b>0.98</b>	<b>0.97</b>	<b>0.96</b>	<b>0.10</b>
2	ResNet50	0.96	0.95	0.95	0.93	0.16
3	VGG16	0.92	0.95	0.91	0.92	0.17
4	VGG19	0.91	0.94	0.90	0.89	0.19
5	DenseNet169	0.93	0.95	0.93	0.91	0.17
6	ResNet152	0.92	0.95	0.94	0.92	0.23

**Fig. 12** Training and Validation Accuracies**Fig. 13** Training and Validation AUC

performed better than older models like VGG-16 and VGG-19. The same metrics were observed during the test phase as well. The other models, however, closely matched the best-performing models. Their less complex nature, may also be preferred during the



**Fig. 14** Training and Validation Loss

deployment phase for faster predictions and lower memory footprint. Comparing the second approach against the first one, DCNN with Neural Net Classifier highlights a 4% increase in performance by the classical machine learning models when paired with deep feature extraction techniques. This is because in this approach Deep Neural Networks are utilized for feature extraction and classification compared to ML classifiers in the first approach. In the first approach, no fine tuning was performed whereas this approach fine tune the pertained networks according to the requirement of the target domain and thus increases the model's efficiency with high accuracy values. Hyperparameter optimization was also performed which further contributes to enhancement in accuracy in the second approach.

Drawbacks of the study include Large dataset requirements for training Deep Convolutional Networks, the Difference in capturing mechanisms of various Mammogram/MRI machines, the Requirements of heavy computing hardware such as High-performance computing systems or GPU-based devices, and the Model's complexity.

In the future, this framework could be extended for feature fusion of multiple models and could be utilized on other datasets in this field. The proposed method could also be developed on other imaging modalities with several ensemble techniques. The study does not address the class imbalance issues and generalizability of the proposed framework to the unseen data, which could also be incorporated in future avenues.

## 6 Comparison with cutting-edge techniques

Table 8 compares the Proposed CAD model (TransNet) with other reference techniques in the domain. The articles are compared from 2018 to 2023. As we observe in Table 8, the TransNet model (highlighted in the table) performs better than the other cutting-edge techniques for Breast Carcinoma Classification.

## 7 Conclusion

The most prevalent malignancy among women is breast carcinoma. Recent advances in deep learning algorithms have proved that early detection and diagnosis improves mortality rates and thus saves the life of patients. This paper proposes a framework, i.e., TransNet, for diagnosing and classifying breast carcinoma on the CBIS-DDSM dataset. We have trained

**Table 8** Comparative Analysis with Recent Approaches for Breast Carcinoma Diagnosis Using Deep Learning Techniques

Year	Reference	Imaging Modality	Feature extraction/DNN Model Adopted	Classification	Dataset	Accuracy (%)	AUC
2018	[34]	Mammography	CNN	--	MIAS	85.85	--
2018	[35]	Histopathology Images	CNN, LSTM	SVM, Softmax	BreakHis	91	--
2018	[36]	Mammography	Generative Adversarial Network (GAN) + ResNet	--	DDSM		0.89
2019	[37]	Mammography	DCNN-Alex Net	SVM	CBIS-DDSM	87.2	0.94
2019	[38]	Mammography	Deep feature fusion of VGG16, VGG19, Google Net, and Res Net 50	--	CBIS-DDSM	96.6	0.93
2020	[39]	Mammography	Google Net	XGBoost	DDSM	92.8	--
2020	[40]	Mammography	Feature fusion of several Models	SVM, XGBoost, Naïve Bayes, KNN, DT, Ada Boosting	CBIS-DDSM	90.91	--
2021	[27]	Mammography	Deep feature fusion of Alex Net, Google Net, ResNet 18, Res Net 50, ResNet101	SVM	CBIS-DDSM	97.9	1
					MIAS	97.4	1
2021	[41]	Histopathology	VGG19, ResNet34, ResNet50 along with Structural Pruning	--	BreakHis	92.07 (ResNet50)	--
2022	[42]	Mammography	ResNet50, Nasnet-Mobile Network	--	MIAS	89.5 (ResNet50) 70 (NasNet)	--
2022	[43]	Mammography	CoroNet (Based on Xception Net Model)	---	CBIS-DDSM	94.92(4-class) 88.67(2-class)	--
2022	[44]	Histopathology Images	VGG16, VGG19, InceptionResNetV2, DenseNet121, and DenseNet201	--		92.64 (DenseNet121)	--

**Table 8** (continued)

Year	Reference	Imaging Modality	Feature extraction/ DNN Model Adopted	Classification	Dataset	Accuracy (%)	AUC
2023	[45]	Mammography	ResNet50	KNN, Random Forest and SVM	Private Dataset	85 (KNN)	0.89
2023	[46]	Histopathology	K-Means for Segmentation and ResNet18 for feature extraction	SVM	BreakHis	92.6(200x magnification)	--
2023	[47]	Histopathology	3D U Net Model	--	Private Dataset	97	--
2023	Proposed CAD Model (TransNet)	Mammography	<b>First Approach:</b> VGG16, VGG19, ResNet50, ResNet 152	ML Classifiers: KNN, SVM, Random Forest, Ada Boost, XGB	CBIS-DDSM	<b>Best Result: Train Set: 100 (Random Forest &amp; XGB)</b> <i>Test Set: 98 (ResNet152 with Random Forest Classifier)</i>	--
<b>Second Approach:</b> Mobile Net, VGG16, VGG19, Res Net50, Res Net152, Dense Net169				The same Models were used for the Classification	CBIS-DDSM	<b>Best Result: Train Set: 99 Test Set: 98 (Mobile Net)</b>	<b>Best Result: Train Set &amp; Test Set: 0.99 &amp; 0.98 (Mobile Net)</b>

our model using VGG16, VGG19, Mobile Net, Res Net 50, Res Net 152, and Dense Net 169 pre-trained networks. Two experiments were performed: In the first approach, namely Deep feature fusion with ML Classifier Head, pre-trained networks are deployed as feature extractors, and then the derived features are supplied to machine learning classifiers for classification. The second approach, called Deep feature fusion with Neural Net classifiers, utilizes these models for feature extraction and categorization. The findings reveal that KNN and XGB classifiers perform best in the first approach yielding an accuracy of 100% on all the networks in the training phase. On the other hand, ResNet152 outperforms the other pre-trained models by producing 6% increase in accuracy on the test dataset. In the second experiment, Mobile Net, ResNet50, and DenseNet169 performed best in the training and testing phase with Accuracy and AUC of 97%(+2%) and 0.97(+0.02). The minimum loss, however, was exhibited by Mobile Net in both the train and test phases. The second approach performed better than the first, thus, improving all the evaluation metrics. In the future, this framework could be extended for feature fusion of multiple models and could be utilized on other datasets in this field. The proposed method could also be developed on other imaging modalities.

**Data availability** CBIS-DDSM dataset used in this research could be referred in [28]

## Declarations

**Conflict of interest** The authors certify that they have no competing interest.

## References

1. Lai X, Yang W, Li R (2020) DBT Masses Automatic Segmentation Using U-Net Neural Networks. <https://doi.org/10.1155/2020/7156165>
2. Pang T, Wong JHD, Ng WL, Chan CS (2020) Deep learning radiomics in breast cancer with different modalities: Overview and future. *Expert Syst Appl* 158: <https://doi.org/10.1016/j.eswa.2020.113501>
3. GLOBOCAN 2020: New Global Cancer Data | UICC. <https://www.uicc.org/news/globocan-2020-new-global-cancer-data>. Accessed 11 Jun 2022
4. Yadavendra CS (2020) A comparative study of breast cancer tumor classification by classical machine learning methods and deep learning method. *Mach Vis Appl* 31:1–10. <https://doi.org/10.1007/s00138-020-01094-1>
5. Breast cancer. <https://www.who.int/news-room/fact-sheets/detail/breast-cancer>. Accessed 11 Jun 2022
6. Breast Cancer Facts and Statistics. <https://www.breastcancer.org/facts-statistics>. Accessed 21 Jul 2022
7. Krithiga R, Geetha P (2020) Deep learning-based breast cancer detection and classification using fuzzy merging techniques. *Mach Vis Appl* 31: <https://doi.org/10.1007/s00138-020-01122-0>
8. Chugh G, Kumar S, Singh N (2021) Survey on Machine Learning and Deep Learning Applications in Breast Cancer Diagnosis. *Cognit. Comput*
9. Shaikh TA, Ali R, Beg MMS (2020) Transfer learning privileged information fuels CAD diagnosis of breast cancer. *Mach Vis Appl* 31: <https://doi.org/10.1007/s00138-020-01058-5>
10. Sahiner B, Pezeshk A, Hadjiiski LM et al (2019) Deep learning in medical imaging and radiation therapy. *Med Phys* 46:e1–e36. <https://doi.org/10.1002/MP.13264>
11. Gu S, Chen Y, Sheng F et al (2019) A novel method for breast mass segmentation: from superpixel to subpixel segmentation. *Mach Vis Appl* 30:1111–1122. <https://doi.org/10.1007/s00138-019-01020-0>
12. Rahimeto S, Debelee TG, Yohannes D, Schwenker F (2021) Automatic pectoral muscle removal in mammograms. *Evol Syst* 12:519–526. <https://doi.org/10.1007/S12530-019-09310-8>
13. Samala RK, Chan HP, Hadjiiski L et al (2016) Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. *Med Phys* 43:6654–6666. <https://doi.org/10.1118/1.4967345>



14. Li H, Zhuang S, Dao L et al (2019) Benign and malignant classification of mammogram images based on deep learning. *Biomed Signal Process Control* 51:347–354. <https://doi.org/10.1016/j.bspc.2019.02.017>
15. Bevilacqua V, Brunetti A, Guerriero A et al (2019) A performance comparison between shallow and deeper neural networks supervised classification of tomosynthesis breast lesions images. *Cogn Syst Res* 53:3–19. <https://doi.org/10.1016/j.cogsys.2018.04.011>
16. Agarwal P, Yadav A, Mathur P (2022) Breast Cancer Prediction on BreakHis Dataset Using Deep CNN and Transfer Learning Model. *Lect Notes Networks Syst* 238:77–88. [https://doi.org/10.1007/978-981-16-2641-8\\_8](https://doi.org/10.1007/978-981-16-2641-8_8)
17. Hassan NM, Hamad S, Mahar K (2022) Mammogram breast cancer CAD systems for mass detection and classification: a review. *Multimed Tools Appl*
18. Sharma S, Mehra R (2020) Conventional Machine Learning and Deep Learning Approach for Multi-Classification of Breast Cancer Histopathology Images—a Comparative Insight. *J Digit Imaging*. <https://doi.org/10.1007/s10278-019-00307-y>
19. Antropova N, Huynh BQ, Giger ML (2017) A Deep Feature Fusion Methodology for Breast Cancer Diagnosis Demonstrated on Three Imaging Modality Datasets. *Med Phys* 44:5162–5171. <https://doi.org/10.1002/mp.12453>
20. Rana M, Bhushan M (2022) Machine learning and deep learning approach for medical image analysis: diagnosis to detection. *Multimed Tools Appl* 26731–26769. <https://doi.org/10.1007/s11042-022-14305-w>
21. Vijayakumar K, Kadam VJ, Sharma SK (2021) Breast cancer diagnosis using multiple activation deep neural networks. *Concurr Eng Res Appl* 29:275–284. <https://doi.org/10.1177/1063293X211025105>
22. Senan EM, Alsaade FW, Al-Mashhadani MIA et al (2021) Classification of histopathological images for early detection of breast cancer using deep learning. *J Appl Sci Eng* 24:323–329. [https://doi.org/10.6180/JASE.202106\\_24\(3\).0007](https://doi.org/10.6180/JASE.202106_24(3).0007)
23. Al-antari MA, Al-masni MA, Choi MT et al (2018) A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification. *Int J Med Inform* 117:44–54. <https://doi.org/10.1016/j.ijmedinf.2018.06.003>
24. Al-masni MA, Al-antari MA, Park JM et al (2018) Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. *Comput Methods Programs Biomed* 157:85–94. <https://doi.org/10.1016/j.cmpb.2018.01.017>
25. Convolutional Neural Network: An Overview. <https://www.analyticsvidhya.com/blog/2022/01/convolutional-neural-network-an-overview/>. Accessed 11 Jun 2022
26. Deep view on Transfer learning with Image classification Pytorch | by purnasai gudikandula | Medium. <https://purnasaigudikandula.medium.com/deep-view-on-transfer-learning-with-image-classification-pytorch-5cf963939575>. Accessed 11 Jun 2022
27. Ragab DA, Attallah O, Sharkas M et al (2021) A framework for breast cancer classification using Multi-DCNNs. *Comput Biol Med* 131:104245. <https://doi.org/10.1016/j.compbiomed.2021.104245>
28. Lee RS, Gimenez F, Hoogi A, et al (2017) Data Descriptor: A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci Data* 4:. <https://doi.org/10.1038/SDATA.2017.177>
29. Segmenting Abnormalities in Mammograms (Part 2 of 3) | by Cleon W | Towards Data Science. <https://towardsdatascience.com/can-you-find-the-breast-tumours-part-2-of-3-1d43840707fc>. Accessed 11 Jun 2022
30. MobileNet Convolutional neural network Machine Learning Algorithms | Analytics Vidhya. <https://medium.com/analytics-vidhya/image-classification-with-mobilenet-cc6fbb2cd470>. Accessed 11 Jun 2022
31. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition
32. He K, Zhang X, Ren S, Sun J Deep Residual Learning for Image Recognition
33. Steen M, Downe S, Bamford N, Edozien L (2018) DenseNet:Densely Connected Convolutional Networks arXiv:1608.06993v5. Arxiv 28:362–371
34. Tan YJ, Sim KS, Ting FF (2018) Breast cancer detection using convolutional neural networks for mammogram imaging system. *Proceeding 2017 Int Conf Robot Autom Sci ICORAS 2017 2018-March*:1–5. <https://doi.org/10.1109/ICORAS.2017.8308076>
35. Al Nahid A, Mehrabi MA, Kong Y (2018, 2018) Histopathological breast cancer image classification by deep neural network techniques guided by local clustering. *Biomed Res Int*. <https://doi.org/10.1155/2018/2362108>
36. Wu E, Wu K, Cox D, Lotter W (2018) Conditional infilling GANs for data augmentation in mammogram classification. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 11040 LNCS:98–106. [https://doi.org/10.1007/978-3-030-00946-5\\_11](https://doi.org/10.1007/978-3-030-00946-5_11)

37. Ragab DA, Sharkas M, Marshall S, Ren J (2019) Breast cancer detection using deep convolutional neural networks and support vector machines. *PeerJ* 2019:e6201. <https://doi.org/10.7717/PEERJ.6201/TABLE-8>
38. Khan HN, Shahid AR, Raza B, et al (2019) Multi-View Feature Fusion Based Four Views Model for Mammogram Classification Using Convolutional Neural Network. In: IEEE Access. <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8897609>. Accessed 14 Jun 2022
39. Song R, Li T, Wang Y (2020) Mammographic Classification Based on XGBoost and DCNN with Multi Features. *IEEE Access* 8:75011–75021. <https://doi.org/10.1109/ACCESS.2020.2986546>
40. Zhang H, Wu R, Yuan T, et al (2020) DE-Ada \* : A Novel Model for Breast Mass Classification Using Cross-modal Pathological Semantic Mining and Organic Integration of Multi-feature Fusions
41. Choudhary T, Mishra V, Goswami A, Sarangapani J (2021) A transfer learning with structured filter pruning approach for improved breast cancer classification on point-of-care devices. *Comput Biol Med* 134:104432. <https://doi.org/10.1016/j.compbiomed.2021.104432>
42. Alruwaili M, Gouda W (2022) Automated Breast Cancer Detection Models Based on Transfer Learning. *Sensors* 22:1. <https://doi.org/10.3390/s22030876>
43. Mobark N, Hamad S, Rida SZ (2022) CoroNet: Deep Neural Network-Based End-to-End Training for Breast Cancer Diagnosis. *Appl Sci* 12:1. <https://doi.org/10.3390/app12147080>
44. Sujatha R, Chatterjee JM, Angelopoulou A, et al (2022) A transfer learning-based system for grading breast invasive ductal carcinoma. *IET Image Process* 1979–1990. <https://doi.org/10.1049/ipr2.12660>
45. Bai Y, Li M, Ma X, et al (2023) Recognizing breast tumors based on mammograms combined with pre-trained neural networks. *Multimed Tools Appl* 27989–28008. <https://doi.org/10.1007/s11042-023-14708-3>
46. Sahu Y, Tripathi A, Gupta RK, et al (2022) A CNN-SVM-based computer-aided diagnosis of Breast Cancer using histogram K-means segmentation technique. *Multimed Tools Appl* 14055–14075. <https://doi.org/10.1007/s11042-022-13807-x>
47. Khalil S, Nawaz U, Zubariah et al (2023) Enhancing Ductal Carcinoma Classification Using Transfer Learning with 3D U-Net Models in Breast Cancer Imaging. *Appl Sci* 13:1–20. <https://doi.org/10.3390/app13074255>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

## REVIEW

# Unraveling the intricate relationship: Influence of microbiome on the host immune system in carcinogenesis

Saksham Garg | Nikita Sharma | Bharmjeet | Asmita Das 

Department of Biotechnology, Delhi  
Technological University, Delhi, India

**Correspondence**

Asmita Das, Department of Biotechnology,  
Delhi Technological University, Main Bawana  
Road, Shahbad Daulatpur, Delhi 110042, India.  
Email: [asmitadas1710@dce.ac.in](mailto:asmitadas1710@dce.ac.in)

**Funding information**

Delhi Technological University

**Abstract**

**Background:** Cancer is an outcome of various disrupted or dysregulated metabolic processes like apoptosis, growth, and self-cell transformation. Human anatomy harbors trillions of microbes, and these microbes actively influence all kinds of human metabolic activities, including the human immune response. The immune system which inherently acts as a sentinel against microbes, curiously tolerates and even maintains a distinct normal microflora in our body. This emphasizes the evolutionarily significant role of microbiota in shaping our adaptive immune system and even potentiating its function in chronic ailments like cancers. Microbes interact with the host immune cells and play a part in cancer progression or regression by modulating immune cells, producing immunosuppressants, virulence factors, and genotoxins.

**Recent Findings:** An expanding plethora of studies suggest and support the evidence of microbiome impacting cancer etiology. Several studies also indicate that the microbiome can supplement various cancer therapies, increasing their efficacy. The present review discusses the relationship between bacterial and viral microbiota with cancer, discussing different carcinogenic mechanisms influenced by prokaryotes with special emphasis on their immunomodulatory axis. It also elucidates the potential of the microbiome in transforming the efficacy of immunotherapeutic treatments.

**Conclusion:** This review offers a thorough overview of the complex interaction between the human immune system and the microbiome and its impact on the development of cancer. The microbiome affects the immune responses as well as progression of tumor transformation, hence microbiome-based therapies can vastly improve the effectiveness of cancer immunotherapies. Individual variations of the microbiome and its dynamic variability in every individual impacts the immune modulation and cancer progression. Therefore, further research is required to understand these underlying processes in detail, so as to design better microbiome-immune system axis in the treatment of cancer.

**KEYWORDS**

cancer, dysbiosis, gut microbiome, immunotherapy, tumor microenvironment



## 1 | INTRODUCTION

In the last couple of decades, our apprehension of the microbial world has increased significantly. The understanding of microbes especially those associated with the human body became important as several studies suggested the microbes being associated with numerous metabolic activities, including healthy and stress-induced conditions. The microbiota is such a crucial component of the human body that it is seldom referred to as a “forgotten organ”.<sup>1</sup> An estimate of a trillion microbes inhabits human anatomy and our comprehensive understanding and appreciation of the prokaryotic kingdom are attributed to the revolutionizing technologies over time. Recent evidence has proposed that an imbalance of the whole microbial ecosystem (dysbiosis) is responsible rather than a particular microorganism for various conditions. This disturbed microbiome is also recognized in fatal ailments such as cancer.<sup>2,3</sup>

The role of both inflammatory reactions and metabolic products of microbes in cancer was soon acknowledged and was found to be tumor-promoting as well as tumor-suppressing.<sup>4,5</sup> Although the role of microbes in cancer was hypothesized more than a century ago by Virchow,<sup>6</sup> ongoing research demonstrated that about 18% of all cancers in the world are related to infectious agents.<sup>7,8</sup> *Helicobacter pylori*, a common inhabitant of the gut microbiome, is officially given the status of class I carcinogen by International Agency for Research on Cancer (IARC),<sup>9,10</sup> and many other microbes and viruses are under scrutiny for their connection with different types of cancers.<sup>11–14</sup> The relationship is very complex and difficult between cancer and microbes. Generally, cancer is considered to be a disease of host genetics and environmental factors. The microbiota present in the human gut helps in the detoxification of dietary components, reducing inflammation, and maintaining balance in proliferation and host cell growth.<sup>15</sup> However, scientists have given evidence to suggest a strong connection between microbial infections and cancer. The tumor can develop due to inflammatory responses that are induced by bacteria, secretion of bacterial toxins and enzymes, and oncogenic peptides.<sup>16</sup> The present review discusses the fine balance between the microbiome of the body and the triggering of carcinogenesis and also the unique mechanism by which the modulation of the immune system by the microbiome can be cleverly used as a combinatorial therapeutic mechanism for cancer.

## 2 | MECHANISMS CONTROLLING THE HOST INTERACTION WITH MICROBIOTA INVOLVED IN CARCINOGENESIS

The human body harbors different microbial species that have their own set of relationships such as commensalism, parasitism, and mutualism.<sup>17,18</sup> The human gut is a major habitat for an infinite number of bacterial strains and species, showing a symbiotic relationship between humans and microbes. An anatomical barrier that is, primarily the epithelial lining is essential to carry out the smooth functioning of this symbiotic relationship. The multi-level barrier ensures this

separation, and perforation or disruption of this barrier can lead to several diseases and also cancer. In addition to barrier protection, other additional physiological features like mucous layer, low pH, and stratum corneum also contribute to either maintenance of the barrier or directly shaping the microbial community inside the host. The host has also developed some special cells, such as goblet cells, paneth, and keratinocytes, which monitor the bacterial community and actively participate in shaping the microbial population by secreting antibacterial peptides.<sup>19,20</sup> Elements of the immune system are also a part of the barrier system for protection. The invading bacteria also respond in order to survive and protect themselves in the hostile environment. The mucin layer and bacteriocins production are the main mechanisms that bacteria implement to proliferate in a hostile environment.<sup>21</sup> In the competition for resources, bacteria produce bacteriocins, which in turn are helpful for the human too as bacteriocins suppress the pathobionts, thus limiting the pathogenicity of the microbes.<sup>22</sup> A supportive narrative is given by the fact that germ-free mice, showed increased susceptibility to the invading bacteria and infections.<sup>23</sup>

As is often observed, the microbiota-host interaction is hanging by a string of equilibrium, the disturbance in this equilibrated system is the cause that makes micro-organisms cause disease. The failure is analogous to the domino effect; the defective working of even one mechanism (including both host and microbial) is followed by a defect in another mechanism. With all the dominos falling, overall equilibrium is disrupted, which is the essential cause for many diseases and often carcinogenesis too. *H. pylori* is a good example of the domino effect as its infection of the host harms not only the barrier cells but also causes increased inflammation, disrupts the microbiota causing dysbiosis, and also changes the gastric environment.<sup>24</sup> The relationships between certain microorganisms and different kinds of cancer are shown in Table 1. This table emphasizes the microbial dysbiosis seen in many tumors and offers details on the possible involvement of these bacteria in the development of cancer.

## 3 | MICROBIOME INFLUENCING HOST IMMUNE RESPONSE

For the survival of the individual organism or species, developing defense mechanisms against diseases is a crucial part of evolution. The crosstalk between microbiome and human cells plays a vital role in the development of both innate and acquired immunity.<sup>45</sup> It is estimated that a human host trillions of microbes at various sites in the body.<sup>46</sup> This results in extensive interactions and the immune system develops itself to attack the harmful microorganisms.<sup>47</sup>

Right after birth, the immune system begins to take shape, and it continues to mature over the course of a person's lifespan. In just a few weeks after birth, the developing newborn serves as the host to the whole microbial population. The research indicated that nursing on mother's milk is a crucial source for gut microbiota colonization in a newborn, while it is unclear exactly what should characterize the initial encounter with the microorganisms.<sup>48</sup> Specific microbial strains

**TABLE 1** Various microbes involved in different types of cancers.

S. No.	Type of cancer	Influencing pathogen	Mode of action	References
1.	Lung cancer	<i>Bacillus</i> sp.	Chronic inflammation and the generation of metabolites that can lead to cancer	25
		<i>Mycoplasma</i> sp.	Triggering persistent inflammation, altering immunological responses, and perhaps having an impact on carcinogenic pathways	
		<i>Staphylococcus epidermis</i>	Indirectly influence immunological responses and the general makeup of the lung microbiota, which in turn may lead to lung cancer.	
		<i>Capnocytophaga</i>	By aspirating oral infections into the lungs and causing subsequent chronic inflammation	26
		<i>Veillonella</i>	Associated with respiratory infections	
		<i>Klebsiella</i>	Exacerbate tumor growth, compromise immunological function, and cause persistent inflammation	27,28
		<i>Moraxella catarrhalis</i>	Having it in the respiratory system may cause persistent inflammation and encourage the growth of lung cancer	
2.	Pancreatic cancer	<i>H. pylori</i>	Via oncogenic signaling cascade activation, persistent inflammation, and host immune response modification	29–31
		<i>Pseudomonas aeruginosa</i>	Through the generation of virulence factors, activation of chronic inflammation, and possibly disruption of anticancer immune responses	32
		<i>Fusobacterium</i> species	Provoke chronic inflammation, modify the tumor microenvironment, encourage cell invasion, and affect immune responses	33
		<i>Streptococcus mitis</i>	Encouraging immunological dysregulation, and inflammation	34,35
		<i>Porphyromonas gingivalis</i>	Cause prolonged inflammation, and interfere with the immunological response	
		<i>Neisseria elongate</i>	Immunological dysregulation and persistent inflammation.	36
3.	Gall bladder cancer	<i>Salmonella typhi</i>	Chronic inflammatory response resulting in DNA damage and cell growth	37
		<i>Helicobacter pylori</i>	Oncogenic signaling pathways and chronic inflammation.	38
4.	Ovarian cancer	<i>Mycoplasma</i> sp.	Immune control, long-term inflammation, and direct contact with ovarian cells	39
		<i>Chlamydia pneumonia</i>	Immune control, long-term inflammation, and direct contact with ovarian cells	40
5.	T-cell leukemia	Human T-cell leukemia virus-1	DNA insertion into the host cell, cellular function impairment, and cell division	41–43
6.	Nasopharyngeal cancer	Epstein–Barr virus	Epithelial cell infection, transformation, and immune response evasion	
7.	Kaposi sarcoma	Kaposi sarcoma-associated herpesvirus	Immune and endothelial cell infection, as well as growth factor and pro-inflammatory cytokine release	
8.	Merkel cell carcinoma	Simian virus 40	DNA insertion into the host cell, interference with cellular control, impairment of cell growth and survival	
		Merkel cell polyomavirus	DNA insertion into the host cell, interference with cell cycle regulation, and cellular transformation	
9.	Skin cancer	Merkel cell polyomavirus	Cellular transformation, viral oncogene expression, and integration into the DNA of the host cell	44

like *Lactobacillus*, *Enterococcus*, *Bifidobacterium*, and *Staphylococcus*, are commonly found in breast milk and infant stool samples.<sup>49,50</sup> Moreover, the similarity of bacterial strains in the stool sample has been shown to be increasing proportionately with an increased intake of breast milk.<sup>51</sup> These studies suggested a clear transfer of microbes from breast milk to

the infant's gut, which is inevitable and dynamic. After initial exposure to the prokaryotic kingdom, the interaction and interplay increase with time thus, paving the path for development and increasing the tolerance of the human immune system. Important immune-modulating effects of the microbiome on cancer development are shown in Table 2.

**TABLE 2** Key immune-modulating effects of the microbiome.

Immune modulation	Mechanism
Tumor immune surveillance	APC activation and T cell reactions
Regulatory T cell function	Inhibition of regulatory T cell induction
Inflammation	Modulation of inflammatory signals both pro and anti
Immune checkpoint regulation	Influence on the expression and activation of immunological checkpoints
Response to immunotherapy	The effectiveness and responsiveness to immunotherapies are affected

With the discovery of pattern-recognition receptors (PRRs) and our growing understanding of the complex interaction of immune cells with the microbial community, humans have realized that despite the capability of our immune system to selectively destroy all prokaryotic microbes evolution has taught our immune system to tolerate the harboring of the normal flora. The PRR superset consists of a subset-families with known characteristic pathways for each subset including Toll-like receptor (TLRs), nucleotide-binding oligomerization (NOD)-like receptor (NLRs), the RIG-I-like receptor, the C-type lectin receptor, the absent in melanoma 2 (AIMS2)-like receptors, and the OAS-like receptors.<sup>52</sup> These receptors produce a specific response to an active stimulus. This sensing system of multiple levels and components has an analogy with a rheostat; according to the processed information at different levels of physiology, the appropriate adjustments are made to fit the host with the neighboring microbial ecosystem.<sup>45</sup> At times, disruption in this intricate balance between humans and microbes is observed and called dysbiosis. In a nutshell, dysbiosis is an unstable microbiota that can lead to opportunistic infections.<sup>53</sup> External agents such as probiotics are often supplemented for the re-establishment of beneficial microbes or eubiosis, which strengthens the immune responses to treat dysbiosis.<sup>54</sup> Overall, the immune system can be considered as a buffer to metabolic activities and foreign interactions as it configures both the response against and for the proliferation of microbial community by promoting the growth of microbiota that is helpful to the host and simultaneously shaping the existing microbiota, which is tolerable by the host.

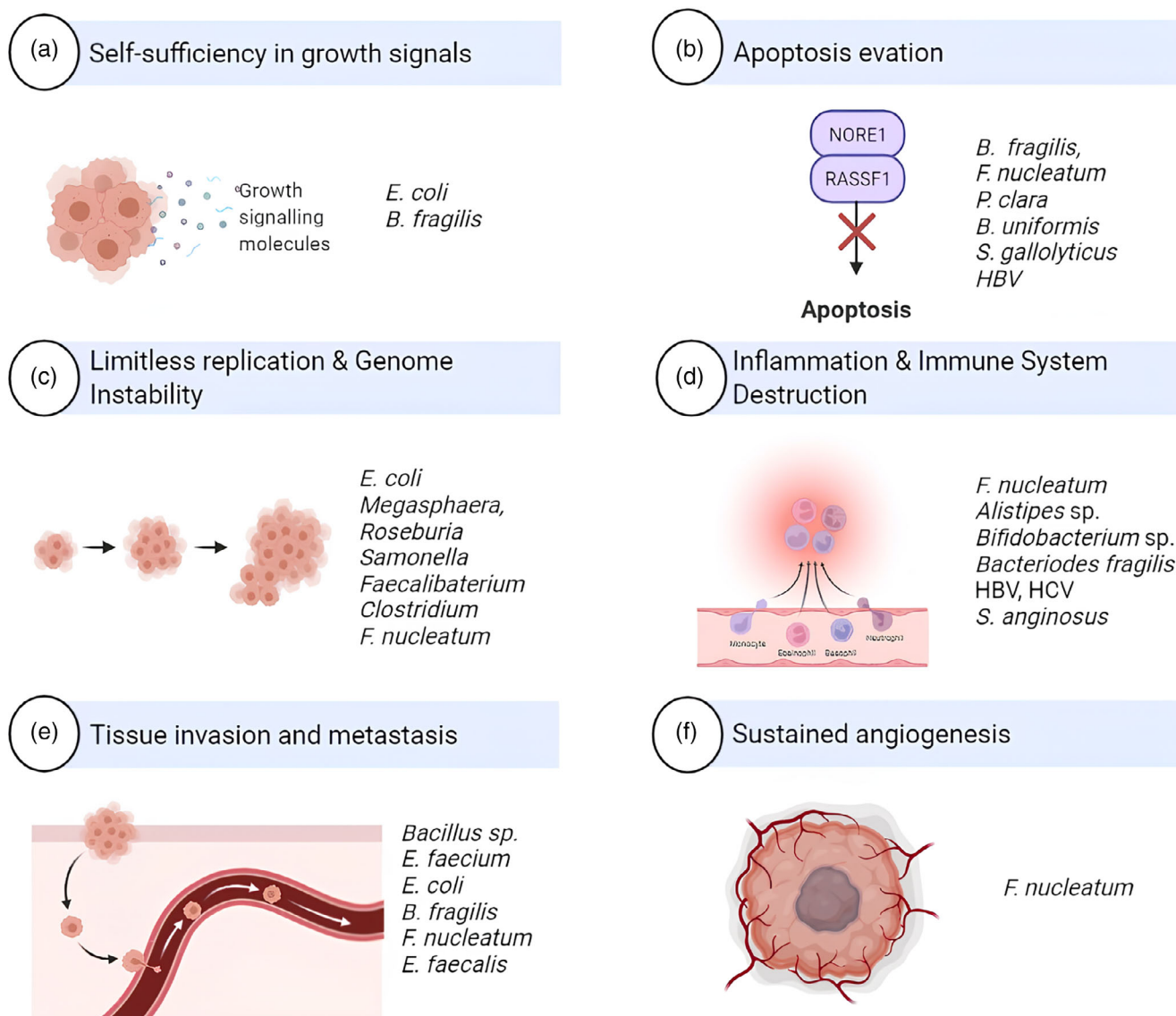
TLRs are a component of the innate immune system and can produce high-intensity pro-inflammatory stimuli,<sup>55</sup> depending on the pathway taken by the TLRs and the bacterial cell MAMPs they recognize, can either promote or protect against carcinogenesis. While TLR2 promotes gastric cancer,<sup>56</sup> lipopolysaccharide and TLR4, increase colorectal, liver, pancreatic, and skin cancers.<sup>57–60</sup> TLRs are connected to cancer formation and survival, but it is not apparent if a direct or indirect role is being played. In order to promote pro-carcinogenic pathways, NF- $\kappa$ B and STAT3 are activated,<sup>56,57,61</sup> and mitogens such as amphiregulin, epiregulin, and hepatocyte growth factor (HGF) produced by TLR-expressing fibroblast cells have cancer-causing effects.<sup>57,61–63</sup> Microflora-induced TLR activation on myeloid cells results in the activation of cancer-causing pathways such as IL-17 and IL-23.

The carcinogenic nature of these cytokine networks is confirmed by the evidence that decreased IL-17 and IL-23 expression either by genetic suppression of *Tlr2*, *Tlr4*, *Myd88*, or *Tlr9* or rooting out the microflora by antibiotics, resulted in the reduced propensity for carcinogenesis in tissues.<sup>64,65</sup> As demonstrated by the case of MYD88, where MYD88 inactivation suppresses cancer but may potentially result in colorectal cancer (CRC) linked to the damaged epithelium, TLRs also play a protective function.<sup>59,66</sup> Through the downregulation of pro-inflammatory IL-22 and promotion of epithelial damage repair, the protective IL-18 pathway, which is controlled by the MYD88 protein, confers a protective character to the colonic epithelium. The development of colitis-associated cancer (CAC) is shown in IL-18-deficient mouse models, which is equivalent to the outcomes of *Myd88*<sup>-/-</sup> mice models.<sup>67</sup> Therefore, the absence of an IL-18 pathway might be the cause of colon cancer. A subgroup of the PRR superfamily called NOD-like receptors (NLRs) has a NACHT domain situated at the center, which is crucial for oligomerization and leucine-rich repeats are present toward C-terminal facilitating ligand identification.<sup>68,69</sup> A group of NLRs known as NLRP has a pyrin domain (PYD) out of which NLRP6 has been extensively studied in host-microbiome interaction and is best known for inducing inflammasomes.<sup>70</sup> and contributes to immunity against bacteria, which is represented by dysbiosis in *Nlrp6*<sup>-/-</sup> mice, making them more prone to colitis and CRC progression.<sup>71–73</sup> The proposed molecular mechanism in *Nlrp6* deficit mice is the reduced stimulus to inflammasomes and IL-18 production. Greater susceptibility to CRC is shown by *Asc*<sup>-/-</sup> and IL-18 deficient mice confirming the proposed mechanism.<sup>70</sup> A muramyl dipeptide-sensing NLR called NOD2 aids in preserving immunity against microorganisms.<sup>74</sup> Dysbiosis brought on by mutations in NLRP6 and NOD2 can result in CRC.<sup>75,76</sup> umari et al showed that the intestine of *Nod2*<sup>-/-</sup> mice model is usually found in a state of dysbiotic distress.<sup>77</sup> All these pieces of evidence shift the belief that the involvement of NOD2 in cancer is via dysbiosis as this increased cancer propensity was seen to be transferable too in co-housing.<sup>76</sup> Both NOD2 and NLRP6 share IL-6<sup>78</sup> as a common mediator, and treatment with IL-6-neutralizing antibodies to knockout models reduces the development of CRC.<sup>70,76</sup> By promoting inflammation and inducing CRC, NOD1 deficiency also influences carcinogenesis. More research is required to determine the precise mechanism behind the carcinogenic impact of other NLR family members.

#### 4 | INFLUENCE OF MICROBIOTA ON CARCINOGENESIS THROUGH VIRULENCE FACTORS AND GENOTOXINS

Microbes are extremely important for the development, progression, and even response to therapy of cancer. New research reveals that the human microbiome, which consists of a wide variety of bacteria living in and on our bodies, might affect the development and course of cancer. At the beginning of the disease, some bacteria can directly cause DNA damage or encourage persistent inflammation, fostering conditions that allow healthy cells to develop into malignant cells. Through a variety of processes, including immune system regulation,





**FIGURE 1** Microorganisms at different stages of cancer.

modification of the tumor microenvironment, and creation of chemicals that support tumor cell survival and proliferation, the microbiome can influence tumor development, invasion, and metastasis as cancer progresses. Furthermore, the microbiome can change the way drugs are metabolized, regulate immune responses, and possibly affect how well immunotherapy works. Figure 1 emphasizes how bacteria play a role in the initiation of DNA damage, the promotion of chronic inflammation, the influence on the growth and spread of tumors, and the modulation of therapeutic responses.

Virulence factors are transcribed products that mediate the establishment of a pathogen in the host cell, increasing its pathogenicity. These may include bacterial toxins, hydrolytic enzymes, surface protein, and carbohydrates.<sup>79</sup> Some factors such as vacuolating cytotoxin A (VacA), cytotoxin-associated gene A (CagA), and FadA have been extensively researched for their potential to cause cancer.<sup>6</sup> CagA and VacA are developmental agents of gastric cancer in mice models.

FadA mediates the adhesion of bacterium to the host cell. However, FadA has been observed to interact with other compounds too, such as E-cadherin activating  $\beta$ -catenin signal leading to CRC.<sup>80,81</sup>

Metabolism of xenobiotics is considered to play a crucial role in gut microflora.<sup>82</sup> Such metabolism affects the activity and toxicity of the drugs and therapeutic compounds that are administered. One example is the case of irinotecan (CPT-11), a drug used for the treatment of colon and pancreatic cancer.<sup>83</sup> In the liver, the metabolism of irinotecan terminates by its inactivation, but in the presence of  $\beta$ -glucuronidase, the compound is reactivated and thus contributes to additional severe side effects like diarrhea, which also limit the potency of the drug. The inactivation of bacterial enzymes or treatment with antibiotics are some of the measures that are taken to reduce these complications and increase the potency of the particular drug molecule.<sup>84</sup>

Microflora present in the intestine regulates bile metabolism using hydrolase enzymes as the medium of action. Deoxycholic acid



(DCA) is one of the modulated bile salts. In a diet model with high-fat content, DCA is an appurtenance to another Hepatocellular carcinoma (HCC) causing agents. Diet changes and reduction in DCA-producing bacteria using antibiotics decrease the chances of developing HCC. DCA is being implicated in colon and esophageal cancers also,<sup>85–87</sup> suggesting bacterial involvement in a multitude of cancers.

Around 3.6% of all cancers, including the rectum, pharynx, oral cavity, colon, female breast, esophagus, and liver, are implicated by alcohol, and the breakdown of alcohol to acetaldehyde is largely dependent on the microbiome present.<sup>88</sup> Observations in germ-free mice revealed a significantly lower concentration of acetaldehyde, thereby showing genotoxic effects and disease promotion. This conversion is also hypothesized to be crucial in oral cavity cancers as the concentration of acetaldehyde in the oral cavity is 10–100 times higher than in blood vessels.<sup>88,89</sup>

Additionally, DNA instability and genotoxicity caused by bacterial toxins and metabolic products can result in cancer.<sup>90</sup> Bacterial toxins such as cytolethal distending toxin (CDT), colibactin, and cytotoxic necrotizing factors contribute to DNA malfunctioning and cytotoxicity disrupting the mediating cellular pathways. Bacteria-induced inflammation is synchronized with reactive oxygen species (ROS) production, which also contributes to genotoxicity and has the potential to form tumors.<sup>65,91–94</sup> A toxin produced by *Bacteroides fragilis* (*B. fragilis*) that causes bowel inflammatory diseases has been linked to CRC neoplasia and cancer.<sup>95,96</sup> Infection with *Salmonella* sp. has been reported to have a genotoxic effect by suppressing the adenomatous polyposis coli (APC) gene, a tumor suppressor gene thus, increasing the risk of generation and proliferation of cancer.<sup>97</sup> Both CDT and colibactin toxins exert extensive DNA damage by ataxia-telangiectasia mutated (ATM)-CHK2 pathway and histone H2AX phosphorylation leading to seizing cell cycle and swollen cells.<sup>6</sup>

CDT genotoxin can cause various cancers, including CRC, gastric, and gallbladder.<sup>98</sup> CDT is an AB toxin with three subunits as CdtA, CdtB, and CdtC. CdtB is the only subunit that enters the cell cytoplasm, from where the active subunit invades the nucleus and confers DNA damage.<sup>92</sup> The damaging subunit is highly homologous to the DNase I site, and this homology results in reduced damage control of DNA by the normal functioning enzymes.<sup>92,99</sup> In nuclear factor- $\kappa$ B (NF- $\kappa$ B) lacking mice models, the CdtB mutant bacteria failed to induce any cell growth in the intestine. Administration of the same bacteria to IL10<sup>−/−</sup> mice models also failed to show any signs of dysplasia. These studies concluded that CDT-mediated DNase activity is essential for certain groups of bacteria to form tumors and can be potentially carcinogenic.<sup>100,101</sup>

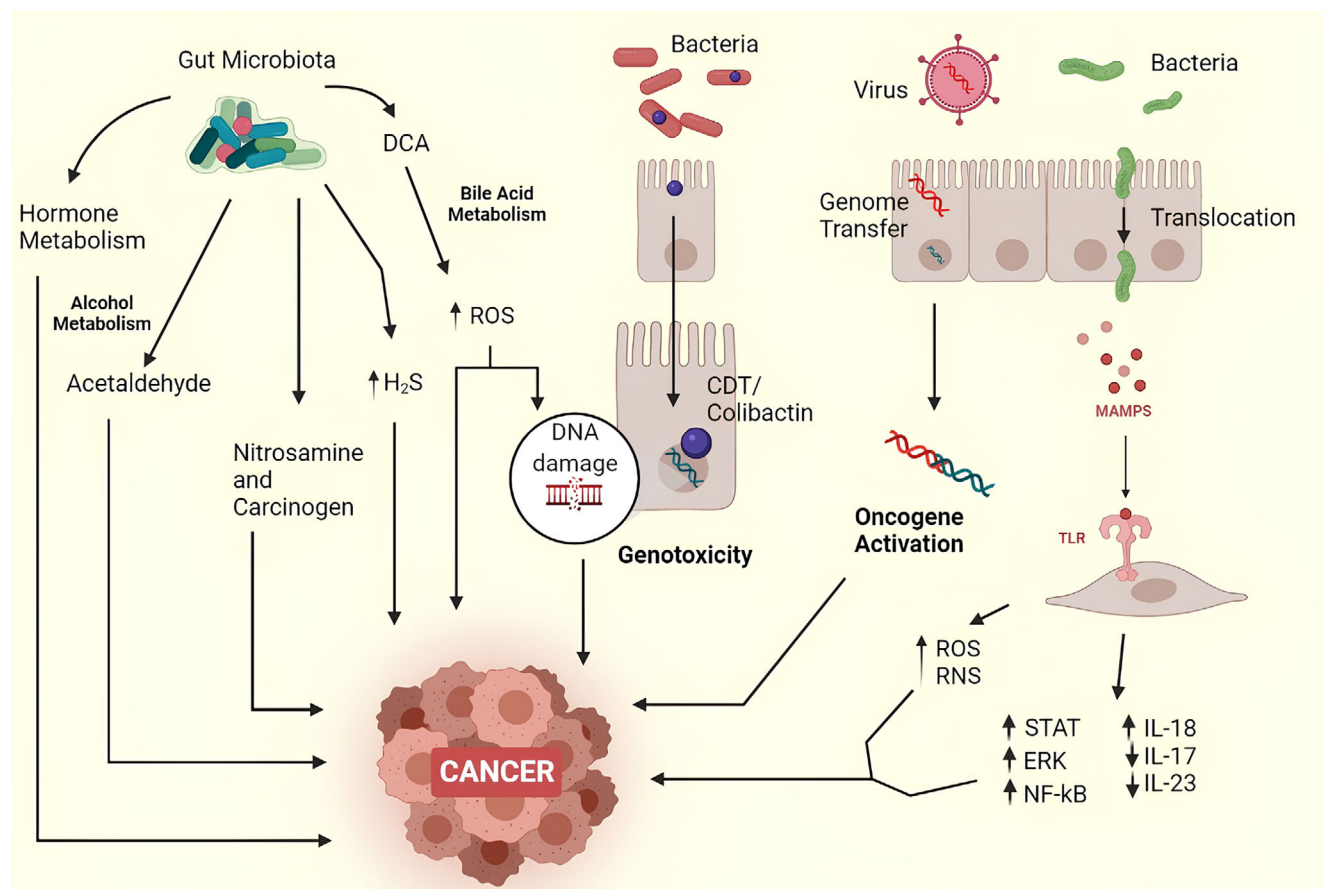
Colibactin is encoded by 54 kb polyketide synthase (PKS).<sup>102</sup> Some studies have linked the toxin with CRC development using the IL10<sup>−/−</sup> mice model. *E. coli* with functionally active PKS islands have been isolated from the developing CRC tissues.<sup>91,103</sup> Colibactin is formed as a result of the inclusive activity of eight out of nine accessory genes and also includes non-ribosomal peptide synthetase (NRPS). All these components are essential to produce functional colibactin. It is also hypothesized that carcinogenicity can be mediated by DNase. This hypothesis is supported by the study, which concluded

that the structural integrity of DNA was affected in the cells having *pks* active bacterial strain.<sup>102</sup> As toxin is rather newly observed, the exact direct relationship of the toxin needs further explanations and pieces of evidence. A thorough explanation of how microorganisms might contribute to cancer progression is given in Figure 2. The figure illustrates important microbiological pathways including persistent inflammation, genomic changes, immunological regulation, and synthesis of chemicals that support tumor growth.

At times, metabolic products released by the bacteria also contribute to genotoxicity and instability of DNA.<sup>104,105</sup> Prime examples of such cases are the production of hydrogen sulfide, superoxide compounds, and N-nitroso compounds (NOC). Diet in such cases is a crucial point to consider as the gut microflora widely is dependent on the type of dietary components consumed.<sup>106</sup> Sulfate-reducing bacteria (SRB) are the cause of the H<sub>2</sub>S released in the gut, which is a genotoxic and cytotoxic gas.<sup>104,107,108</sup> It damages the DNA as well as reduces ATP formation.<sup>109</sup> SRB is a constituent part of gut microbiota. However, a diet with sulfur-containing amino acids or sulfur-polluted water can increase the concentration of SRB thereby, increasing the risk of CRC.<sup>106,110</sup> Superoxide radicals are observed to promote tumor formation in IL10<sup>−/−</sup> mice, specifically in the gut.<sup>111,112</sup> NOC shows alkylation of DNA, therefore, transitioning GC to AT causes major DNA damage in the K-ras gene promoting CRC.<sup>113</sup> Nitrogen-reducing bacteria are again a part of normal gut microflora, but consumption of red meat diet increases the requirement of nitrogen-reducing bacteria and, in turn, increases the metabolic NOC.<sup>114,115</sup> Detoxification and microflora management by control of diet and elimination process genotoxic compounds are likely to excrete out and accordingly affect carcinogenesis.

## 5 | MICROBES AS PERSONALIZED CANCER TREATMENT

For the past 10 years, researchers have studied microbes as a therapy or in conjunction with other medicines. However, Dr. Coley first identified their application as anti-cancer drugs in the 19th century, which is when their usage as cancer therapies began. The basis for the use of microorganisms in cancer therapy was laid by an experiment involving the injection of *Streptococcus pyogenes* in a patient with bone sarcoma that showed promising tumor shrinkage.<sup>116,117</sup> The use of microorganisms in the treatment of cancer is currently receiving fresh attention. Because various types of cancers exhibit heterogeneity and patients respond differently to therapies, precision medicine is an attractive cure for cancer. Table 3 shows that a number of clinical trials have looked at the potential of microbiome-based therapies in the treatment of cancer. The research looked at several treatments, including fecal microbiota transplantation (FMT) for colorectal cancer, probiotic supplements for breast cancer, and the effect of antibiotic medication on the effectiveness of immunotherapy for lung cancer. By influencing the immune system, particularly inflammation pathways, microbe-wide genomic investigations have demonstrated the significance of bacteria in both well-being and disease. Despite the challenges involved in the process, recognizing this effect is essential



**FIGURE 2** Microbial mechanisms in cancer progression.

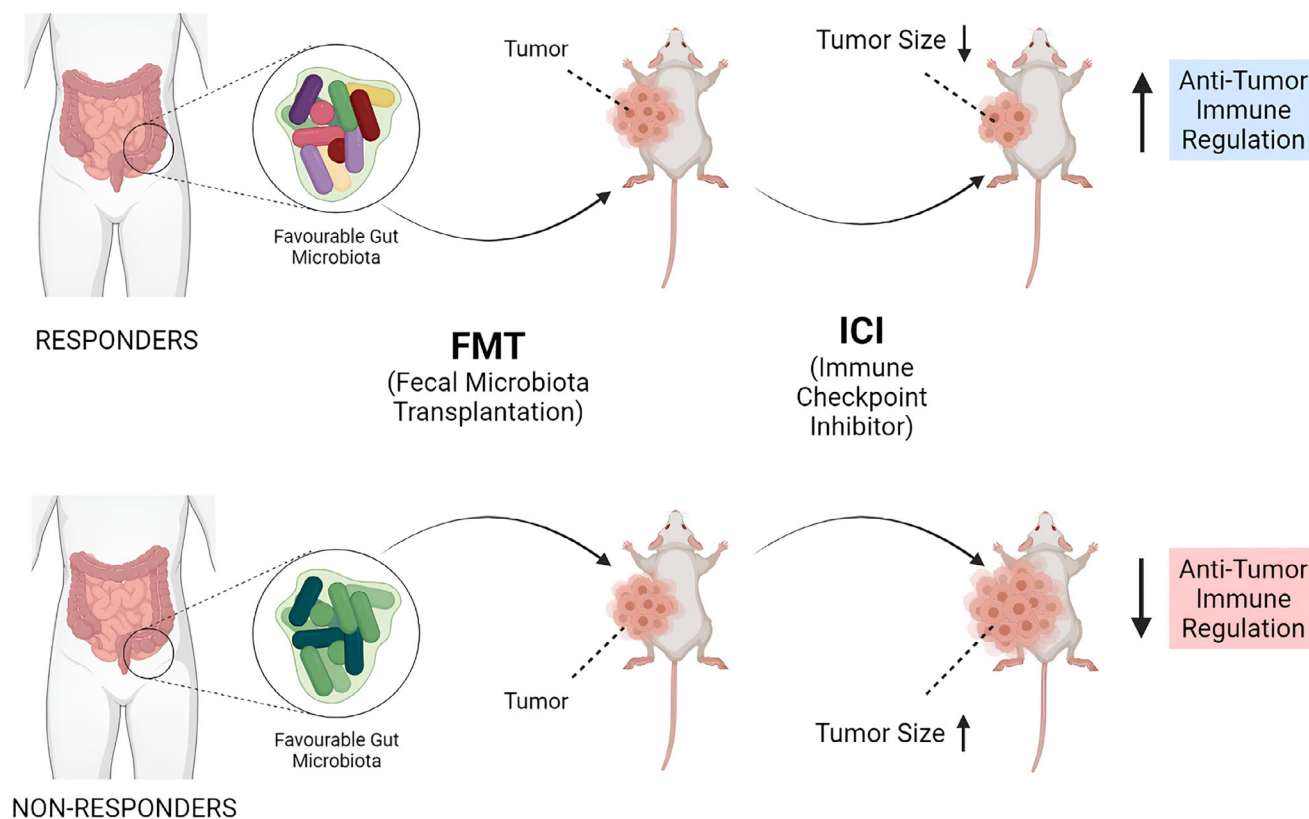
**TABLE 3** Clinical trials examining cancer treatments based on the microbiome.

Cancer type	Intervention	Study design	Findings
Colorectal cancer	FMT	Randomized controlled trial	FMT decreased tumor development and increased survival.
Lung cancer	Antibiotic treatment	Retrospective cohort study	Use of antibiotics was linked to a decreased response to immunotherapy.
Breast cancer	Probiotic supplementation	Prospective interventional study	Probiotics increased the variety of bacteria in the stomach and decreased adverse effects from therapy.
Pancreatic cancer	Gut microbiota modulation	Randomized controlled trial	Patient survival and treatment responsiveness were both enhanced by gut microbiota modification.
Prostate cancer	Prebiotic supplementation	Single-arm pilot study	Supplementing with prebiotics has the potential to alter the gut flora and lower inflammation.
Hepatocellular carcinoma	Microbiota targeted therapy	Pilot clinical trial	Therapy that targets the microbiota has the potential to enhance patient outcomes and liver function.
Ovarian cancer	Probiotic supplementation	Case-control study	Probiotics improved treatment outcomes

for improving outcomes, and identifying targets, biomarkers as well as diagnostics.<sup>118</sup>

The effectiveness of traditional cancer treatments has been enhanced with the use of combinatorial cancer medicines,<sup>119</sup> and microbial therapies have been found to be effective when combined.

Microbes and the immune system are intertwined, and they can affect how well immunotherapeutic treatments work. According to studies, the resident gut microbiota, which includes “favorable” and “unfavorable” microorganisms, might influence how differently people respond to the same medicine (Figure 3).



**FIGURE 3** Microbial diversity and population present in the gut influence the regulation and response of immune checkpoint inhibitors (ICI).

By eliciting pro-inflammatory immune responses, *Bacteroides fragilis* and *Bacteroides thetaiotaomicron*, for instance, have been demonstrated to improve the effectiveness of CTLA-4-based blocking treatment.<sup>120</sup> Fecal microbiota transplant FMT has been shown to increase anti-tumor immunity and the overall effectiveness of PD-L1 inhibition by reducing tumor development and boosting T-cell responsiveness in non-responding models.<sup>121</sup> Patients with microbiota enriched in *Faecalibacterium* exhibited better results, which is indicative of the relative variety of microorganisms in the gut.<sup>122,123</sup> These results have been shown in solid tumors such as urothelial, renal cell, and lung carcinomas, where the presence of a lot of *Akkermansia muciniphila* favored a better prognosis.<sup>124</sup>

Since the microbial populations in the gut can enhance current treatments, such as immune checkpoint inhibition, they can be investigated for potential therapeutic benefits in cancer as well as other chronic illnesses. By identifying and modifying the gut microbiota, personalized medicine can be designed using sequencing and FMT that may improve cancer therapies augmentatively.

## 6 | CONCLUSION

The complex interaction between the microbiome and the human immune system during carcinogenesis is now widely acknowledged as a key element in the onset and spread of cancer. Through a variety of processes, the microbiome affects the host immune system, having an

impact on immune cell activity, immunological signaling pathways, and the tumor microenvironment. Dysbiosis and predominance of certain microbial communities have been linked to an increased risk of developing cancer, highlighting the significance of the makeup of the microbiome in the development of cancer.

Significant impacts on cancer prevention, diagnosis, and therapy may result from an understanding of the intricate interactions between the microbiota and the human immune system. An intriguing direction for future study is using the microbiome to modify immune responses and improve the efficiency of immunotherapies. Innovative treatment approaches to enhance patient outcomes can be created by using the microbiome-immune system axis. However, there are a number of issues that must be resolved. Further research is necessary to fully understand the processes by which the microbiome affects immune responses and the development of cancer. Additionally, it is difficult to pinpoint specific microbial signatures linked to cancer because of the complexity of the microbiome makeup and the inter-individual variability.

Nevertheless, the role of the microbiome-immune system nexus in the treatment of cancer cannot be overstated. Future studies should concentrate on elucidating the complex processes behind this link, creating standardized microbiome analysis procedures, and carrying out clinical trials to evaluate the effectiveness of microbiome-based treatments in preventing cancer and therapy. In conclusion, the impact of the microbiome on the human immune system in the development of cancer is an exciting field of study with broad ramifications.



A changeable element that can be addressed for customized cancer treatments is the microbiome. Continued research in this area will surely improve our knowledge of cancer biology and open the door to cutting-edge methods for treating this deadly condition.

## AUTHOR CONTRIBUTIONS

**Saksham Garg:** Data curation (lead); writing – original draft (lead). **Nikita Sharma:** Writing – original draft (supporting). **Bharmjeet:** Writing – original draft (equal). **Asmita Das:** Conceptualization (lead); project administration (lead); writing – review and editing (lead).

## ACKNOWLEDGMENTS

All studies have been conducted in Delhi Technological University with funding provided to Dr. Asmita Das.

## CONFLICT OF INTEREST STATEMENT

The authors have stated explicitly that there are no conflicts of interest in connection with this article.

## DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

## ETHICS STATEMENT

The present work has been done in compliance with the ethical guidelines of Delhi Technological University.

## ORCID

Asmita Das  <https://orcid.org/0000-0001-9846-1005>

## REFERENCES

- O'Hara AM, Shanahan F. The gut flora as a forgotten organ. *EMBO Rep.* 2006;7(7):688-693. doi:10.1038/sj.embor.7400731
- Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JL. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature.* 2006;444(7122):1027-1031. doi:10.1038/nature05414
- Smith MI, Yatsunenkov T, Manary MJ, et al. Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. *Science (New York, N.Y.).* 2013;339(6119):548-554. doi:10.1126/science.1229000
- Colditz GA, Sellers TA, Trapido E. Epidemiology – identifying the causes and preventability of cancer? *Nat Rev Cancer.* 2006;6(1):75-83. doi:10.1038/nrc1784
- Peto J. Cancer epidemiology in the last century and the next decade. *Nature.* 2001;411(6835):390-395. doi:10.1038/35077256
- Schwabe RF, Jobin C. The microbiome and cancer. *Nat Rev Cancer.* 2013;13(11):800-812. doi:10.1038/nrc3610
- Balkwill F, Mantovani A. Inflammation and cancer: back to Virchow? *Lancet (London, England).* 2001;357(9255):539-545. doi:10.1016/S0140-6736(00)04046-0
- Trinchieri G. Cancer and inflammation: an old intuition with rapidly evolving new concepts. *Annu Rev Immunol.* 2012;30:677-706. doi:10.1146/annurev-immunol-020711-075008
- Huang JQ, Sridhar S, Chen Y, Hunt RH. Meta-analysis of the relationship between *Helicobacter pylori* seropositivity and gastric cancer. *Gastroenterology.* 1998;114(6):1169-1179. doi:10.1016/S0016-5085(98)70422-6
- Parkin DM. The global health burden of infection-associated cancers in the year 2002. *Int J Cancer.* 2006;118(12):3030-3044. doi:10.1002/ijc.21731
- Burnett-Hartman AN, Newcomb PA, Potter JD. Infectious agents and colorectal cancer: a review of *Helicobacter pylori*, *Streptococcus bovis*, JC virus, and human papillomavirus. *Cancer Epidemiol Biomarkers Prev.* 2008;17(11):2970-2979. doi:10.1158/1055-9965.EPI-08-0571
- Castellari M, Warren RL, Freeman JD, et al. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res.* 2012;22(2):299-306. doi:10.1101/gr.126516.111
- Goodwin AC, Destefano Shields CE, Wu S, et al. Polyamine catabolism contributes to enterotoxigenic *Bacteroides fragilis*-induced colon tumorigenesis. *Proc Natl Acad Sci U S A.* 2011;108(37):15354-15359. doi:10.1073/pnas.1010203108
- Marchesi JR, Dutilh BE, Hall N, et al. Towards the human colorectal cancer microbiome. *PLoS One.* 2011;6(5):e20447. doi:10.1371/journal.pone.0020447
- Garrett WS. Cancer and the microbiota. *Science (New York, N.Y.).* 2015;348(6230):80-86. doi:10.1126/science.aaa4972
- Laliani G, Ghasemian Sorboni S, Lari R. Bacteria and cancer: different sides of the same coin. *Life Sci.* 2020;246:117398. doi:10.1016/j.lfs.2020.117398
- Bäckhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JL. Host-bacterial mutualism in the human intestine. *Science (New York, N.Y.).* 2005;307(5717):1915-1920. doi:10.1126/science.1104816
- Dethlefsen L, McFall-Ngai M, Relman DA. An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature.* 2007;449(7164):811-818. doi:10.1038/nature06245
- Salzman NH, Underwood MA, Bevins CL. Paneth cells, defensins, and the commensal microbiota: a hypothesis on intimate interplay at the intestinal mucosa. *Semin Immunol.* 2007;19(2):70-83. doi:10.1016/j.smim.2007.04.002
- Nestle FO, Di Meglio P, Qin J-Z, Nickoloff BJ. Skin immune sentinels in health and disease. *Nat Rev Immunol.* 2009;9(10):679-691. doi:10.1038/nri2622
- van Nood E, Vrieze A, Nieuwdorp M, et al. Duodenal infusion of donor feces for recurrent *Clostridium difficile*. *N Engl J Med.* 2013;368(5):407-415. doi:10.1056/NEJMoa1205037
- Cornforth DM, Foster KR. Competition sensing: the social side of bacterial stress responses. *Nat Rev Microbiol.* 2013;11(4):285-293. doi:10.1038/nrmicro2977
- Kamada N, Kim YG, Sham HP, et al. Regulated virulence controls the ability of a pathogen to compete with the gut microbiota. *Science (New York, NY).* 2012;336(6086):1325-1329. doi:10.1126/science.1222195
- Fox JG, Wang TC. Inflammation, atrophy, and gastric cancer. *J Clin Invest.* 2007;117(1):60-69. doi:10.1172/JCI30111
- Apostolou P, Tsantsaridou A, Papasotiriou I, Toloudi M, Chatziioannou M, Giamouzis G. Bacterial and fungal microflora in surgically removed lung cancer samples. *J Cardiothorac Surg.* 2011;6(1):137. doi:10.1186/1749-8090-6-137
- Yan X, Yang M, Liu J, et al. Discovery and validation of potential bacterial biomarkers for lung cancer. *Am J Cancer Res.* 2015;5(10):3111-3122.
- Lin FC, Huang JY, Tsai SC, et al. The association between human papillomavirus infection and female lung cancer. *Medicine.* 2016;95(23):e3856. doi:10.1097/MD.0000000000003856
- Greathouse KL, White JR, Vargas AJ, et al. Interaction between the microbiome and TP53 in human lung cancer. *Genome Biol.* 2018;19(1):123. doi:10.1186/s13059-018-1501-6
- Rabelo-Gonçalves EM. Extragastric manifestations of *Helicobacter pylori* infection: possible role of bacterium in liver and pancreas diseases. *World J Hepatol.* 2015;7(30):2968. doi:10.4254/wjh.v7.i30.2968



30. Bulajic M. *Helicobacter pylori* and pancreatic diseases. *World J Gastrointest Pathophysiol.* 2014;5(4):380. doi:[10.4291/wjgp.v5.i4.380](https://doi.org/10.4291/wjgp.v5.i4.380)
31. Goni E, Franceschi F. *Helicobacter pylori* and extragastric diseases. *Helicobacter.* 2016;21:45-48. doi:[10.1111/hel.12340](https://doi.org/10.1111/hel.12340)
32. Kavita RP, Suresh RN, Babu VV. Probiotics, prebiotics and synbiotics – a review. *J Food Sci Technol.* 2015;52(12):7577-7587. doi:[10.1007/s13197-015-1921-1](https://doi.org/10.1007/s13197-015-1921-1)
33. Greiner AK, Papineni RVL, Umar S. Chemoprevention in gastrointestinal physiology and disease. Natural products and microbiome. *Am J Physiol Gastrointest Liver Physiol.* 2014;307(1):G1-G15. doi:[10.1152/ajpgi.00044.2014](https://doi.org/10.1152/ajpgi.00044.2014)
34. Wang C, Li J. Pathogenic microorganisms and pancreatic cancer. *Gastrointest Tumors.* 2015;2(1):41-47. doi:[10.1159/000380896](https://doi.org/10.1159/000380896)
35. Fan X, Alekseyenko AV, Wu J, et al. Human oral microbiome and prospective risk for pancreatic cancer: a population-based nested case-control study. *Gut.* 2018;67(1):120-127. doi:[10.1136/gutjnl-2016-312580](https://doi.org/10.1136/gutjnl-2016-312580)
36. Michaud DS, Izard J, Wilhelm-Benartzi CS, et al. Plasma antibodies to oral bacteria and risk of pancreatic cancer in a large European prospective cohort study. *Gut.* 2013;62(12):1764-1770. doi:[10.1136/gutjnl-2012-303006](https://doi.org/10.1136/gutjnl-2012-303006)
37. Choi SJ, Kim Y, Jeon J, et al. Association of microbial dysbiosis with gallbladder diseases identified by bile microbiome profiling. *J Korean Med Sci.* 2021;36(28):e189. doi:[10.3346/jkms.2021.36.e189](https://doi.org/10.3346/jkms.2021.36.e189)
38. de Martel C, Plummer M, Parsonnet J, van Doorn L-J, Franceschi S. *Helicobacter* species in cancers of the gallbladder and extrahepatic biliary tract. *Br J Cancer.* 2009;100(1):194-199. doi:[10.1038/sj.bjc.6604780](https://doi.org/10.1038/sj.bjc.6604780)
39. Di Tucci C, De Vito I, Muzii L. Immune-onco-microbiome: a new revolution for gynecological cancers. *Biomedicine.* 2023;11(3):782. doi:[10.3390/biomedicine11030782](https://doi.org/10.3390/biomedicine11030782)
40. Shanmughapriya S, SenthilKumar G, Vinodhini K, Das BC, Vasanthi N, Natarajaseenivasan K. Viral and bacterial aetiologies of epithelial ovarian cancer. *Eur J Clin Microbiol Infect Dis.* 2012;31(9):2311-2317. doi:[10.1007/s10096-012-1570-5](https://doi.org/10.1007/s10096-012-1570-5)
41. Pagano JS, Blaser M, Buendia MA, et al. Infectious agents and cancer: criteria for a causal relation. *Semin Cancer Biol.* 2004;14(6):453-471. doi:[10.1016/j.semcancer.2004.06.009](https://doi.org/10.1016/j.semcancer.2004.06.009)
42. Weitzman MD, Weitzman JB. What's the damage? The impact of pathogens on pathways that maintain host genome integrity. *Cell Host Microbe.* 2014;15(3):283-294. doi:[10.1016/j.chom.2014.02.010](https://doi.org/10.1016/j.chom.2014.02.010)
43. Xu W, Liu Z, Bao Q, Qian Z. Viruses, other pathogenic microorganisms and esophageal cancer. *Gastrointest Tumors.* 2015;2(1):2-13. doi:[10.1159/000380897](https://doi.org/10.1159/000380897)
44. Liu W, MacDonald M, You J. Merkel cell polyomavirus infection and Merkel cell carcinoma. *Curr Opin Virol.* 2016;20:20-27. doi:[10.1016/j.coviro.2016.07.011](https://doi.org/10.1016/j.coviro.2016.07.011)
45. Thaïss CA, Zmora N, Levy M, Elinav E. The microbiome and innate immunity. *Nature.* 2016;535(7610):65-74. doi:[10.1038/nature18847](https://doi.org/10.1038/nature18847)
46. Sender R, Fuchs S, Milo R. Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol.* 2016;14(8):e1002533. doi:[10.1371/journal.pbio.1002533](https://doi.org/10.1371/journal.pbio.1002533)
47. Gopalakrishnan V, Helmink BA, Spencer CN, Reuben A, Wargo JA. The influence of the gut microbiome on cancer, immunity, and cancer immunotherapy. *Cancer Cell.* 2018;33(4):570-580. doi:[10.1016/j.ccell.2018.03.015](https://doi.org/10.1016/j.ccell.2018.03.015)
48. DeWeerd S. How baby's first microbes could be crucial to future health. *Nature.* 2018;555(7695):S18-S19. doi:[10.1038/d41586-018-02480-6](https://doi.org/10.1038/d41586-018-02480-6)
49. Makino H, Kushi A, Ishikawa E, et al. Transmission of intestinal *Bifidobacterium longum* subsp. *longum* strains from mother to infant, determined by multilocus sequencing typing and amplified fragment length polymorphism. *Appl Environ Microbiol.* 2011;77(19):6788-6793. doi:[10.1128/AEM.05346-11](https://doi.org/10.1128/AEM.05346-11)
50. Martin V, Maldonado-Barragán A, Moles L, et al. Sharing of bacterial strains between breast milk and infant feces. *J Hum Lact.* 2012;28(1):36-44. doi:[10.1177/0890334411424729](https://doi.org/10.1177/0890334411424729)
51. Pannaraj PS, Li F, Cerini C, et al. Association between breast milk bacterial communities and establishment and development of the infant gut microbiome. *JAMA Pediatr.* 2017;171(7):647. doi:[10.1001/jamapediatrics.2017.0378](https://doi.org/10.1001/jamapediatrics.2017.0378)
52. Thaïss CA, Levy M, Itav S, Elinav E. Integration of innate immune signaling. *Trends Immunol.* 2016;37(2):84-101. doi:[10.1016/j.it.2015.12.003](https://doi.org/10.1016/j.it.2015.12.003)
53. Frosali S, Pagliari D, Gambassi G, Landolfi R, Pandolfi F, Cianci R. How the intricate interaction among toll-like receptors, microbiota, and intestinal immunity can influence gastrointestinal pathology. *J Immunol Res.* 2015;2015:489821. doi:[10.1155/2015/489821](https://doi.org/10.1155/2015/489821)
54. Huda MN, Lewis Z, Kalanetra KM, et al. Stool microbiota and vaccine responses of infants. *Pediatrics.* 2014;134(2):e362-e372. doi:[10.1542/peds.2013-3937](https://doi.org/10.1542/peds.2013-3937)
55. Moresco EMY, LaVine D, Beutler B. Toll-like receptors. *Curr Biol.* 2011;21(13):R488-R493. doi:[10.1016/j.cub.2011.05.039](https://doi.org/10.1016/j.cub.2011.05.039)
56. Tye H, Kennedy CL, Najdovska M, et al. STAT3-driven upregulation of TLR2 promotes gastric tumorigenesis independent of tumor inflammation. *Cancer Cell.* 2012;22(4):466-478. doi:[10.1016/j.ccr.2012.08.010](https://doi.org/10.1016/j.ccr.2012.08.010)
57. Fukata M, Chen A, Vamadevan AS, et al. Toll-like receptor-4 promotes the development of colitis-associated colorectal tumors. *Gastroenterology.* 2007;133(6):1869-1869.e14. doi:[10.1053/j.gastro.2007.09.008](https://doi.org/10.1053/j.gastro.2007.09.008)
58. Mittal D, Saccheri F, Vénéreau E, Pusterla T, Bianchi ME, Rescigno M. TLR4-mediated skin carcinogenesis is dependent on immune and radioresistant cells. *Embo J.* 2010;29(13):2242-2252. doi:[10.1038/emboj.2010.94](https://doi.org/10.1038/emboj.2010.94)
59. Ochi A, Nguyen AH, Bedrosian AS, et al. MyD88 inhibition amplifies dendritic cell capacity to promote pancreatic carcinogenesis via Th2 cells. *J Exp Med.* 2012;209(9):1671-1687. doi:[10.1084/jem.20111706](https://doi.org/10.1084/jem.20111706)
60. Littman DR, Rudensky AY. Th17 and regulatory T cells in mediating and restraining inflammation. *Cell.* 2010;140(6):845-858. doi:[10.1016/j.cell.2010.02.021](https://doi.org/10.1016/j.cell.2010.02.021)
61. Dapito DH, Mencin A, Gwak GY, et al. Promotion of hepatocellular carcinoma by the intestinal microbiota and TLR4. *Cancer Cell.* 2012;21(4):504-516. doi:[10.1016/j.ccr.2012.02.007](https://doi.org/10.1016/j.ccr.2012.02.007)
62. Neufert C, Becker C, Türeci Ö, et al. Tumor fibroblast-derived epiregulin promotes growth of colitis-associated neoplasms through ERK. *J Clin Invest.* 2013;123(4):1428-1443. doi:[10.1172/JCI63748](https://doi.org/10.1172/JCI63748)
63. Brandl K, Sun L, Neppel C, et al. MyD88 signaling in nonhematopoietic cells protects mice against induced colitis by regulating specific EGF receptor ligands. *Proc Natl Acad Sci U S A.* 2010;107(46):19967-19972. doi:[10.1073/pnas.1014669107](https://doi.org/10.1073/pnas.1014669107)
64. Grivennikov SI, Wang K, Mucida D, et al. Adenoma-linked barrier defects and microbial products drive IL-23/IL-17-mediated tumour growth. *Nature.* 2012;491(7423):254-258. doi:[10.1038/nature11465](https://doi.org/10.1038/nature11465)
65. Wu S, Rhee KJ, Albesiano E, et al. A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. *Nat Med.* 2009;15(9):1016-1022. doi:[10.1038/nm.2015](https://doi.org/10.1038/nm.2015)
66. Salcedo R, Worschech A, Cardone M, et al. MyD88-mediated signaling prevents development of adenocarcinomas of the colon: role of interleukin 18. *J Exp Med.* 2010;207(8):1625-1636. doi:[10.1084/jem.20100199](https://doi.org/10.1084/jem.20100199)
67. Salcedo R, Cattaïsson C, Hasan U, Yuspa SH, Trinchieri G. MyD88 and its divergent toll in carcinogenesis. *Trends Immunol.* 2013;34(8):379-389. doi:[10.1016/j.it.2013.03.008](https://doi.org/10.1016/j.it.2013.03.008)
68. Ting JP, Lovering RC, Alnemri ES, et al. The NLR gene family: a standard nomenclature. *Immunity.* 2008;28(3):285-287. doi:[10.1016/j.immuni.2008.02.005](https://doi.org/10.1016/j.immuni.2008.02.005)



69. Yeretsian G. Effector functions of NLRs in the intestine: innate sensing, cell death, and disease. *Immunol Res.* 2012;54(1-3):25-36. doi:[10.1007/s12026-012-8317-3](https://doi.org/10.1007/s12026-012-8317-3)
70. Hu B, Elinav E, Huber S, et al. Microbiota-induced activation of epithelial IL-6 signaling links inflammasome-driven inflammation with transmissible cancer. *Proc Natl Acad Sci.* 2013;110(24):9862-9867. doi:[10.1073/pnas.1307575110](https://doi.org/10.1073/pnas.1307575110)
71. Chen GY, Liu M, Wang F, Bertin J, Núñez G. A functional role for Nlrp6 in intestinal inflammation and tumorigenesis. *J Immunol.* 2011;186(12):7187-7194. doi:[10.4049/jimmunol.1100412](https://doi.org/10.4049/jimmunol.1100412)
72. Normand S, Delanoye-Crespin A, Bressenot A, et al. Nod-like receptor pyrin domain-containing protein 6 (NLRP6) controls epithelial self-renewal and colorectal carcinogenesis upon injury. *Proc Natl Acad Sci.* 2011;108(23):9601-9606. doi:[10.1073/pnas.1100981108](https://doi.org/10.1073/pnas.1100981108)
73. Elinav E, Strowig T, Kau AL, et al. NLRP6 inflammasome regulates colonic microbial ecology and risk for colitis. *Cell.* 2011;145(5):745-757. doi:[10.1016/j.cell.2011.04.022](https://doi.org/10.1016/j.cell.2011.04.022)
74. Khor B, Gardet A, Xavier RJ. Genetics and pathogenesis of inflammatory bowel disease. *Nature.* 2011;474(7351):307-317. doi:[10.1038/nature10209](https://doi.org/10.1038/nature10209)
75. Allen IC, Wilson JE, Schneider M, et al. NLRP12 suppresses colon inflammation and tumorigenesis through the negative regulation of noncanonical NF- $\kappa$ B signaling. *Immunity.* 2012;36(5):742-754. doi:[10.1016/j.immuni.2012.03.012](https://doi.org/10.1016/j.immuni.2012.03.012)
76. Couturier-Maillard A, Secher T, Rehman A, et al. NOD2-mediated dysbiosis predisposes mice to transmissible colitis and colorectal cancer. *J Clin Invest.* 2013;123(2):700-711. doi:[10.1172/JCI62236](https://doi.org/10.1172/JCI62236)
77. Rehman A, Sina C, Gavrilova O, et al. Nod2 is essential for temporal development of intestinal microbial communities. *Gut.* 2011;60(10):1354-1362. doi:[10.1136/gut.2010.216259](https://doi.org/10.1136/gut.2010.216259)
78. Kumari N, Dwarakanath BS, Das A, Bhatt AN. Role of interleukin-6 in cancer progression and therapeutic resistance. *Tumor Biol.* 2016;37(9):11553-11572. doi:[10.1007/s13277-016-5098-7](https://doi.org/10.1007/s13277-016-5098-7)
79. Chen L, Xiong Z, Sun L, Yang J, Jin Q. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res.* 2012;40:D641-D645. doi:[10.1093/nar/gkr989](https://doi.org/10.1093/nar/gkr989)
80. Han YW, Ikegami A, Rajanna C, et al. Identification and characterization of a novel adhesin unique to oral fusobacteria. *J Bacteriol.* 2005;187(15):5330-5340. doi:[10.1128/JB.187.15.5330-5340.2005](https://doi.org/10.1128/JB.187.15.5330-5340.2005)
81. Rubinstein MR, Wang X, Liu W, Hao Y, Cai G, Han YW. Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/ $\beta$ -catenin signaling via its FadA adhesin. *Cell Host Microbe.* 2013;14(2):195-206. doi:[10.1016/j.chom.2013.07.012](https://doi.org/10.1016/j.chom.2013.07.012)
82. Haider HJ, Turnbaugh PJ. Is it time for a metagenomic basis of therapeutics? *Science.* 2012;336(6086):1253-1255. doi:[10.1126/science.1224396](https://doi.org/10.1126/science.1224396)
83. Hahn RZ, Antunes MV, Verza SG, et al. Pharmacokinetic and pharmacogenetic markers of irinotecan toxicity. *Curr Med Chem.* 2018;26(12):2085-2107. doi:[10.2174/0929867325666180622141101](https://doi.org/10.2174/0929867325666180622141101)
84. Wallace BD, Wang H, Lane KT, et al. Alleviating cancer drug toxicity by inhibiting a bacterial enzyme. *Science.* 2010;330(6005):831-835. doi:[10.1126/science.1191175](https://doi.org/10.1126/science.1191175)
85. Bernstein C, Holubec H, Bhattacharyya AK, et al. Carcinogenicity of deoxycholate, a secondary bile acid. *Arch Toxicol.* 2011;85(8):863-871. doi:[10.1007/s00204-011-0648-7](https://doi.org/10.1007/s00204-011-0648-7)
86. Quante M, Bhagat G, Abrams JA, et al. Bile acid and inflammation activate gastric cardia stem cells in a mouse model of Barrett-like metaplasia. *Cancer Cell.* 2012;21(1):36-51. doi:[10.1016/j.ccr.2011.12.004](https://doi.org/10.1016/j.ccr.2011.12.004)
87. Yoshimoto S, Loo TM, Atarashi K, et al. Obesity-induced gut microbial metabolite promotes liver cancer through senescence secretome. *Nature.* 2013;499(7456):97-101. doi:[10.1038/nature12347](https://doi.org/10.1038/nature12347)
88. Seitz HK, Stickel F. Molecular mechanisms of alcohol-mediated carcinogenesis. *Nat Rev Cancer.* 2007;7(8):599-612. doi:[10.1038/nrc2191](https://doi.org/10.1038/nrc2191)
89. Seitz HK, Simanowski UA, Garzon FT, et al. Possible role of acetaldehyde in ethanol-related rectal cocarcinogenesis in the rat. *Gastroenterology.* 1990;98(2):406-413. doi:[10.1016/0016-5085\(90\)90832-L](https://doi.org/10.1016/0016-5085(90)90832-L)
90. Pharmaceu Sci GJ, Saks M, Upreti S, Dang R. Genotoxicity: mechanisms, testing guidelines and methods. *Glob J Pharm Pharm Sci.* 2017;1:133-138. doi:[10.19080/GJPPS.2017.02.555575](https://doi.org/10.19080/GJPPS.2017.02.555575)
91. Arthur JC, Perez-Chanona E, Mühlbauer M, et al. Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science (New York, N.Y.).* 2012;338(6103):120-123. doi:[10.1126/science.1224820](https://doi.org/10.1126/science.1224820)
92. Nesić D, Hsu Y, Stebbins CE. Assembly and function of a bacterial genotoxin. *Nature.* 2004;429(6990):429-433. doi:[10.1038/nature02532](https://doi.org/10.1038/nature02532)
93. Cuevas-Ramos G, Petit CR, Marcq I, Boury M, Oswald E, Nougayrède J-P. Escherichia coli induces DNA damage in vivo and triggers genomic instability in mammalian cells. *Proc Natl Acad Sci U S A.* 2010;107(25):11537-11542. doi:[10.1073/pnas.1001261107](https://doi.org/10.1073/pnas.1001261107)
94. Travaglione S, Fabbri A, Fiorentini C. The rho-activating CNF1 toxin from pathogenic *E. coli*: a risk factor for human cancer development? *Infect Agents Cancer.* 2008;3:4. doi:[10.1186/1750-9378-3-4](https://doi.org/10.1186/1750-9378-3-4)
95. Boleij A, Hechenbleikner EM, Goodwin AC, et al. The *Bacteroides fragilis* toxin gene is prevalent in the colon mucosa of colorectal cancer patients. *Clin Infect Dis.* 2015;60(2):208-215. doi:[10.1093/cid/ciu787](https://doi.org/10.1093/cid/ciu787)
96. Prindiville TP, Sheikh RA, Cohen SH, Tang YJ, Cantrell MC, Silva J. *Bacteroides fragilis* enterotoxin gene sequences in patients with inflammatory bowel disease. *Emerg Infect Dis.* 2000;6(2):171-174. doi:[10.3201/eid0602.000210](https://doi.org/10.3201/eid0602.000210)
97. Martin OCB, Bergonzini A, D'Amico F, et al. Infection with genotoxin-producing salmonella enterica synergises with loss of the tumour suppressor APC in promoting genomic instability via the PI3K pathway in colonic epithelial cells. *Cell Microbiol.* 2019;21(12):e13099. doi:[10.1111/cmi.13099](https://doi.org/10.1111/cmi.13099)
98. Smith JL, Bayles DO. The contribution of cytolethal distending toxin to bacterial pathogenesis. *Crit Rev Microbiol.* 2006;32(4):227-248. doi:[10.1080/10408410601023557](https://doi.org/10.1080/10408410601023557)
99. Elwell CA, Dreyfus LA. DNase I homologous residues in CdtB are critical for cytolethal distending toxin-mediated cell cycle arrest. *Mol Microbiol.* 2000;37(4):952-963. doi:[10.1046/j.1365-2958.2000.02070.x](https://doi.org/10.1046/j.1365-2958.2000.02070.x)
100. Shen Z, Feng Y, Rogers AB, et al. Cytolethal distending toxin promotes helicobacter cinaedi-associated typhlocolitis in interleukin-11-deficient mice. *Infect Immun.* 2009;77(6):2508-2516. doi:[10.1128/IAI.00166-09](https://doi.org/10.1128/IAI.00166-09)
101. Fox JG, Rogers AB, Whary MT, et al. Gastroenteritis in NF-kappaB-deficient mice is produced with wild-type *Campylobacter jejuni* but not with *C. jejuni* lacking cytolethal distending toxin despite persistent colonization with both strains. *Infect Immun.* 2004;72(2):1116-1125. doi:[10.1128/iai.72.2.1116-1125.2004](https://doi.org/10.1128/iai.72.2.1116-1125.2004)
102. Nougayrède JP, Homburg S, Taieb F, et al. Escherichia coli induces DNA double-strand breaks in eukaryotic cells. *Science (New York, N.Y.).* 2006;313(5788):848-851. doi:[10.1126/science.1127059](https://doi.org/10.1126/science.1127059)
103. Buc E, Dubois D, Sauvanet P, et al. High prevalence of mucosa-associated *E. coli* producing cyclomodulin and genotoxin in colon cancer. *PLoS One.* 2013;8(2):e56964. doi:[10.1371/journal.pone.0056964](https://doi.org/10.1371/journal.pone.0056964)
104. Huycke MM, Gaskins HR. Commensal bacteria, redox stress, and colorectal cancer: mechanisms and models. *Exp Biol Med (Maywood).* 2004;229(7):586-597. doi:[10.1177/153537020422900702](https://doi.org/10.1177/153537020422900702)
105. Carbonero F, Benefiel AC, Alizadeh-Ghamsari AH, Gaskins HR. Microbial pathways in colonic sulfur metabolism and links with health and disease. *Front Physiol.* 2012;3:448. doi:[10.3389/fphys.2012.00448](https://doi.org/10.3389/fphys.2012.00448)
106. Hullar MAJ, Burnett-Hartman AN, Lampe JW. Gut microbes, diet, and cancer. 2014;377-399. doi:[10.1007/978-3-642-38007-5\\_22](https://doi.org/10.1007/978-3-642-38007-5_22)



107. Attene-Ramos MS, Wagner ED, Plewa MJ, Gaskins HR. Evidence that hydrogen sulfide is a genotoxic agent. *Mol Cancer Res*. 2006; 4(1):9-14. doi:[10.1158/1541-7786.MCR-05-0126](https://doi.org/10.1158/1541-7786.MCR-05-0126)
108. Decroos K, Vanhemmens S, Cattoir S, Boon N, Verstraete W. Isolation and characterisation of an equol-producing mixed microbial culture from a human faecal sample and its activity under gastrointestinal conditions. *Arch Microbiol*. 2005;183(1):45-55. doi:[10.1007/s00203-004-0747-4](https://doi.org/10.1007/s00203-004-0747-4)
109. Christl SU, Eisner HD, Dusel G, Kasper H, Scheppach W. Antagonistic effects of sulfide and butyrate on proliferation of colonic mucosa: a potential role for these agents in the pathogenesis of ulcerative colitis. *Dig Dis Sci*. 1996;41(12):2477-2481. doi:[10.1007/BF02100146](https://doi.org/10.1007/BF02100146)
110. Deplancke B, Finster K, Graham WV, Collier CT, Thurmond JE, Gaskins HR. Gastrointestinal and microbial responses to sulfate-supplemented drinking water in mice. *Exp Biol Med (Maywood)*. 2003;228(4):424-433. doi:[10.1177/153537020322800413](https://doi.org/10.1177/153537020322800413)
111. Wang X, Yang Y, Moore DR, Nimmo SL, Lightfoot SA, Huycke MM. 4-hydroxy-2-nonenal mediates genotoxicity and bystander effects caused by *Enterococcus faecalis*-infected macrophages. *Gastroenterology*. 2012;142(3):543-551.e7. doi:[10.1053/j.gastro.2011.11.020](https://doi.org/10.1053/j.gastro.2011.11.020)
112. Balish E, Warner T. *Enterococcus faecalis* induces inflammatory bowel disease in interleukin-10 knockout mice. *Am J Pathol*. 2002; 160(6):2253-2257. doi:[10.1016/S0002-9440\(10\)61172-8](https://doi.org/10.1016/S0002-9440(10)61172-8)
113. Bos JL. ras oncogenes in human cancer: a review. *Cancer Res*. 1989; 49(17):4682-4689.
114. Norat T, Lukanova A, Ferrari P, Riboli E. Meat, fish, and colorectal cancer risk: the European prospective investigation into cancer and nutrition. *J Natl Cancer Inst*. 2005;97(12):906-916. doi:[10.1093/jnci/dji164](https://doi.org/10.1093/jnci/dji164)
115. Silvester KR, Cummings JH. Does digestibility of meat protein help explain large bowel cancer risk? *Nutr Cancer*. 1995;24(3):279-288. doi:[10.1080/01635589509514417](https://doi.org/10.1080/01635589509514417)
116. McCarthy EF. The toxins of William B. Coley and the treatment of bone and soft-tissue sarcomas. *Iowa Orthop J*. 2006;26:154-158.
117. Bickels J, Kollender Y, Merinsky O, Meller I. Coley's toxin: historical perspective. *Isr Med Assoc J*. 2002;4(6):471-472.
118. Jobin C. Precision medicine using microbiota. *Science*. 2018; 359(6371):32-34. doi:[10.1126/science.aar2946](https://doi.org/10.1126/science.aar2946)
119. Bhatia K, Bhumika, Das A. Combinatorial drug therapy in cancer - new insights. *Life Sci*. 2020;258:118134. doi:[10.1016/j.lfs.2020.118134](https://doi.org/10.1016/j.lfs.2020.118134)
120. Vétizou M, Pitt JM, Daillère R, et al. Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. *Science*. 2015; 350(6264):1079-1084. doi:[10.1126/science.aad1329](https://doi.org/10.1126/science.aad1329)
121. Sivan A, Corrales L, Hubert N, et al. Commensal Bifidobacterium promotes antitumor immunity and facilitates anti-PD-L1 efficacy. *Science*. 2015;350(6264):1084-1089. doi:[10.1126/science.aac4255](https://doi.org/10.1126/science.aac4255)
122. Chaput N, Lepage P, Coutzac C, et al. Baseline gut microbiota predicts clinical response and colitis in metastatic melanoma patients treated with ipilimumab. *Ann Oncol*. 2017;28(6):1368-1379. doi:[10.1093/annonc/mdx108](https://doi.org/10.1093/annonc/mdx108)
123. Gopalakrishnan V, Spencer CN, Nezi L, et al. Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. *Science*. 2018;359(6371):97-103. doi:[10.1126/science.aan4236](https://doi.org/10.1126/science.aan4236)
124. Routy B, Le Chatelier E, Derosa L, et al. Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science*. 2018;359(6371):91-97. doi:[10.1126/science.aan3706](https://doi.org/10.1126/science.aan3706)

**How to cite this article:** Garg S, Sharma N, Bharmjeet, Das A. Unraveling the intricate relationship: Influence of microbiome on the host immune system in carcinogenesis. *Cancer Reports*. 2023;e1892. doi:[10.1002/cnr2.1892](https://doi.org/10.1002/cnr2.1892)

# Unravelling the Ultralow Thermal Conductivity of Ternary Antimonide Zintl Phase $\text{RbGaSb}_2$ : A First-principles Study

Sangeeta, Rajesh Kumar, Ramesh Kumar, Kulwinder Kumar & Mukhtiyar Singh\*

Department of Applied Physics, Delhi Technological University, Delhi 110 042, India

Received 28 June 2023; accepted 14 August 2023

The recent discovery of antimonide based Zintl phase compounds has sparked the research in finding high-performance thermoelectric materials. In present study, a ternary antimonide Zintl phase  $\text{RbGaSb}_2$  is investigated using First-principles calculations. A good agreement observed between our computed results, such as lattice parameter and thermal conductivity, with the experimental report validating our theoretical framework. A direct band gap of 1.17 eV is obtained using Tran Blaha modified Becke Johnson approach. The negative value of Seebeck coefficient indicates its n-type character. We propose a strategy for enhancing power factor via carrier concentration optimization. The calculated results reveal the anisotropic transport properties. The intrinsic ultralow lattice thermal conductivity about  $0.094 \text{ Wm}^{-1}\text{K}^{-1}$  along the x-direction, and  $0.019 \text{ Wm}^{-1}\text{K}^{-1}$  along z-direction at room temperature is obtained. The ZT value can reach 0.90 (in x-direction) and 0.85 (in z-direction) for n-type doping at 900 K, indicating  $\text{RbGaSb}_2$  as promising thermoelectric material.

**Keywords:**  $\text{RbGaSb}_2$ ; Ternary Antimonide Zintl Phase; Thermal Conductivity; DFT

## 1 Introduction

Thermoelectric (TE) materials incorporate an approach that transform thermal energy into electricity without any moving parts and have been identified as highly promising candidates for energy harvesting<sup>1,2</sup>. The efficiency of these materials relies on the figure of merit, ZT, which is represented as  $S^2\sigma T / k_e + k_l$ . Here,  $S$ ,  $\sigma$ ,  $k_e$ , and  $k_l$  denote the Seebeck coefficient, electrical conductivity, electronic thermal conductivity, and lattice thermal conductivity, respectively. Currently, extensive research has been taking pace on various materials<sup>3-5</sup> to explore their potential application in thermoelectricity. Among these, Zintl compounds possess intriguing properties like mix chemical bonding, narrow band gap, and high density of material. Their complex structure leads to low lattice thermal conductivity due to large phonon scattering. Also, these compounds exhibit Phonon-Glass Electron-Crystal behaviour<sup>6,7</sup>. Motivated by the recent experimental synthesis of ternary antimonide  $\text{RbGaSb}_2$ <sup>8</sup>, we examine the structural, electronic and transport properties. This work presents an effective n-type Zintl compound with remarkable promise as a future TE material across a large temperature range.

## 2 Computational Methods

The properties of  $\text{RbGaSb}_2$ , were analysed using density functional theory based wien2k code<sup>9</sup>. The optimisation of the structure was performed using generalised gradient approximation method proposed by Perdew-Burke-Ernzerhof<sup>10</sup>. An energy convergence of 0.0001 Ry was achieved when the Kohn-Sham equations were solved in a self-consistent manner. For Brillouin zone sampling, a Monkhorst-Pack k-mesh of  $17 \times 17 \times 9$  was used. The Tran Blaha modified Becke Johnson (TB-mBJ) approach was chosen to perform calculations pertaining to the electronic and transport properties<sup>11</sup>. The transport properties were obtained via solving Boltzmann Transport equation (BTE) as implemented in BoltzTraP code<sup>12</sup>. The phono3py code was used for the computation of anharmonic third-order inter atomic force constants<sup>13</sup>. The lattice thermal conductivity was obtained by solving phonon BTE employing a dense  $13 \times 13 \times 7$  q-mesh.

## 3 Results and Discussion

### 3.1 Structural and electronic properties

$\text{RbGaSb}_2$  ternary antimonide Zintl phase crystallizes in tetragonal structure with space group  $P4_2/nmc$  (space group no. 137). The crystal structure as shown in Fig. 1(a) inset, of the investigated compound comprises of two-dimensional  $[\text{GaSb}_2]^-$

\*Corresponding authors: (E-mail: msphysik09@gmail.com)

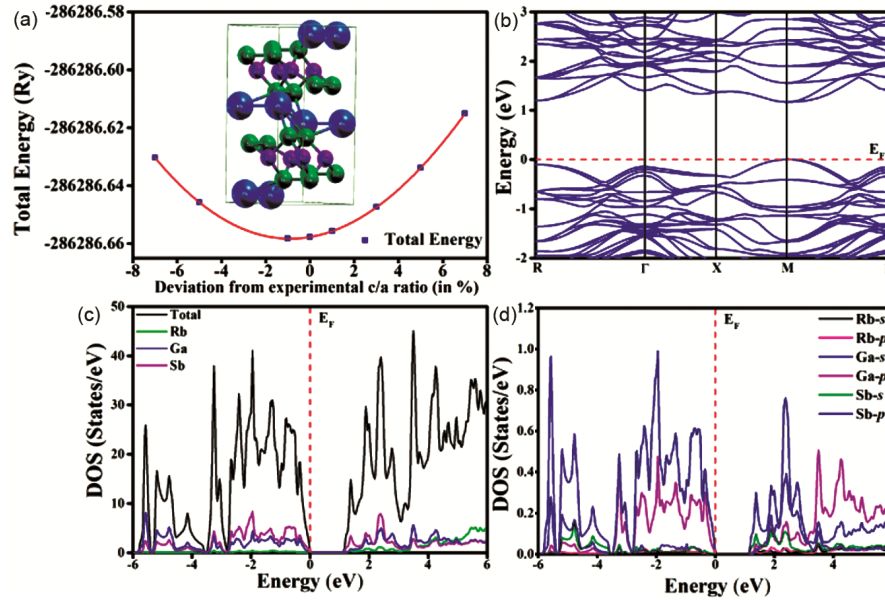


Fig. 1 — (a) The optimization of total energy versus deviation from experimental c/a ratio, Crystal structure (inset). Calculated (b) electronic band structure (c) and (d) density of states of RbGaSb<sub>2</sub> using TB-mBJ potential.

tetrahedral layers separated by layer of Rb<sup>+</sup> cations. There are two [GaSb<sub>2</sub>]<sup>-</sup> layers in the unit cell that are shifted along the [010] direction. This structure results in high degree of anisotropy. The calculated total energy versus deviation from experimental c/a ratio is shown in Fig. 1(a). The initial value of c/a ratio is taken from experimental data (1.857) and optimized value is 1.841. The optimized crystal structure parameters and comparison with the experimental ones are presented in Table 1. The optimised structure is further used to evaluate the electronic and transport properties.

The electronic band structure of RbGaSb<sub>2</sub> obtained using TB-mBJ is shown in Fig. 1 (b). The conduction band minima (CBM) and the valence band maxima (VBM), both present at M, indicating a direct band gap. The calculated energy band gap of 1.17 eV is close to previously estimated value of 1.0 eV. It is experimentally found that RbGaSb<sub>2</sub> tend to form n-type semiconductor near room temperature<sup>8</sup>, therefore we only focus on the lower part of CB that can determine the transport properties of n-type RbGaSb<sub>2</sub>. The band structure exhibits less dispersion at M k-point in the lower part of CB, suggesting that the electron effective mass is large and high *S* can be obtained. The density of states is obtained as shown in Fig. 1(c-d). Both VB and CB are predominantly contributed by Sb atoms, while negligible contribution of Rb atoms is observed. The sharp peaks observed at VBM and CBM. Also, the upper part of

Table 1— Calculated lattice parameters, volume, and total energy.

	a=b (Å)	c (Å)	Volume (Å <sup>3</sup> )	Total energy (Ry)
This work	8.359	15.390	1075.56	-286007.0047
Exp.[8]	8.335	15.483	1075.60	-

VB and lower part of CB are majorly contributed by the *p* orbital of Sb atoms.

### 3.2 Transport and properties

The transport parameters of RbGaSb<sub>2</sub> are computed through the solution of the BTE. We employed rigid band approximation which assumes that temperature and doping concentration have no impact on electronic band structure, and constant relaxation time approximation, which states that *S* is independent of scattering rate. We have investigated how temperature and electron concentration affect TE coefficients. The variation of the *S*,  $\sigma$ ,  $k_e$  and ZT is obtained as functions of electron concentration for both n-type RbGaSb<sub>2</sub> in the range 10<sup>18</sup>-10<sup>21</sup> cm<sup>-3</sup> along the [100] (x-direction) and [001] (z-direction) crystallographic directions for temperature 300, 600 and 900 K. The *S* value decreases with carrier concentration in both directions as shown in Fig. 2(a-b). The negative value of *S* signifies n-type behaviour of RbGaSb<sub>2</sub>.

The maximum *S* value obtained is -633.66  $\mu$ VK<sup>-1</sup> in x-direction at 900 K for electron concentration of 1 $\times$ 10<sup>18</sup> cm<sup>-3</sup>. Further utilizing the values of electrical conductivity and *S* we calculated power factor (PF) in terms of relaxation time ( $\tau$ ) at 300, 600, and 900 K in both directions. The PF/ $\tau$  first increase with the



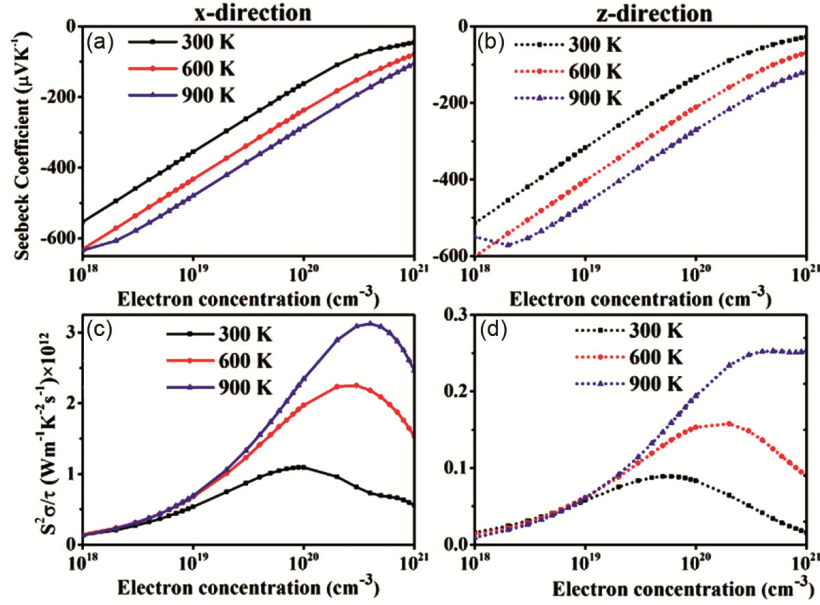


Fig. 2 — Calculated (a-b) Seebeck coefficient, and (c-d) power factor in x- (solid lines) and z- (dotted lines) direction as a function of electron concentration at different temperatures.

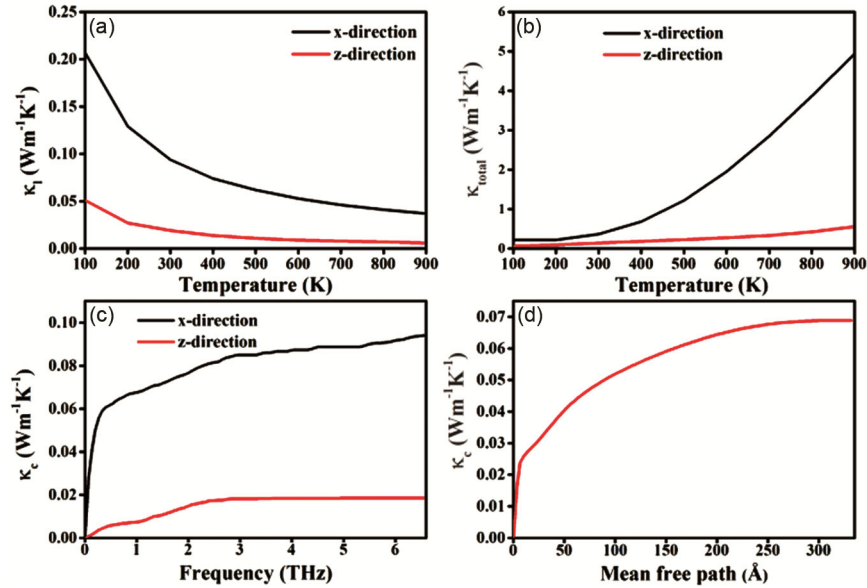


Fig. 3 — Calculated (a)  $k_l$  (b)  $k_{total}$  (c-d)  $k_c$  as a function of frequency and mean free path.

increase in electron concentration then reaches a maximum value, as shown in Fig. 2(c-d). High PF is obtained in x- than z-direction. The optimized PF/ $\tau$  values at  $4 \times 10^{20} \text{ cm}^{-3}$  are  $3.13 \times 10^{11}$  (x-direction) and  $0.25 \times 10^{11}$  (z-direction) at 900 K.

Further, to evaluate ZT we calculate  $k_l$  that decreases with the increase in temperature in both directions as shown in Fig. 3(a). This decrease in  $k_l$  at higher temperatures is attributed to the phenomenon of Umklapp phonon-phonon scattering. The total

thermal conductivity,  $k_{total}$  (Fig. 3(b)) of RbGaSb<sub>2</sub> was found to be 0.38 W/m-K at a temperature of 300 K and found within the range of previously reported values for antimony based TE<sup>14</sup>. The complex crystal structure, and rattling behavior of Rb cation and the presence of heavy elements leads to ultralow thermal conductivity of RbGaSb<sub>2</sub><sup>15</sup>.

The calculated frequency dependent cumulative lattice thermal conductivity ( $k_c$ ) shown in Fig. 3(b) reveals that low frequency phonon modes contribute

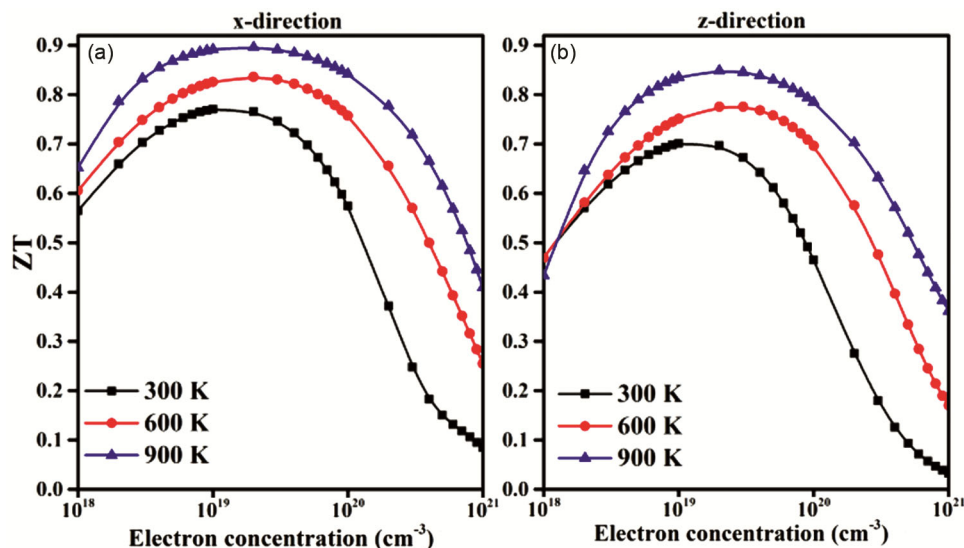


Fig. 4 — Calculated ZT as a function of electron concentration of RbGaSb<sub>2</sub> (a) in x-direction (b) in z-direction.

most to  $k_l$  in the x-direction and then increases slightly. Fig. 3(b) shows the  $k_c$  depends on the mean free path. The variation of ZT with electron concentration at different temperatures is shown in Fig. 4. At a constant temperature, ZT for both directions increase with electron concentration and it reaches optimum value at  $\sim 10^{19} \text{ cm}^{-3}$ , after that significantly decreases. The maximum ZT values at optimized electron concentration are found to be 0.90 (in x-direction) and 0.85 (in z-direction) at 900 K. The high ZT value of RbGaSb<sub>2</sub> is obtained by ultralow  $k_l$  and large value of power factor.

#### 4 Conclusion

In summary, we have investigated the electronic structure and TE properties of n-type RbGaSb<sub>2</sub> using first-principles calculations and BTE. The key advantage of RbGaSb<sub>2</sub> as TE material which is attributed to the complex crystal structure, the potential rattling of Rb cations, and presence of heavy elements. This low thermal conductivity allows for efficient conversion of heat into electricity, making it potential candidate for TE applications.

#### Acknowledgement

We acknowledge NSM for providing computing resources of 'PARAM SMRITI' at NABI, Mohali, which is implemented by C-DAC and supported by

the Ministry of Electronics and Information Technology (MeitY) and Department of Science and Technology (DST), Government of India.

#### References

- 1 Xiao Y & Zhao L-D, *Science*, 367 (2020) 1196.
- 2 He J & Tritt T M, *Science*, 357 (2017) eaak9997.
- 3 Kumar R, Kumar R, Singh M, Meena D & Vij A, *J Phys D: Appl Phys*, 55 (2022) 495302.
- 4 Yadav A, Kumar S, Muruganathan M & Kumar R, *EPL*, 132 (2020) 67003.
- 5 Sangeeta, Kumar R & Singh M, *J Mater Sci*, 57 (2022) 10691.
- 6 Kauzlarich S M, Brown S R & Jeffrey Snyder G, *Dalton Trans*, (2007) 2099.
- 7 Chen C, Feng Z, Yao H, Cao F, Lei B-H, Wang Y, Chen Y, Singh D J & Zhang Q, *Nat Commun*, 12 (2021) 5718.
- 8 Wang J, Owens-Baird B & Kovnir K, *Inorg Chem*, 61 (2022) 533.
- 9 Schwarz K, Blaha P & Madsen G K H, *Comput Phys Commun*, 147 (2002) 71.
- 10 Perdew J P, Burke K & Ernzerhof M, *Phys Rev Lett*, 77 (1996) 3865.
- 11 Tran F & Blaha P, *Phys Rev Lett*, 102 (2009) 226401.
- 12 Madsen G K H & Singh D J, *Comput Phys Commun*, 175 (2006) 67.
- 13 Togo A, Chaput L & Tanaka I, *Phys Rev B*, 91 (2015) 094306.
- 14 Kuo J J, Kang S D, Imasato K, K Tamaki K, Ohno S, Kanno T & Snyder G J, *Energy Environ Sci*, 11 (2018) 429.
- 15 Toberer E S, Zevkink A & Snyder G J, *J Mater Chem*, 21 (2011) 15843.





# Unsupervised sentiment analysis of Hindi reviews using MCDM and game model optimization techniques

NEHA PUNETHA<sup>ID</sup> and GOONJAN JAIN\*

Department of Applied Mathematics, Delhi Technological University, Delhi 110042, India  
e-mail: nehapunetha80@gmail.com; jaingoonjan@outlook.com

MS received 16 July 2022; revised 23 May 2023; accepted 2 August 2023

**Abstract.** In this study, we develop a novel multi-criteria decision-making (MCDM) and game-theoretic mathematical framework for analyzing review sentiment. The ratings and feedback of Hindi-speaking reviewers have been collected in an interactive database. This was accomplished by initially employing a game-theoretic approach to evaluating each review, based on the multi-objective optimization technique Complex Proportional Assessment (COPRAS), and then letting the two participants play the game until they reached a Nash equilibrium. Next, we extract the sentiment label from the inferred Hindi review dataset. To gauge the overall sentiment of a review, we classify it as either good, negative, or neutral. We analyze reviews by assigning a star rating and polarity score to comments written in the HindiSentiWordNet (HSWN) lexicon. To classify unstructured sentiment, we offer a model that optimizes for both polarity and rating scores. Our proposed model achieves comparable results to state-of-the-art models, as evidenced by experimental results on three widely used Hindi review datasets. We also use statistical analysis to determine the importance of the findings. The proposed MCDM Model ensures sound reasoning and consistency. In simulations, the proposed algorithm is seen to outperform the baseline and state-of-the-art approaches. This new benchmark for sentiment analysis was achieved by incorporating the rating and polarity score of the Hindi reviews into the MCDM and game model. Additionally, our technique is very generalizable and can do sentiment analysis across many different domains.

**Keywords.** COPRAS; sentiment analysis; MCDM; game theory; NLP.

## 1. Introduction

Sentiment analysis is employed in Natural Language Processing (NLP) to detect, extract, and quantify subjective data. By using computerized methods, sentiment analysis aims to decipher the sentiments or opinions expressed in a text. The focus of sentiment analysis is to explore the subjective information in a text and understand people's emotions, attitudes, and viewpoints toward a product [1]. The identification of sentiment orientation is crucial because the internet has become a vital resource for customers and companies to compare and evaluate products and services. In the realm of digital commerce, opinion-rich online reviews have become increasingly important for manufacturers and customers. Customers often seek information about popular products, notable features, and the reasons behind positive or negative evaluations. This has led to the development of various sentiment analysis algorithms and approaches. Automated sentiment analysis in NLP often relies on customer evaluations as a typical example. Manual sentiment analysis becomes impractical in certain cases due to the large volume of data and the need

for real-time processing [2]. Customer satisfaction represents how consumers perceive a product or service, reflecting the gap between their expectations and the experience. Online customer evaluations are of utmost importance as they can significantly impact the popularity of a seller's product or service [3]. To access detailed reviews and determine the appropriate sentiment tags for these reviews, it is necessary to identify the sentiment orientations of each review, which takes into account positive, negative, and neutral opinions.

Numerous solutions have been created, combining common sense with Natural Language Processing (NLP) Techniques, usually stated as supervised and unsupervised techniques [4]. Several elements must be considered when completing a Sentiment Analysis. For example, before traveling to a new place, we consulted with locals, but now we rely on online reviews to make decisions. These text data must be processed to establish viewpoint orientation, often known as opinion mining or sentiment classification. Almost two decades have extracted sentiment from English, with essential categories being sentiment categorization, lexicon resource development, etc. However, little work has been done in the sentiment analysis domain of Indian languages. The increasing availability of Indian

\*For correspondence

Published online: 09 September 2023

language data on the internet has raised the necessity of investigating sentiments of Indian language text.

Hindi is the world's 4th most widely spoken language [5]. The expanding amount of user-generated information on the internet fuels the sentiment analysis study. The Hindi language has gotten minimal attention in terms of sentiment analysis. Hindi's information content must be analyzed for industrial use. Most Indians speak Hindi as their first language and therefore prefer to express themselves and give feedback in their language. A recent survey found that over 500 million people speak Hindi in India. People seem eager to provide feedback and ideas in their local language [5]. As a result, more Hindi content is available online in weblogs, blogs, reviews, and recommendations. The need to analyze and extract relevant data has become critical. Unicode (utf-8) has increased the amount of non-English web content. Google search engines now support Hindi scripts. Mining such data and extracting valuable information has become necessary for businesses and people. Individual perspectives can help the government determine if a new policy will be successful and whether the public will be satisfied. Sentiment analysis is used in education, capital markets, product corporations, and many other industries where user reviews are crucial. We aim to create a benchmark framework for developing unsupervised sentiment frameworks using a multi-criteria decision-making optimization model.

### 1.1 Research problem

In recent years, the field of sentiment analysis and opinion mining has experienced significant growth, reflected in the substantial increase in published works and research conducted in this area. This surge in publications indicates the extensive focus and attention given to this field [6]. Due to the availability and accessibility of the Internet and Web 2.0, many individuals can now express their opinions online [7]. Numerous machine-readable data are already readily available online, which can be used to enhance SA and develop more effective algorithms [8]. This study enables businesses to understand the competitive environment in their particular business domain and maintain their supply-demand cycle. The necessary information can be discovered by analyzing the feelings that others have expressed through comments, criticism, and reviews. Depending on a specified number of points, these opinions or feelings are classified as positive, negative, or neutral (for example, 3 or 4, or 5 stars). The fundamental challenge in evaluating emotions is deciding how they should be portrayed in text and if a given sentence offers a positive or negative evaluation of its subject [5]. Therefore, to conduct a sentimental analysis, it is necessary to confirm the authenticity of the feelings being conveyed and their applicability to the topic

at hand. The systems use fundamental methods that necessitate a powerful computer and sufficient resources, yet this is not enough to deliver customer-focused solutions [9]. Even just extracting sentiment features from the text requires more data. The current study is focussing on these objectives.

- (1) Improving the algorithm's ability to identify sentiment.
- (2) Reducing the amount of necessary manual work involved in content analysis. In terms of both space and temporal complexity, it is the most effective.
- (3) A model that is unsupervised and unconstrained by training or language.
- (4) An unsupervised, language and training-independent model.
- (5) To create a system that facilitates clients' decision-making regarding purchases when you can foresee their thoughts and the most important aspects.

The unsupervised approach uses the MCDM and game model method for Hindi text, namely Complex Proportional Assessment (COPRAS) and game model [10], to calculate the total performance score of each review and then deduce the sentiment tag (positive, negative, neutral), using the game model to each review. The proposed methodology's applicability is tested with three different review datasets.

### 1.2 Contribution

Some of the article's most significant contributions and novelties are as follows:

- (1) This is the first study we are aware of to use MCDM and game theory in association with sentiment categorization to analyze Hindi reviews. This study evaluates the viability of coupled MCDM and game models for NLP applications like sentiment analysis of Hindi text using unique MCDM and game theory models.
- (2) This research aims to introduce a sentiment tagger that is used in the proposed technique to correctly categorize the sentiment of each review. Since the proposed model is unsupervised, it can be used with any low-resource language dataset, regardless of subject matter.
- (3) The proposed approach uses optimization strategies for Hindi text to achieve optimal precision with minimal resource usage. We classify sentiment on Hindi review datasets into three groups using both star ratings and textual input. For the proposed unsupervised approach to accurate sentiment tagging of reviews and state-of-the-art performance, we compiled three domain-specific Hindi review and rating datasets.
- (4) To evaluate the validity and robustness of the proposed structure, statistical significance is analyzed. The

efficiency of the suggested paradigm is examined using several performance evaluation metrics.

### 1.3 Organisation of the study

The structure of the paper is divided into the following categories: In section 1, we discuss an overview of sentiment analysis, the research problem, and our contribution. Section 2 provides a summary of prior sentiment analysis research. Section 3 contains the introductions to MCDM and game theory. The proposed algorithm for sentiment analysis is described in section 4. Section 5 contains the performance of the proposed model on three review datasets with that of cutting-edge techniques. We also discuss the statistical significance of validating the proposed method over the Hindi dataset. In sections 6 and 7, discussion and conclusions are provided.

## 2. Related work

In this section, we discuss the literature on sentiment analysis of Hindi text, MCDM, and game theory.

### 2.1 Sentiment analysis of Hindi text

Amitava Das and Bandyopadhyay proposed a technique for expanding SentiWordNet (Bengali) using a bilingual English-Bengali dictionary, Lexicons. Das *et al* [11] investigated the autonomous generation of sentiment lexicons using SentiWordNet. IIT Bombay developed HSWN. To create this resource, an English lexical resource called SentiWordNet was used in conjunction with the English-Hindi WordNet linkage. Annotations were added to each synset in SentiWordNet, such as positive, negative, and objective scores. The HSWN is used to determine the polarity of each term in the document, and the aggregate of opinions determines the absolute polarity. Balamurari *et al* introduce [12] multilingual data in the Hindi and Marathi languages and data in a single domain. Arora *et al* [5] built a subjective lexicon using WordNet and a small pre-annotated seed list.

Mittal *et al* [13] classified reviews in Hindi based on their sentiment after dealing with negation and discourse relationships. Jha *et al* [14] built a Hindi-language opinion mining method for a data set of Bollywood film reviews. Overall, they classified positive and negative materials with an accuracy of 87.1 percent. The shared task SAIL dataset, which includes tweets in Hindi, Bengali, and Tamil, has been the subject of several research papers [15]. These investigations were conducted on datasets from Twitter (SAIL 2015), IIT-Patna product and service reviews, and IIT-Patna movie reviews. Yakshi Sharma *et al* [16] presented a subjective lexicon-based system for sentiment

analysis of Hindi tweets linked to the topics “JAIHIND” and “World Cup 2015.” Modi *et al* [17] demonstrated how a rule-based system can be used to create a system for labeling parts of speech. Jha *et al* [18] suggested a reputation system capable of assessing trust among all legitimate eBay merchants and effectively ranking them.

Sarkar [19] performed the sentiment analysis of Hindi and Bengali tweets. Jha *et al* [20] investigated sentence-level subjectivity and attained an accuracy of nearly 80% on the Hindi dataset. On the IITP-Movie and IIT-Product review datasets, Akhtar *et al* [21] proposed a hybrid deep learning architecture with an average accuracy of 51.11%. Garg *et al* [22] performed sentiment analysis on tweets related to Prime Minister Mr. Narendra Modi’s radio broadcast “Mann Ki Baat” using a lexicon-based technique. Kunchukuttan *et al* [23] published the IndicNLP word embedding for the IITP-Movie and IIT-Product evaluation datasets in 2020. Augmentation GAN was presented by Pandey *et al* [24] as a tool for deep-learning models to generate coherent syntactic phrases. They validated their model on data in English, Hindi, and Bengali. Lin *et al* [25] generate stories using a multi-channel word embedding, a technique that combines classic vectorization techniques with their form of BERT. Akhtar *et al* [26] developed an ensemble framework for elucidating mood and emotion in text.

Several supervised and unsupervised methodologies have been proposed in the literature. However, due to the lack of standard trained Hindi datasets, supervised approaches are less reliable and neither domain nor language-independent. Their training is a daunting task. On the contrary, unsupervised algorithms need datasets only for evaluating the algorithm. However, the results generated by unsupervised algorithms are inferior to supervised algorithms.

There have been studies showing both supervised and unsupervised methods for sentiment analysis. Research in sentiment analysis employing MCDM and game theory on various Data sets has been undertaken to recommend the most suitable product or alternative. None of the research, however, has used the MCDM+ game model methodology in the Hindi reviews to create a sentiment orientation tagger for text. In other words, this is the first time the Hindi dataset has been used with COPRAS and a non-cooperative game model. An optimistic, pessimistic, and neutral emotion orientation tagger was created for this research. The purpose, methodology, and results of the proposed study are described in detail below.

### 2.2 Related work of the MCDM and game theory

Recently the author [27] used game theory and MCDM techniques for sentiment classification of the English text. Deng *et al* [28] used DEMATEL and Game Theory for supplier selection. A case study on the tea industry is done where the evolutionary game model and MCDM technique are used for the problem of decision-making in the Indian

Tea Industry [29]. The aspect-based and sentiment orientation-based sentiment classification is done on Indian restaurants using SAW and the cooperative game model [30]. The study [31] introduces Strategy selection in higher education using a game-theoretic model informed by multi-criteria decision-making. A study [32] on climate-adaptive multi-agent decision-making framework for assessing alternative plans for managing water and other environmental resources. The study introduces Strategies for Green Supply Chain Management Using Computational Models of the Demand Process [33]. GRA-based sentiment analysis of news headlines is introduced recently [34]. Bayesian game model-based sentiment categorization of reviews is done [35].

### 3. Preliminaries

In this section, we discuss various preliminaries of game theory, MCDM, and the relationship between game theory and MCDM.

#### 3.1 Game theory

Game theory is a mathematical framework for analyzing strategic interactions among rational decision-makers. It utilizes mathematical equations and models to study how players' choices and strategies impact the outcomes of a game. The primary mathematical tool used in game theory is the concept of a game form, represented by a tuple  $(N, A, U)$ ,

where:

$N$  represents the set of players involved in the game.

$A$  denotes the set of actions available to each player.

$U$  defines the utility function that assigns a numerical value to each player's payoff based on the combination of actions chosen by all players.

In a game, each player aims to maximize their utility or payoff. To do so, they must consider the strategies of other players and anticipate their actions. Strategies can be pure (where a player selects a specific action) or mixed (where a player randomizes their actions according to a probability distribution). One of the key solution concepts in game theory is the Nash equilibrium, named after mathematician John Nash. A Nash equilibrium represents a stable state where no player has an incentive to unilaterally change their strategy, given the strategies chosen by others. Mathematically, a Nash equilibrium is defined as a combination of strategies ( $s^*$ ) for all players, such that no player can unilaterally deviate and improve their payoff.

$$u_i(s_i^*, s_{-i}) \geq u_i(s_i, s_{-i}) \quad \forall i \in N \quad \text{and} \quad s_i \neq s_{-i}^* \quad (1)$$

Here in equation (1)  $u_{-i}$  represents the utility function for player  $i$ ,  $s_{-i}^*$  represents the player  $i$ 's strategy at the equilibrium,  $s_{-i}$  represents the strategies of all other players, and  $u_{-i}(s_i, s_{-i})$  represents the utility of player  $i$  when choosing strategy  $s_i$  while others choose strategies  $s_{-i}$ . Game theory also encompasses various solution concepts and mathematical techniques to analyze specific types of games, such as dominant strategies, extensive form games, cooperative games, and repeated games. These concepts and techniques involve additional mathematical equations and models tailored to specific game scenarios. In summary, game theory employs mathematical equations and models to study strategic interactions, player decision-making, and their impact on outcomes. It provides a rigorous framework for analyzing rational behavior in a wide range of situations, using concepts like game forms, utility functions, strategies, and solution concepts such as Nash equilibrium.

**3.1.1 Non-cooperative game:** Non-cooperative game theory simulates and assesses situations where each player's best decisions depend on his views or expectations about his opponent's behavior. Non-cooperative games often aim to forecast player actions and outcomes and achieve Nash equilibria. According to non-cooperative models, players cannot form legally binding agreements outside of the confines of the game's predetermined rules. The focus of non-cooperative solutions is on how to act strategically. The following describes a basic non-cooperative game theory.

#### 3.2 MCDM techniques

Multi-Criteria Decision Making (MCDM) is a methodology that facilitates the amalgamation of multiple qualitative and quantitative criteria to effectively assess and choose optimal solutions, achieving unanimous agreement among all relevant alternatives [36]. Within the realm of MCDM, two distinct types of criteria are discernible: benefit criteria, which necessitates maximization, and cost criteria, which demand minimization to attain the most desirable outcomes. A typical MCDM problem with  $m$  alternatives ( $A1, A2, \dots, Am$ ) and  $n$  criteria ( $C1, C2, \dots, Cn$ ) can be presented in the following form. Where  $M$  is the decision matrix and  $W$  is the weights assigned to each criterion.

$$M = [m_{ij}]_{m \times n}, \quad W = [w_j]_n$$

MCDM is a method that selects the most suitable solution from a set of available alternatives based on their performance against a set of evaluation criteria. The applications



of MCDM problems include a variety of fields, economy [37], education [38], management [39], production [40] sustainable development, construction, and so on. The fundamental components of an MCDM model are criteria, alternatives, and alternate performances against each criterion. These components have a one-to-one correspondence with the fundamental elements of a game, viz., players, strategies, and payoffs from likely outcomes, respectively. One of the studies demonstrates that MCDM and game theory components have one-to-one correspondence. This study aims to develop a novel decision support system that tags each review with a sentiment tag. In this study, we used the COPRAS MCDM technique.

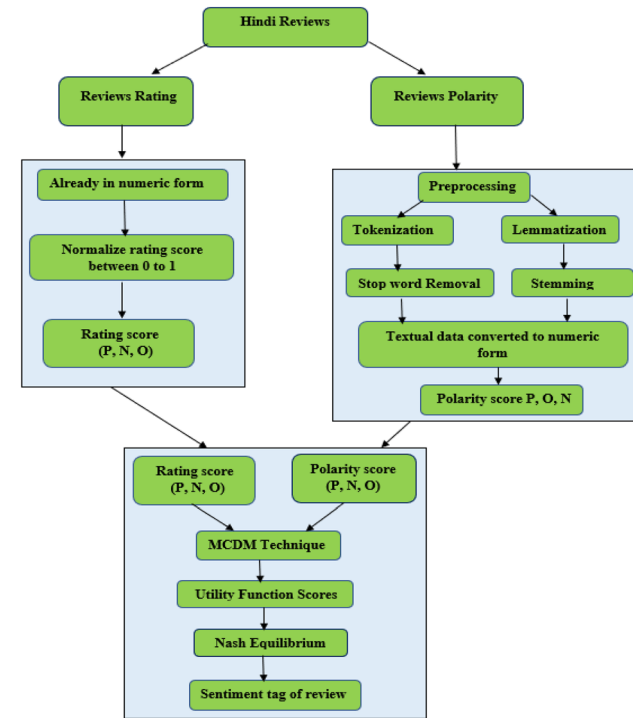
The author [41] presented the COPRAS technique. In addition, the COPRAS approach can calculate both maximization and minimization criteria. It is now possible to

derive both qualitative and quantitative metrics using this approach. The COPRAS approach has many advantages over other multi-criteria decision-making methods. The COPRAS method begins by constructing the problem's decision matrix. The decision matrix's columns and rows contain the criteria and the alternatives. We first construct the decision matrix  $x_{ij}$ . Then we calculate the normalized decision matrix ( $X_{ij}^*$ ) and weighted normalized decision matrix ( $d_{ij}$ ). In the fourth step, we calculate the Maximum ( $S_{i+}$ ) and Minimum Indexes ( $S_{i-}$ ) for Each Alternative. Then we calculate weighted averages ( $Q_i$ ). In the sixth step, we calculate the order of alternatives ( $A^{\square}$ ), at the end we calculate the polarity based on Utility value ( $U_i$ ). The complete MCDM technique is demonstrated by algorithm 2.

#### 4. Proposed methodology

This study uses the COPRAS model to present a framework for tagging reviews with sentiment based on ratings and customer comments. The model generally consists of three phases: (i) extracting data characteristics, such as customer evaluations and ratings; (ii) creating a decision matrix based on attributes; and (iii) carrying out an MCDM analysis. This is the first step in extracting features from competitors' products. The extracted characteristics are transformed into a decision matrix in the second phase. In the third phase, we assigned each criterion an equal weight of 0.5 and then ranked the reviews using the COPRAS method. The final tag for this review is the alternative that receives the highest rank. Each of these phases' steps is described below in figure 1.

**Step 1:** In the first step, we determine the polarity score of textual comments using (HSWN) [13]. The HSWN database stores a list of words that can be either positively or negatively charged. To determine the reviews' polarity, we pull the positive and negative polarity values from HSWN for each POS-tagged word. Figure 2 shows the excerpt of HSWN with the degree of positive, negative, and neutral degrees equal to  $(1 - (\text{degree of positive} + \text{degree of negative}))$ . We follow algorithm 1 to generate the sentiscores of the Hindi comments. These sentiscores give the performance value of the positive, negative, and neutral sentiment alternatives for polarity criteria.



**Figure 1.** A pipeline of the proposed model for sentiment analysis of reviews.

POS_TAG	ID	POS	NEG	LIST_OF_WORDS
a	10363	0.0	0.0	अनौपचारिक
a	2627	0.0	0.75	मृत
a	11476	0.125	0.0	परवर्ती
a	28106	0.25	0.375	अच्छा, बढ़िया
a	1156	0.875	0.0	सौभाग्यशाली, खुशकिस्मत, खुशानसीब, तक्रदीर, वाला, नसीब, वाला, भाग्यवान, भाग्यशाली, खुशकिस्मत, खुशानसीब
a	2279	0	1.0	दुर्भाग्यशाली, अभाग, बदनसीब, भाग्यहीन, मनहूस, बदकिस्मत, मंदभाग्य, बदकिस्मत, दईमारा, कमबख्त, कमबख्त, अधन्य, अभाग
a	2384	0.0	0.875	आवासहीन, आश्रयहीन, गृहहीन, गृहविहीन, बेघर, बेघरबार, अगतिक, ओह, अनिकेत
a	4714	0.25	0.125	सुराहित, सुराक्षित, सुराबुद्ध, सुराधपूर्ण, सुराभित, अधिवासित, सुराबुद्ध
a	1488	0.0	0.75	बदबुद्ध, दुर्गंधपूर्ण, दुर्गंधयुक्त, दुर्गंधित
a	29150	0.0	0.0	लगा, लगा, हुआ

**Figure 2.** Excerpt of HSWN lexicon.

**Algorithm 1: Evaluate polarity score of the review**

**Input:** Let  $W$  represent a collection of open-class and parts-of-speech-tagged words that were taken from the provided sentence. SWL is SentiWordNet list of words with each word's positive, negative and neutral sentiment value.

**Output:** Polarity Score of  $i^{\text{th}}$  review  $C = \{P, N, O\}$ , where  $P$  = positive sentiment value,  $N$  = negative sentiment value and  $O$  = Neutral sentiment value.

- 1: Initialize,  $P = 0, N = 0$  and  $O = 0$ ; such that  $C_i = \{0, 0, 0\}$ .
- 2: Take  $W = (w_1, w_2, \dots, w_i, \dots, w_n)$  where  $w_i$  represents the  $i^{\text{th}}$  ( $1 \leq i \leq K$ ) word in the input review.
- 3: Repeat step 4 for each word of  $W$
- 4: If ( $w_i \in \text{SWL}$ ), then
  - $P_{\text{Sentiscore}} = p$  + positive sentiment score of  $w_i$
  - $N_{\text{Sentiscore}} = n$  + negative sentiment score of  $w_i$ .
  - $O_{\text{Sentiscore}} = (1 - (P_{\text{Sentiscore}} + N_{\text{Sentiscore}}))$

**Step 2:** We aggregate all the criteria and alternative numeric values in this step. There are two criteria for polarity and rating; each criterion has three alternative degrees of positive, degree of negative, and degree of neutral. The alternative scores pertaining to polarity criteria are evaluated using the HSWN lexicon, which ranges between 0 and 1. The alternative scores for rating criteria are evaluated using equations (2), (3), and (4). Where  $p$  is the given rating of the product, we consider a positive rating ( $P$ ). Similarly, following equations (2), (3), and (4), we evaluate negative ( $N$ ) and neutral ratings ( $O$ ). We normalized these alternative scores using equations (5), (6), and (7) and renamed them  $dP$ ,  $dN$ , and  $dO$ . Now, these values lie between 0 and 1.

$$\text{Positive rating (P)} = p \quad (2)$$

$$\text{Negative rating (N)} = 5 - p \quad (3)$$

$$\text{Neutral rating (O)} = 5 - (P - N) \quad (4)$$

$$\text{Degree of Positive rating } dP = \frac{P}{P + N + O} \quad (5)$$

$$\text{Degree of Negative rating } dN = \frac{5 - p}{P + N + O} \quad (6)$$

$$\text{Degree of Neutral rating } dO = \frac{5 - (P - N)}{P + N + O} \quad (7)$$

**Step 3:** We perform the COPRAS method following

algorithm 2. Algorithm 2 generates the utility value by integrating rating and polarity scores which will work as the payoff for the non-cooperative game model.

**Algorithm 2: MCDM Method for sentiment tagging of reviews**

**Input:** Rating and Polarity scores of the review.

**Output:** Utility value of the reviews.

- 1: Construct a decision matrix  $x_{ij}$ .
- 2: Normalize the decision matrix ( $X_{ij}^*$ ), 
$$X_{ij}^* = \frac{x_{ij}}{\sum_{i=1}^m x_{ij}}$$
- 3: Calculate Normalized Decision-Making ( $d_{ij}$ ),  $d_{ij} = X_{ij}^* \cdot w_j$
- 4: Calculate Maximum ( $S_{i+}$ ) and Minimum Indexes ( $S_{i-}$ ) for Each Alternative.

$$S_{i+} = \sum_{j=1}^k d_{ij} \quad \text{and} \quad S_{i-} = \sum_{j=k+1}^n d_{ij}$$

$$\text{5: Calculate Weighted averages } (Q_i), \quad Q_i = S_{i+} + \frac{\min_i S_{-i} \sum_{i=1}^m S_{-i}}{S_{-i} \sum_{i=1}^m \frac{\min_i S_{-i}}{S_{-i}}} \cong S_{i+} + \frac{\sum_{i=1}^m S_{-i}}{\sum_{i=1}^m \left( \frac{1}{S_{-i}} \right)}$$

$$\text{6: Calculate the order of alternatives } (A^{\otimes}), \quad A^{\otimes} = \{A_i \mid \max_i Q_i\}$$

$$\text{7: Calculate the Utility value } (U_i), \quad U_i = \frac{Q_i}{Q_{\max}} \times 100\%$$



**Table 1.** Normal form representation of game played between two reviews.

	Positive	Negative	Neutral
Positive	$(\lambda_1, \omega_1)$	$(\lambda_1, \omega_2)$	$(\lambda_1, \omega_3)$
Negative	$(\lambda_2, \omega_1)$	$(\lambda_2, \omega_2)$	$(\lambda_2, \omega_3)$
Neutral	$(\lambda_3, \omega_1)$	$(\lambda_3, \omega_2)$	$(\lambda_3, \omega_3)$

**Step 4:** The non-cooperative game between two players (R1 and R2) is now played. Two players (R1 and R2) with three different strategies (Positive, Negative, and Neutral) are required to play the non-cooperative game. Utility values are calculated using algorithm 2 and will be taken as a payoff for R1 and R2. Utility value of R1 is  $\lambda_1, \lambda_2, \lambda_3$ , and Utility value of R2 is  $\omega_1, \omega_2, \omega_3$ . So possible combinations of Utility values of R1 and R2 are shown in table 1. We follow algorithm 3 to reach the Nash equilibrium. After achieving Nash equilibrium, each review's determined tag is comprised of the strategies that correlate to these payoffs of Nash equilibrium.

**Table 3.** Numeric Scores of Criteria and Alternative.

Alternative	Criteria	
	Polarity (C1)	Rating (C2)
R1 (4 star)		
Positive (A1)	0.075949	0
Negative (A2)	0.047059	1
Neutral (A3)	0.971109	0
R2 (1 star)		
Positive (A1)	0.89	0.571
Negative (A2)	0.21	0.143
Neutral (A3)	0.11	0.43

reviews. Using Algorithm 2, we obtain the utility value for R1 and R2 which is shown in table 4. The non-cooperative game concept is then put into practice using Algorithm 3.

---

**Algorithm 3: Deduce sentiment tag for review**


---

**Input:** Utility value  $\{\lambda_1, \lambda_2, \lambda_3\}$  for review  $R_i$  and  $\{\omega_1, \omega_2, \omega_3\}$  for review  $R_j$

**Output:** Sentiment Tag for  $R_i$  and  $R_j$ , i.e.,  $\{R_i, R_j\} \in \{P, N, O\}$ .

1: Generate a normal form matrix for players  $R_i$  and  $R_j$  using the appraisement scores.

2: Compute dominant strategies for  $R_i$  i.e. ( $DR_i$ ) and  $R_j$  i.e. ( $DR_j$ ), where  $DR_i, DR_j$  belongs to  $\{P, N, O\}$

3: Compute Nash equilibrium ( $NE$ ), where  $NE = DR_i$  intersection  $DR_j$ .

4: The strategies corresponding to  $NE$  are the sentiment tags for reviews  $R_i$  and  $R_j$ .

---

#### 4.1 Illustrative example

We take a Hindi review comment and rating from the online review dataset depicted below and use the COPRAS method to generate a sentiment tag, which is in detail below. COPRAS requires criteria that affect the alternatives in their computations. Tables 2 and 3 show that the criteria chosen in this study are polarity (C1) and rating (C2).

Both criteria are equally important. We considered equivalent weights, i.e. 0.5 for polarity and 0.5 is rating. We follow algorithm 3 to perform sentiment tagging of

**Table 2.** Criteria description.

Criteria	Weights	Type
Polarity (C1)	0.5	Beneficial
Rating (C2)	0.5	Beneficial

**R1: (4 star)** “मुझे यह पसंद है! मेरे फोन का हेडसेट जैक छोटा है, इसलिए यह वाफ व मे एक अलग एडेप्टर की आक कता के बिना फिट बैठता है। यह सीधे जैक में जुड़ा वाह! मैं आकार भूल जाता हूँ, लेकिन मुझे यकीन है कि यह इस ऊँ पाद की जानकारी के विनिर्देश अनुभाग में है। यह वही करता है जो मैं चाहता था।”

“I am liking this! My phone's headset jack is small, so this actually fit WITHOUT needing a separate adapter. It connected right into the jack. Yay! I forget the sizes, but I am sure it is in the specification section of this product's information. It does just what I wanted.”

“mujhe yah pasand hai! mere phon ka hedaset jaik chhota hai, isalie yah vaastav mein ek alag edeptar kee aavashyakata ke bina phit baiithata hai. yah seedhe jaik mein juda. vaah! main aakaar bhoool jaata hoon, lekin mujhe yakeen hai ki yah is utpaad kee jaanakaaree ke vinirdesh anubhaag mein hai. yah vahee karata hai jo main chaahata tha.”

**R2: (1 star)** “हमने कॉफ़ चीज़ बॉक्स, मैचो सूप और पनीर शशलकि सिज़लर ऑर किए। सिज़लर बासी था। पनीर की महक आ रही थी और वेंटर इतनी बदतमीजी कर रहा था कि गलती तक वीकार नहीं कर सका। फिर कभी नहीं जा रहा।”

---

“We ordered Corn Cheese Balls, Mancho Soup and Paneer Shashlik Sizzler. The sizzler was stale. The smell of cheese was coming and the waiter was so abusive that he could not even admit the mistake. never going again”

“hamane korn cheez bols, maincho soop aur paneer shashalik sizalar ordar kie. sizalar baasee tha. paneer kee mahak aa rahee thee aur vetar itanee badatameejee kar raha tha ki galatee tak sveekaar nahin kar saka. phir kabhee nahin ja raha”

The game model is presented in table 5 in normal form. Using the idea of Nash equilibrium, we next ascertained the sentiment orientation of both reviews, which is presented in table 6.

**Table 4.** Utility value score from MCDM.

Alternative	Performance score
<i>R1 (4star)</i>	
Positive (A1)	0.00161
Negative (A2)	0.00664
Neutral (A3)	0.30510
<i>R2 (1 Star)</i>	
Positive (A1)	0.03797
Negative (A2)	0.476471
Neutral (A3)	0.48555

**Table 5.** Non-cooperative game model for deducing sentiment tag.

R <sub>1</sub>	R <sub>2</sub>		
	Positive (A1)	Negative (A2)	Neutral (A3)
Positive (A1)	(0.00161, 0.3797)	<b>(0.00161, 0.476471)</b>	(0.00161, 0.48555)
Negative (A2)	(0.00664, 0.3797)	(0.00664, 0.476471)	(0.00664, 0.48555)
Neutral (A3)	(0.30510, 0.3797)	(0.30510, 0.476471)	(0.30510, 0.48555)

**Table 6.** Deduced tag using the game model.

R <sub>1</sub>	R <sub>2</sub>	
	Positive (A1)	Negative (A2)
Positive (A1)	(0.00161, 0.3797)	<b>(0.00161, 0.476471)</b>

**Table 7.** Data statistics of different datasets.

Data set	Language	Positive	Negative	Neutral
Movies reviews	Hindi	512	350	138
Hotel reviews	Hindi	657	132	211
Electronics reviews	Hindi	576	292	132

The Nash equilibrium of the game model played between two players is (0.00161, 0.476471), and the strategies corresponding to the payoff are positive and negative. Deduced tag of the R1 is positive, and R2 is negative. In the following way, we deduced the sentiment tag of each review. We deduced the sentiment tag using the COPRAS technique of MCDM and the game model. We combined the polarity and rating scores of reviews, evaluated the performance score, and deduced each review's sentiment orientation.

## 5. Result and discussion

In this section, we discuss various datasets, different evaluation metrics, and a comparison of various techniques with the proposed model over different datasets, and at the last, we present the statistical significance of the proposed model.

### 5.1 Data collection

We applied the suggested technique to three sets of data that included ratings and comments written in Hindi. The first dataset was the movie reviews dataset crawled from online sources<sup>1</sup>. Second, we crawled hotel reviews and ratings from online<sup>2</sup> sources. The third dataset is the subset of the Amazon electronics<sup>3</sup> dataset. Each dataset contains 1000 reviews and ratings. Table 7 shows the data statistics of the three collected datasets.

To assess the proposed model's effectiveness, we used various evaluation metrics. Figure 3 depicts the performance of various evaluation metrics across the Hindi review dataset.

### 5.2 Evaluation on the Movie Dataset

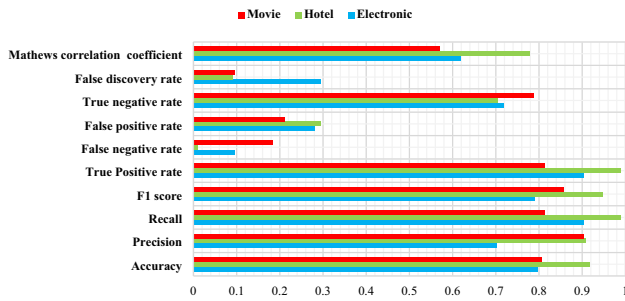
We calculated accuracy, precision, recall, and the F-measure on movie review datasets to compare performance (table 8). Essentially, all of these measures assess the same qualities and, as a result, generate remarkably similar values for a dataset.

Singh *et al* [42] proposed SentiWordNet techniques for sentiment classification of Hindi movie reviews with an accuracy of 63.42%, shown in table 8. Bhoir *et al* [43] proposed two models, Naïve Bayes, having an accuracy of 71%. Joshi *et al* [44] developed the Hindi-SentiWordNet (HSWN) lexical resource for sentiment analysis of a Hindi movie dataset, with an accuracy of 60%. Seshadri *et al* [45] proposed RNN model whose accuracy is 72%. Akhtar *et al*

<sup>1</sup>“Internet Movie Database, <http://www.imdb.com>”.

<sup>2</sup><https://www.tripadvisor.in/Restaurants>.

<sup>3</sup><https://jmcauley.ucsd.edu/data/amazon/>.



**Figure 3.** Performance of Evaluation Metric over three dataset reviews.

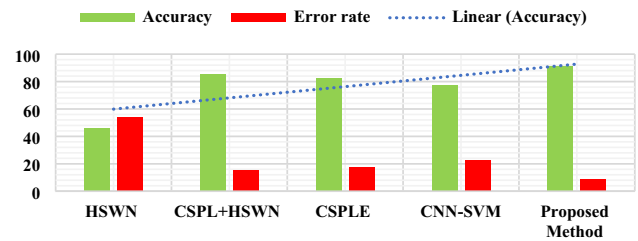
**Table 8.** Comparison of different approaches with the proposed approach.

Unsupervised method	Accuracy	F-measure	Precision	Recall
SWN(VS+APS) [42]	0.63	0.64	0.63	0.63
Naïve bayes [43]	0.71	0.75	0.75	75
HSWN [44]	0.60	0.46	0.60	37.5
RNN [45]	0.72	0.70	0.72	0.71
CNN-SVM [46]	0.65	0.64	0.6	0.66
CSPL+HSWN [47]	0.76	0.73	0.75	0.74
LSTM+CNN [48]	0.78	0.76	0.77	0.76
Proposed method	0.80	0.85	0.87	0.81

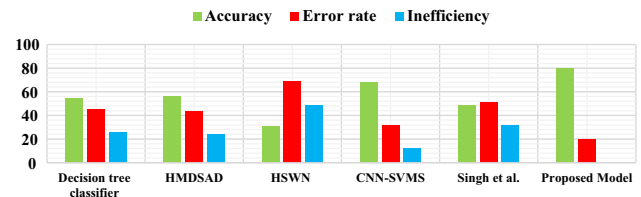
[46] proposed a CNN-SVM model whose accuracy is 65.96%. Mishra *et al* [47] created the (CSPLE + HSWN) model with 76.5% accuracy. Jain *et al* [48] have introduced the Hindi Text classification using an optimization technique. The proposed work outperforms all the techniques for the Hindi dataset.

### 5.3 Evaluation on the hotel dataset

Mishra *et al* [47] introduced the HSWN, CSPL+HSWN, and CSPLE, with respective accuracy rates of 46%, 85%, and 82.5%, and error rates of 54%, 15%, and 17.5%; among these models, CSPL+HSWN had the greatest accuracy rate of 85%. Akhtar *et al* [21] embedded vectors from the CNN. The sentiment-augmented optimized vector obtained at the end is used for SVM training for the proposed model's sentiment classification accuracy of 77.16%. Below, figure 4. depicts the accuracy and error rate of all the approaches where green bars denote these models' accuracy in predicting the sentiment tagging and red bars denote the error rate compared to the proposed model. Figure 4 illustrates that the proposed model's accuracy is 91%, greater than the other approaches, and recorded the lowest error rate of 9%, indicating that the results predicted by the proposed MCDM method are more accurate.



**Figure 4.** Comparison of existing approaches with a proposed model.

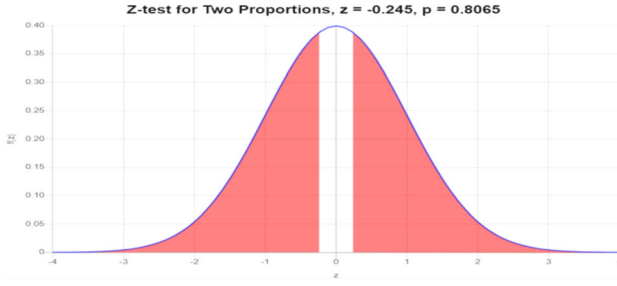


**Figure 5.** Comparison of existing approaches with the proposed model in terms of accuracy, error rate, and inefficiency.

### 5.4 Evaluation on the electronic dataset

Figure 5 depicts the performance comparisons of the proposed approach with other models on the same dataset that outperforms the other model in all respects. Jha *et al* [49] proposed the HMDSAD dictionary-based approach to classify unlabelled reviews from the target domain into positive, negative, and neutral categories. This approach had an accuracy of 56%, and the error rate recorded was 44% over the electronics reviews<sup>4</sup> dataset. HSWN [11] is a comprehensive lexicon covering the polarity of words in the Hindi language. It contains the following fields: POS tag, Synset ID (WordNet ID in Hindi), Positive score, Negative score, and Related Terms (separated by a comma). The accuracy recorded for this lexicon-based method is 31% and 69%, which is highest than the existing approaches. Akhtar *et al* [21] developed the embedded vectors from a convolutional neural network (CNN). The sentiment-augmented optimized vector is used to train the SVM for sentiment classification, and the supervised CNN-SVMS model has an accuracy of 68.04%, and an error rate is around 32%. Singh *et al* [50] pre-processed the Hindi texts and identified their English equivalents, and a summary review was then calculated using the HSWN database. As is evident from the table, the technique offered by Singh *et al* is 31.48% less effective than the proposed model with 48% accuracy. The classification accuracy of the decision trees classifier is 54%, and this supervised method is 25% less efficient than the proposed model, and the recorded rate is 45% [51].

<sup>4</sup><https://jmcauley.ucsd.edu/data/amazon/>.



**Figure 6.** A visual illustration of the above hypothesis's crucial area.

**Table 9.** Efficacy of the proposed model in various situations.

Efficiency	Best case	Average case	Worst case
Time complexity	$\Omega(m+n)$	$\theta(mn)$	$O(mn)$
Space complexity	$\Omega(m+n)$	$\theta(mn)$	$O(mn)$

### 5.5 Statistical validation

We performed the statistical Z test on two proportions to verify the effectiveness of the suggested model. The dataset for the review was divided into two samples of various sample sizes. We collected 1000 reviews ( $n_1$ ) for sample 1. 900 reviews out of 1000 were accurately tagged. ( $p_1$ ) = 0.9009 for the sample proportion. 500 reviews were included in sample 2, and 453 of those reviews were correctly classified, yielding a sample proportion of 0.906. For two population proportions ( $p_1$  and  $p_2$ ), the Z-test was performed. We looked into the following null and alternate hypotheses for the population proportion.

$$H_0: p_1 = p_2$$

$$H_a: p_1 \neq p_2$$

where  $H_0$  is the null hypothesis, and  $H_a$  is the alternative hypothesis. The value of the pooled proportion is computed using equation (8).

$$P = \frac{X_1 + X_2}{N_1 + N_2} = \frac{900 + 452}{1000 + 500} = 0.9013 \quad (8)$$

We utilized a z-test for two population proportions and a two-tailed test. The z-statistic's calculation is shown in equation (9).

$$z = \frac{p_1 - p_2}{\sqrt{P(1-P)(1/n_1 + 1/n_2)}} = \frac{0.9 - 0.904}{\sqrt{0.9013(1 - 0.9013)(1/1000 + 1/500)}} = -0.245 \quad (9)$$

The accepted and critical regions of the aforementioned hypothesis are depicted graphically in figure 6. We failed to reject the null hypothesis  $H_0$ . As a result, there is insufficient evidence to assert that the population proportion  $p_1$  differs

from  $p_2$  at the  $\alpha$  significance level. Up to a 77% level of significance, the null hypothesis is accepted, but at 78% null hypothesis is rejected. This suggests that our model's performance holds across a wide range of datasets and sample sizes.

## 6. Discussion

In comparison to other approaches, each algorithm has advantages and limitations. Similarly, the proposed algorithm has advantages and disadvantages. The following are the most prominent benefits and drawbacks.

(1) We have only evaluated a limited handful of centrality measurements in evaluating the performance. We will have more opportunities to use various centrality measures in the future as new researchers emerge daily.

(2) For the evaluation of an algorithm's efficacy and efficiency, its time and space complexity is critical. The number of operations an algorithm must carry out concerning the size of the input dataset and its temporal complexity. The algorithm's space complexity gauges how little space it needs to operate with various input sizes. When  $m$  is the number of possibilities and  $n$  is the number of criteria, table 9 displays the algorithm's runtime and space complexity in various scenarios, where  $m$  and  $n$

**Table 10.** Examples where the proposed model fails.

Reviews	Actual	Predicted
“मुझे ट्रिलर और हॉरर फिल्म नापसन्द नहीं है।” (“Mujhe triller aur horror film napasand hai”) (“I do not dislike triller and horror movie”) (“अंतिम एपिसोड अंत में एक भयानक मोड़ के साथ अश्चरजनक था।”) (“Antim episode ant me ek bayanak mor ke sath ashchrajanak tha”) (“The final episodes was surprising with a terrible twist at the end “) (negative term used in positive way)	Positive	Negative
“फिल्म देखना आसान था लेकिन मैं इसे अपने दोस्तों को सिफारिश नहीं करूंगा।” (“Film dekhna aasan tha lekin me isse apne dosto ki sifarish nhi karunga”) (“The film was easy to watch but I would not recommend it my friends”) (“कालिल और जहरीली दिख रही है एक्ट्रेस”) (“katil aur jahreli dikh rahi hai actress”) (“Actress is looking killer and poison”) (Irony)	Positive	Negative

**Table 11.** Sentiment classification of English Text.

Sl. no.	Dataset	Recall	F-1 score	Accuracy	Precision
1	English Movie reviews <sup>a</sup>	0.814	0.802	0.791	0.785
2	English Restaurant reviews <sup>b</sup>	0.785	0.772	0.752	0.768
3	English Electronic reviews <sup>c</sup>	0.767	0.754	0.743	0.7449

<sup>a</sup> <http://www.imdb.com>

<sup>b</sup> <https://www.kaggle.com/code/residentmario/exploring-tripadvisor-uk-restaurant-reviews/notebook>

<sup>c</sup> <https://www.kaggle.com/datasets/datafiniti/amazon-and-best-buy-electronics>

represent the number of possibilities and criteria, respectively. Table 9 shows level of difficulty belongs to the category of P-class problems that can be solved in polynomial time. In addition, it is a deterministic algorithm, which means that it always calculates the correct response. The proposed approach applies to any other language dataset with ease, in addition to being useful in analyzing metrics and complexity.

Table 10 shows that the use of game theory has resulted in a 50% reduction in the linear and polynomial time complexity of time and space, respectively. This is because, in the current study, we were able to use game theory to play the game between two players (R1 and R2) simultaneously, cutting the time needed to tag two reviews in half.

(3) HSWN lexicon is the foundation for the sentiment scoring of opinion words of Hindi text. The primary shortcoming of HSWN is that insufficient words are covered, and some words do not receive the appropriate HSWN score. A significant disadvantage of the lexicon-based method is that the system cannot correctly classify consumer feedback if a word or polarity shifter is missing from the sentiment lexicon. Certain sentiment words and polarity shifters cannot be accurately categorized by the proposed system. Table 10 presents a few examples.

(4) *Results on English dataset:* We created three general domain small datasets of English languages to test the robustness of the suggested model. Table 9 provides the results of implementing the proposed approach on the English review dataset. The results are promising as shown in table 11. Hence this implies that the proposed approach is language-independent and can be adaptive to any language whose lexicon database is available.

## 7. Conclusion

This research examined the mathematical underpinnings, as well as the use of MCDM and game theory, for determining the tone of reviews written in Hindi. The

presented model approach is an innovative unsupervised method for sentiment tagging reviews. We evaluated our proposed model on three distinct Hindi review datasets and compared its performance to that of both supervised and unsupervised methods. We also used Z-tests to verify the statistical validity of the suggested model. The suggested model achieved superior results compared to both state-of-the-art supervised and unsupervised methods. We suggest enhancing the model so that it can be used to categorize sentiments and emotions into multiple classes. Since it is unsupervised, the suggested model can readily be modified to classify emotion in additional low-resource languages like Hindi, Bengali, Urdu, etc. The presented model paved the way for several optimization strategies used in a wide range of Natural Language Processing (NLP) applications, including WSD, query expansion, sarcasm detection summarization, and so on. Non-cooperative Nash games and other optimization methods will be the primary focus of our future investigation as we work to build unsupervised approaches.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

## Abbreviations

COPRAS	Complex Proportional Assessment
HSWN	HindiSentiWordNet
MCDM	Multi-criteria decision-making
NLP	Natural Language Processing
CNN	Convolutional neural network

## References

- [1] Hussein D M E D M 2018 A survey on sentiment analysis challenges. *Journal of King Saud University Engineering Sciences* 30: 330–338
- [2] Ilieska K 2013 *Importance of Customer Satisfaction*. *TEM Journal* 2(4): 327–331
- [3] Kang D and Park Y 2014 Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. *Expert Systems with Applications* 41: 1041–1050
- [4] Schouten K, van der Weijde O, Frasinca F and Dekker R 2017 Supervised and Unsupervised Aspect Category Detection for Sentiment Analysis With Co-Occurrence Data. *IEEE Transactions on Cybernetics* 48: 1263–1275
- [5] Kulkarni D S, Sunil D R and Rodd S 2021 Sentiment Analysis in Hindi-A Survey on the State-of-the-art Techniques. *ACM Trans. Asian Low-Resour Lang. Inf. Process.* 21: 345–358
- [6] Han H, Zhang J and Yang J 2018 Generate domain-specific sentiment lexicon for review sentiment analysis. *Multimedia Tools and Applications* 77: 21265–21280



- [7] Yang L, Li Y, Wang J and Sherratt R S 2020 Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning. *IEEE Access* 8: 23522–23530
- [8] Mladenović M, Mitrović J, Krstev C and Vitas D 2016 Hybrid sentiment analysis framework for a morphologically rich language. *Journal of Intelligent Information Systems* 46: 599–620
- [9] Huang F, Yuan C and Bi Y 2022 Multi-granular document-level sentiment topic analysis for online reviews. *Applied Intelligence* 52: 7723–7733
- [10] Akhavan P, Barak S, Maghsoudlou H and Antuchevičienė J 2015 FQSPM-SWOT for strategic alliance planning and partner selection; case study in a holding car manufacturer company. *Technological and Economic Development of Economy* 21: 165–185
- [11] Das A 2010 SentiWordNet for Indian languages. *Aclanthology*: 21–22
- [12] A.R. B, Joshi A and Bhattacharyya P 2012 Cross-Lingual Sentiment Analysis for Indian Languages using Linked WordNets. *Proceedings of COLING 2012: Posters* 1:73–82
- [13] Mittal N, Agarwal B and Chouhan G 2013 Sentiment Analysis of Hindi Review based on Negation and Discourse Relation. *International Joint Conference on Natural Language Processing*:14–18
- [14] Jha V, Manjunath N and Shenoy P D, 2015 HOMS: Hindi opinion mining system. *International Conference on Recent Trends in Information Systems, ReTIS* :366–371
- [15] Patra B, Das D and Das A 2015 Shared task on sentiment analysis in indian languages (sail) tweets-an overview. *Springer* 9468: 650–655
- [16] Sharma Y, Mangat V and Kaur M 2016 A practical approach to Sentiment Analysis of Hindi tweets. *International Conference on Next Generation Computing Technologies, NGCT* :677–680
- [17] Modi D and Nain N 2016 Part-of-Speech Tagging of Hindi Corpus Using Rule-Based Method. In: *Proceedings of the International Conference on Recent Cognizance in Wireless Communication & Image Processing*: 241–247
- [18] Jha V, Shenoy S R P and Venugopal K R 2016 Reputation system: Evaluating reputation among all good sellers. *Human Language Technologies* 93: 115–121
- [19] Sarkar K 2020 Heterogeneous classifier ensemble for sentiment analysis of Bengali and Hindi tweets. *Sadhana - Academy Proceedings in Engineering Sciences* 45: 1–17
- [20] Jha V and R S G and Deepa Shenoy P R V K, 2016 Generating Multilingual Subjectivity Resources using English Language. *International Journal of Computer Applications* 152: 975–8887
- [21] Akhtar M S, Kumar A, Ekbal A and Bhattacharyya P 2016 A hybrid deep learning architecture for sentiment analysis. *Proceedings of COLING 2016: Technical Papers*: 482–493
- [22] Garg K 2019 Sentiment analysis of Indian PM's "Mann Ki Baat." *International Journal of Information Technology* 12: 37–48
- [23] Kakwani D, Kunchukuttan A and Golla S 2020 IndicNLP-Suite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP* 445: 4948–4961
- [24] Pandey S, Akhtar M S and Chakraborty T 2021 Syntactically Coherent Text Augmentation for Sequence Classification. *IEEE Transactions on Computational Social Systems* 8: 1323–1332
- [25] Lin J W and Chang R G 2022 Chinese story generation of sentence format control based on multi-channel word embedding and novel data format. *Soft Computing* 26: 2179–2196
- [26] Akhtar M S, Ghosal D, Ekbal A, Bhattacharya P and Kurohashi S 2022 All-in-One: Emotion, Sentiment and Intensity Prediction Using a Multi-Task Ensemble Framework. *IEEE Transactions on Affective Computing* 13: 285–297
- [27] Punetha N and Jain G 2023 Game theory and MCDM-based unsupervised sentiment analysis of restaurant reviews. *Applied Intelligence*:516–534.
- [28] Liu T, Deng Y and Chan F 2017 Evidential Supplier Selection Based on DEMATEL and Game Theory. *International Journal of Fuzzy Systems* 20: 1321–1333
- [29] Debnath A, Bandyopadhyay A, Roy J and Kar S 2018 Game theory based multi criteria decision making problem under uncertainty: a case study on Indian Tea Industry. *Journal of Business Economics and Management* 19: 154–175
- [30] Punetha N and Jain G 2023 Aspect and orientation-based sentiment analysis of customer feedback using mathematical optimization models. *Knowledge and Information Systems*:1–30
- [31] Ekinci Y, Orbay B Z and Karadayi M A 2022 An MCDM-based game-theoretic approach for strategy selection in higher education. *Socio-Economic Planning Sciences* 81: 101–186
- [32] Motlaghzadeh K, Eyni A and Behboudian M 2023 A multi-agent decision-making framework for evaluating water and environmental resources management scenarios under climate change. *Science of The Total Environment* 864: 161–168
- [33] Jaiswal A, Negi P and Singh N 2023 MCDM Computational Approaches for Green Supply Chain Management Strategies. *6th International Conference on Information Systems and Computer Networks (ISCON)*.1–9
- [34] Punetha N and Jain G 2023 Sentiment Analysis of Stock Prices and News Headlines Using the MCDM Framework. *AIST,IEEE Access*:1–4
- [35] Punetha N and Jain G 2023 Bayesian game model based unsupervised sentiment analysis of product reviews. *Expert Systems with Applications* 214: 119–128
- [36] Kolios A, Mytilinou V, Lozano-Minguez E and Salonitis K 2016 A comparative study of multiple-criteria decision-making methods under stochastic inputs. *Energies* 9: 23–45
- [37] Ghadikolaei A S, Esbouei S K and Antuchevičienė J 2014 Applying fuzzy MCDM for financial performance evaluation of Iranian companies. *Technological and Economic Development of Economy* 20: 274–291
- [38] Marzouk M and Control EA-S 2016 undefined Establishing multi-level performance condition indices for public schools maintenance program using AHP and fuzzy logic. *sic.ici.ro*:78–87
- [39] Li M, Jin L and Wang J 2014 A new MCDM method combining QFD with TOPSIS for knowledge management system selection from the user's perspective in intuitionistic fuzzy environment. *Applied Soft Computing*:28–37



- [40] Pavlovskis M and JA-S 2016 undefined Application of MCDM and BIM for evaluation of asset redevelopment solutions. *sic.ici.ro*:98-112
- [41] Beheshti M, Amoozad Mahdiraji H and Zavadskas E K 2016 Strategy portfolio optimisation: a copras g-modm hybrid approach. *Transformations In Business & Economics* 15: 500–519
- [42] Singh V K, Piryani R and Uddin A 2013 Sentiment analysis of textual reviews: Evaluating machine learning, unsupervised and sentiwordnet approaches. *International Conference on Knowledge and Smart Technology, KST* :122–127
- [43] Bhoir P and Kolte S 2016 Sentiment analysis of movie reviews using lexicon approach. 2015 *IEEE International Conference on Computational Intelligence and Computing Research, ICCIC*:78-87
- [44] Joshi A, R B A and Bhattacharyya P 2010 A Fall-back Strategy for Sentiment Analysis in Hindi: a Case Study. *ICON*:56-67
- [45] Kumar, Article Analyzing Sentiment In Indian Languages Micro Text Using Recurrent Neural Network. *IIOABJ*:313-318
- [46] Akhtar M S, Ekbal A and Bhattacharyya P 2016 Aspect based sentiment analysis in Hindi: Resource creation and evaluation, *LREC* 2703–2709
- [47] Mishra D, Venugopalan M and Gupta D 2016 Context Specific Lexicon for Hindi Reviews. *Procedia Computer Science* 93: 554–563
- [48] Jain V and Kashyap K L 2022 Ensemble hybrid model for Hindi COVID-19 text classification with metaheuristic optimization algorithm. *Multimedia Tools and Applications*.1-23
- [49] Jha V, Savitha R and Shenoy P D 2018 A novel sentiment aware dictionary for multi-domain sentiment classification. *Computers and Electrical Engineering* 69: 585–597
- [50] Singh J P, Rana N P and Alkhowaiter W 2015 Sentiment analysis of products' reviews containing English and Hindi texts. *Lecture Notes in Computer Science* 9373: 416–422
- [51] Kulkarni D S and Rodd S S 2021 Sentiment analysis in Hindi —A survey on the state-of-the-art techniques. *Transactions on Asian and Low-Resource Language Information Processing* :21-27

# Viscous fluid dynamics with decaying vacuum energy density

C. P. Singh\* and Vinita Khatri†

*Department of Applied Mathematics,  
Delhi Technological University, Delhi-110042, India*

(Dated: September 22, 2023)

In this work, we investigate the dynamics of bulk viscous models with decaying vacuum energy density (VED) in a spatially homogeneous and isotropic flat Friedmann-Lemaître-Robertson-walker (FLRW) spacetime. We particularly are interested to study the viscous model which considers first order deviation from equilibrium, i.e., the Eckart theory. In the first part, using the different forms of the bulk viscous coefficient, we find the main cosmological parameters, like Hubble parameter, scale factor, deceleration parameter and equation of state parameter analytically. We discuss some cosmological consequences of the evolutions and dynamics of the different viscous models with decaying VED. We examine the linear perturbation growth in the context of the bulk viscous model with decaying VED to see if it survives this further level of scrutiny. The second part of the work is devoted to constrain the viscous model of the form  $\zeta \propto H$ , where  $\zeta$  is the bulk viscous coefficient and  $H$  is the Hubble parameter, using three different combinations of data from type Ia supernovae (Pantheon),  $H(z)$  (cosmic chronometers), Baryon Acoustic Oscillation and  $f(z)\sigma_8(z)$  measurements with Markov Chain Monte Carlo (MCMC) method. We show that the considered model is compatible with the cosmological probes, and the  $\Lambda$ CDM recovered in late-time of the evolution of the Universe. Finally, we obtain selection information criteria (AIC and BIC) to study the stability of the models.

## I. INTRODUCTION

The different observations such as luminosity distances of type Ia supernova, measurements of anisotropy of cosmic microwave background and gravitational lensing have confirmed that our Universe is spatially flat and expanding with an accelerated rate. It has been observed that the Universe contains a mysterious dominant component, called dark energy (DE) with large negative pressure, which leads to this cosmic acceleration [1–7]. In literature, several models have been proposed to explain the current accelerated expansion of the Universe. The two most accepted DE models are that of a cosmological constant and a slowly varying rolling scalar field (quintessence models)[8–11].

The cosmological constant  $\Lambda$  (CC for short), initially introduced by Einstein to get the static Universe, is a natural candidate for explaining DE phenomena with equation of state parameter equal to  $-1$ . The natural interpretation of CC arises as an effect of quantum vacuum energy. Thus, the cold dark matter based cosmology together with a CC, called the  $\Lambda$ CDM cosmology, is preferred as the standard model for describing the current dynamics of the Universe. It is mostly consistent with the current cosmological observations. However, despite of its success, the  $\Lambda$ CDM model has several strong problems due to its inability to renormalize the energy density of quantum vacuum, obtaining a discrepancy of  $\sim 120$  orders of magnitude between its predicted and observed value, so-called CC or fine-tuning problem [12–14]. It also has the coincidence problem, i.e., why the Universe

transition, from decelerated to an accelerated phase, is produced at late times [15].

Many models have been proposed to tackle these issues. One of the possible proposal is to incorporate energy transfer among the cosmic components. In this respect, the models with time-varying vacuum energy density (VED), also known as ‘decaying vacuum cosmology’ seems to be promising. The idea of a time-varying VED models ( $\rho_\Lambda = \Lambda(t)/8\pi G$ ) is physically more viable than the constant  $\Lambda$  [16–19]. Although no fundamental theory exists to describe a time-varying vacuum, a phenomenological technique has been suggested to parametrize  $\Lambda(t)$ . In literature, many authors [20–40] have carried out analysis on decaying vacuum energy in which the time-varying vacuum has been phenomenologically modeled as a function of time in various possible ways, as a function of the Hubble parameter. Such attempts suggest that decaying VED model provides the possibility of explaining the acceleration of the Universe as well as it solves both cosmological constant and coincidence problems.

Shapiro and Solà [41], and Solà [42] proposed a possible connection between cosmology and quantum field theory on the basis of renormalization group (RG) which gives the idea of running vacuum models (RVM), characterized by VED  $\rho_\Lambda$ , see Refs.[32, 35, 39] for a review. The RVM has been introduced to solve the coincidence problem where the term  $\Lambda$  is assumed to be varying with the Hubble parameter  $H$ . Carnerio et al.[27] proposed that the vacuum term is proportional to the Hubble parameter,  $\Lambda(a) \propto H(a)$ . However, this model fails to fit the current CMB data. It is interesting to note that RG in quantum field theory (QFT) provides a time-varying vacuum, in which  $\Lambda(t)$  evolves as  $\Lambda \propto H^2$  [43]. Basilakos [28] proposed a parametrization of the functional form

\* cpsingh@dce.ac.in

† vinitakhatri\_2k20phdam501@dtu.ac.in

of  $\Lambda(t)$  by applying a power series expansion in  $H$  up to the second order. Recently, a large class of cosmologies has been discussed where VED evolves like a truncated power-series in the Hubble parameter  $H$ , see Refs.[44, 45] and references therein.

On the other hand, in recent years, the observations suggest that the Universe is permeated by dissipative fluids. Based on the thermodynamics point of view, phenomenological exotic fluids are supposed to play the role for an alternative DE models. It has been known since long time ago that a dissipative fluid can produce acceleration during the expansion of the Universe [46, 47]. The bulk and shear viscosity are most relevant parts of dissipative fluid. The bulk viscosity characterizes a change in volume of the fluid which is relevant only for the compressed fluids. The shear viscosity characterizes a change in shape of a fixed volume of the fluid which represents the ability of particles to transport momentum. In general, shear viscosity is usually used in connection with the spacetime anisotropy where as bulk viscosity plays the role in an isotropic cosmological models. The dynamics of homogeneous cosmological models has been studied in the presence of viscous fluid and has application in studying the evolution of the Universe.

Eckart [48] extended a classical irreversible thermodynamics from Newtonian to relativistic fluids. He proposed the simplest non-causal theory of relativistic dissipative phenomena of first order which was later modified by Landau and Lifshitz [49]. The Eckart theory has some important limitations. It has been found that all the equilibrium states are unstable [50] and the signals can propagate through the fluids faster than the speed of light [51]. Therefore, to resolve these issues, Israel and Stewart [52] proposed a full causal theory of second order. When the relaxation time goes to zero, the causal theory reduces to the Eckart's first order theory. Thus, taking the advantage of this limit of vanishing relaxation time at late time, it has been used widely to describe the recent accelerated expansion of the Universe. An exhaustive reviews on non-causal and causal theories of viscous fluids can be found in Refs.[53–66]. In recent years, the direct observations indicate for viscosity dominated late epoch of accelerating expansion of the Universe. In this respect, many authors have explored the viability of a bulk viscous Universe to explain the present accelerated expansion of the Universe cf.[67–88].

In Eckart theory, the effective pressure of the cosmic fluid is modeled as  $\Pi = -3\zeta H$ , where  $\zeta$  is bulk viscous coefficient and  $H$  the Hubble parameter. Bulk viscous coefficient can be assumed as a constant or function of Hubble parameter. It allows to explore the presence of interacting terms in the viscous fluid. Since the imperfect fluid should satisfy the equilibrium condition of thermodynamics, the pressure of the fluid must be greater than the one produced by the viscous term. To resolve this condition, it is useful to add an extra fluid such as cosmological constant. Many authors [89–93] have studied viscous cosmological models with constant or with time-

dependent cosmological constant. Hu and Hu [92] have investigated a bulk viscous model with cosmological constant by assuming bulk viscous proportional to the Hubble parameter. Herrera-Zamorano et al. [93] have studied a cosmological model filled with two fluids under Eckart formalism, a perfect fluid as DE mimicking the dynamics of the CC, while a non-perfect fluid as dark matter with viscosity term.

In this paper, we focus on discussing the dynamics of viscous Universe which consider the first order deviation from equilibrium, i.e., Eckart formalism with decaying VED. Using different versions of bulk viscous coefficient  $\zeta$ , we find analytically the main cosmological functions such as the scale factor, Hubble parameter, and deceleration and equation of state parameters. We discuss the effect of viscous model with varying VED in perturbation level. We implement the perturbation equation to obtain the growth of matter fluctuations in order to study the contribution of this model in structure formation. We perform a Bayesian Markov Chain Monte Carlo (MCMC) analysis to constrain the parameter spaces of the model using three different combinations involving observational data from type Ia supernovae (Pantheon), Hubble data (cosmic chronometers), Baryon acoustic oscillations and  $f(z)\sigma_8(z)$  measurements. We compare our model and concordance  $\Lambda$ CDM to understand the effects of viscosity with decaying vacuum by plotting the evolutions of the deceleration parameter, equation of state parameter and Hubble parameter. We also study the selection information criterion such as AIC and BIC to analyze the stability of the model.

The work of the paper is organized as follows. In Section II, we present the basic cosmological equations of Friedmann-Lemaître-Robertson-Walker (FLRW) geometry with bulk viscosity and decaying VED. In Section III, we find the solution of the field equations by assuming the various forms of bulk viscous coefficient. We discuss the growth rate equations that govern the perturbation in Section IV. Section V presents the observational data and method to be used to constrain the proposed model. The results and discussion on the evolution of the various parameters are presented in Section VI. In Section VII, we present the selection information criterion to distinguish the presented model with concordance  $\Lambda$ CDM. Finally, we conclude our finding in Section VIII.

## II. VISCOUS MODEL WITH VARYING- $\Lambda$

Let us start with the Friedmann-Lemaître-Robertson-Walker (FLRW) metric in the flat space geometry as the case favoured by observational data

$$ds^2 = -dt^2 + a^2(t) [dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2)], \quad (1)$$

where  $(r, \theta, \phi)$  are the co-moving coordinates and  $a(t)$  is the scale factor of the Universe. The large scale dynamics of (1) is described by the Einstein field equations, which

include the cosmological constant  $\Lambda$  and is given by

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = 8\pi G(T_{\mu\nu} + g_{\mu\nu}\rho_\Lambda), \quad (2)$$

where  $G_{\mu\nu}$  is the Einstein tensor,  $\rho_\Lambda = \Lambda/8\pi G$  is the vacuum energy density (the energy density associated to CC vacuum term) and  $T_{\mu\nu}$  is the energy-momentum tensor of matter. It is to be noted that for simplicity we use geometrical units  $8\pi G = c = 1$ . We introduce a bulk viscous fluid through the energy-momentum tensor which is given by [94]

$$T_{\mu\nu} = (\rho_m + P)u_\mu u_\nu + g_{\mu\nu}P, \quad (3)$$

where  $u^\mu$  is the fluid four-velocity,  $\rho_m$  is the density of matter and  $P$  is the pressure which is composed of the barotropic pressure  $p_m$  of the matter fluid plus the viscous pressure  $\Pi$ , i.e.,  $P = p_m + \Pi$ . The origin of bulk viscosity is assumed as a deviation of any system from the local thermodynamic equilibrium. According to the second law of thermodynamics, the re-establishment to the thermal equilibrium is a dissipative processes which generates entropy. Due to generation of entropy, there is an expansion in the system through a bulk viscous term.

In homogeneous and isotropic cosmological models, the viscous fluid is characterized by a bulk viscosity. It is mostly based on the Eckart's formalism [48] which can be obtained from the second order theory of non-equilibrium thermodynamics proposed by Israel and Stewart [52] in the limit of vanishing relaxation time. The viscous effect can be defined by the viscous pressure  $\Pi = -3\zeta H$ , where  $\zeta$  is the bulk viscous coefficient and  $H$  is the Hubble parameter. The bulk viscous coefficient  $\zeta$  is assumed to be positive on thermodynamical grounds. Therefore, it makes the effective pressure as a negative value which leads to modification in energy-momentum tensor of perfect fluid.

If we denote the total energy-momentum tensor  $T_{\mu\nu} + g_{\mu\nu}\rho_\Lambda$  as modified  $\tilde{T}_{\mu\nu}$  on right hand side of field equations (2), then the modified  $\tilde{T}_{\mu\nu}$  can be assumed the same form as  $T_{\mu\nu}$ , that is,  $\tilde{T}_{\mu\nu} = (\rho + p)u_\mu u_\nu + g_{\mu\nu}p$ , where  $\rho = \rho_m + \rho_\Lambda$  and  $p = p_m - 3\zeta H + p_\Lambda$  are the total energy density and pressure, respectively. Further, we assume that the bulk viscous fluid is the non-relativistic matter with  $p_m = 0$ . Thus, the contribution to the total pressure is only due to the sum of negative viscous pressure,  $-3\zeta H$  and vacuum energy pressure,  $p_\Lambda = -\rho_\Lambda$ .

Using the modified energy-momentum tensor as discussed above, the Einstein field equations (2) describing the evolution of FLRW Universe dominated by bulk viscous matter and vacuum energy yield

$$3H^2 = \rho = \rho_m + \rho_\Lambda, \quad (4)$$

$$2\dot{H} + 3H^2 = -p = 3\zeta H + \rho_\Lambda. \quad (5)$$

where  $H = \dot{a}/a$  is the Hubble parameter and an over dot represents the derivative with respect to cosmic time  $t$ .

In this paper, we propose the evolution of the Universe based on decaying vacuum models, i.e., vacuum energy density as a function of the cosmic time. From (2), the Bianchi identity  $\nabla^\mu G_{\mu\nu} = 0$  gives

$$\nabla^\mu \tilde{T}_{\mu\nu} = 0, \quad (6)$$

or, equivalently,

$$\dot{\rho}_m + 3H(\rho_m + p_m - 3\zeta H + \rho_\Lambda + p_\Lambda) = -\dot{\rho}_\Lambda, \quad (7)$$

which imply that there is a coupling between a dynamical  $\Lambda$  term and viscous CDM. Therefore, there is some energy exchange between the viscous CDM fluid and vacuum. Using the equation of state of the vacuum energy  $p_\Lambda = -\rho_\Lambda$  and  $p_m = 0$ , Eq. (7) leads to

$$\dot{\rho}_m + 3H(\rho_m - 3\zeta H) = -\dot{\rho}_\Lambda. \quad (8)$$

Now, combining Eqs.(4) and (8), we get

$$\dot{H} + \frac{3}{2}H^2 = \frac{1}{2}\rho_\Lambda + \frac{3}{2}\zeta H. \quad (9)$$

The dynamics of the Universe depends on the specific forms of  $\rho_\Lambda$  and  $\zeta$ .

### III. SOLUTION OF FIELD EQUATIONS

The evolution equation (9) has three independent unknown quantities, namely,  $H$ ,  $\zeta$  and  $\rho_\Lambda$ . We get the solution only if  $\zeta$  and  $\rho_\Lambda$  are specified. In this paper, we parameterize the functional form of  $\rho_\Lambda$  as a function of Hubble parameter. The motivation for a function  $\rho_\Lambda = \rho_\Lambda(H)$  can be assumed from different points of view. Although the correct functional form of  $\rho_\Lambda$  is not known, a quantum field theory (QFT) approach within the context of the renormalization group (RG) was proposed in Refs.[95, 96] and further studied by many authors [29, 32, 35, 42, 97, 98]. In Ref. [36], the following ratio has been defined between the two fluid components:

$$\gamma = \frac{\rho_\Lambda - \rho_{\Lambda_0}}{\rho_m + \rho_\Lambda}, \quad (10)$$

where  $\rho_{\Lambda_0}$  is a constant vacuum density. If  $\rho_\Lambda = \rho_{\Lambda_0}$ , then  $\gamma = 0$ , and we get  $\Lambda$ CDM model. On the other hand, if  $\rho_{\Lambda_0} \neq 0$ , then we get

$$\rho_\Lambda = \rho_{\Lambda_0} + \gamma(\rho_m + \rho_\Lambda) = \rho_{\Lambda_0} + 3\gamma H^2. \quad (11)$$

The above proposal was first considered by Shapiro and Sola [41] in context of RG. Many authors have studied the evolution of the Universe by assuming this form [33, 34, 40]. Hereafter, we shall focus on the simplest form of  $\rho_\Lambda$  which evolves with the Hubble rate. Specifically, in this paper we consider

$$\rho_\Lambda = c_0 + 3\nu H^2, \quad (12)$$

where  $c_0 = 3H_0^2(\Omega_{\Lambda 0} - \nu)$  is fixed by the boundary condition  $\rho_{\Lambda}(H_0) = \rho_{\Lambda 0}$ . The suffix ‘0’ denotes the present value of the parameter. The dimensionless coefficient  $\nu$  is the vacuum parameter and is expected to be very small value  $|\nu| \ll 1$ . A non-zero value of it makes possible the cosmic evolution of the vacuum.

The choice of  $\zeta$  generates different viscous models and in literature there are different approaches to assume the evolution of bulk viscosity. In this paper, we consider the most general form of the bulk viscous term  $\zeta$ , which is assumed to be the sum of three terms: the first term is a constant,  $\zeta_0$ , the second term is proportional to the Hubble parameter  $H = \dot{a}/a$  which is related to the expansion and the third term is proportional to the acceleration,  $\ddot{a}/\dot{a}$ . Thus, we assume the parametrization of bulk viscous coefficient in the form[72, 78, 99, 100]

$$\zeta = \zeta_0 + \zeta_1 \frac{\dot{a}}{a} + \zeta_2 \frac{\ddot{a}}{\dot{a}}, \quad (13)$$

where  $\zeta_0$ ,  $\zeta_1$  and  $\zeta_2$  are constants to be determined by the observations. The term  $\ddot{a}/\dot{a}$  in Eq. (13) can be written as  $\ddot{a}/aH$ . The basic idea about the assumption of  $\zeta$  in Eq.(13) is that the dynamic state of the fluid influences its viscosity in which the transport viscosity is related to the velocity and acceleration. In what follows, we study the decaying vacuum model defined in (12) with different forms of bulk viscous coefficient as defined in Eq.(13).

#### A. Cosmology with $\zeta = \zeta_0 = \text{const.}$

This is the simplest parametrization of Eckart’s bulk viscosity model. Many authors [69, 70, 77, 84, 87, 88, 91, 101–103] have studied the viscous cosmological models with constant bulk viscous coefficient. Using the decaying vacuum form (12) and taking  $\zeta = \zeta_0 = \text{const.}$ , where  $\zeta_1 = \zeta_2 = 0$  in Eq.(13), the evolution equation (9) reduces to

$$\dot{H} + \frac{3}{2}(1 - \nu)H^2 - \frac{3}{2}\zeta_0 H = \frac{1}{2}c_0. \quad (14)$$

Solving (14) for  $\nu < 1$ , we get

$$H = \frac{\zeta_0}{2(1 - \nu)} + \sigma \left( \frac{1 + e^{-3(1 - \nu)\sigma t}}{1 - e^{-3(1 - \nu)\sigma t}} \right), \quad (15)$$

where

$$\sigma = \sqrt{\left(\frac{\zeta_0}{2(1 - \nu)}\right)^2 + \frac{H_0^2(\Omega_{\Lambda 0} - \nu)}{(1 - \nu)}}. \quad \text{Here, we have used } c_0 = 3H_0^2(\Omega_{\Lambda 0} - \nu).$$

The above equation simplifies to give

$$H = \frac{\zeta_0}{2(1 - \nu)} + \sigma \coth\left(\frac{3}{2}(1 - \nu)\sigma t\right). \quad (16)$$

It can be observed that the solution reduces to the standard  $\Lambda$  for  $\zeta_0 = 0$  and  $\nu = 0$ , whereas for  $\zeta_0 = 0$  and  $\nu \neq 0$  it gives the solution for  $\Lambda(t)$  model from

quantum field theory[29]. Using the Hubble parameter  $H = \dot{a}/a$ , the scale factor of the model  $a(t)$  with the condition  $a(t_0) = 1$  is given by

$$a(t) = e^{\frac{\zeta_0}{2(1 - \nu)}t} \left( \sinh\left(\frac{3}{2}(1 - \nu)\sigma t\right) \right)^{\frac{2}{3(1 - \nu)}}, \quad (17)$$

which shows that the scale factor increases exponentially as  $t$  increases. From (17), one can observe that, in general, it is not possible to express cosmic time  $t$  in terms of the scale factor  $a$ . It is possible only if  $\zeta_0 = 0$ . In the absence of bulk viscosity, we obtain the result of decaying vacuum model as discussed in Ref.[29]. Further, for constant  $\Lambda$ , the solution reduced to the  $\Lambda$ CDM model with no viscosity.

To discuss the decelerated and accelerated phases and its transition during the evolution of the Universe, we study a cosmological parameter, known as ‘deceleration parameter’,  $q$ , which is defined as

$$q = -\frac{\ddot{a}}{a} \frac{1}{H^2} = -\left(1 + \frac{\dot{H}}{H^2}\right). \quad (18)$$

In cosmology,  $q$  is a dimensionless measure of the cosmic acceleration. The expansion of the Universe decelerates if  $q > 0$ , whereas it accelerates for  $q < 0$  and  $q = 0$  gives the marginal inflation. The time-dependent  $q$  may describe the transition from one phase to another phase. Using (16), the deceleration parameter is calculated as

$$q = -1 + \frac{3}{2} \frac{(1 - \nu)\sigma^2 \csc^2 h\left(\frac{3}{2}(1 - \nu)\sigma t\right)}{\left(\frac{\zeta_0}{2(1 - \nu)} + \sigma \coth\left(\frac{3}{2}(1 - \nu)\sigma t\right)\right)^2}. \quad (19)$$

For sake of completeness, we discuss another important cosmological parameter, known as effective equation of state (EoS) parameter, which is defined as

$$w_{eff} = -1 - \frac{2}{3} \frac{\dot{H}}{H^2}. \quad (20)$$

Using (16), we get

$$w_{eff} = -1 + \frac{(1 - \nu)\sigma^2 \csc^2 h\left(\frac{3}{2}(1 - \nu)\sigma t\right)}{\left(\frac{\zeta_0}{2(1 - \nu)} + \sigma \coth\left(\frac{3}{2}(1 - \nu)\sigma t\right)\right)^2}. \quad (21)$$

#### B. Cosmology with $\zeta = \zeta_1 H$

Let us consider the case where bulk viscous coefficient is proportional to the Hubble parameter, i.e.,  $\zeta = \zeta_1 H$ . Such a form of  $\zeta$  has been studied by many authors [53, 59, 72, 80, 104, 105]. This type of bulk viscous coefficient can be obtained by assuming  $\zeta_0 = \zeta_2 = 0$  in Eq.(13). Thus, using  $\zeta = \zeta_1 H$  and Eq.(12) into Eq.(9), we get the evolution equation for Hubble parameter as

$$\dot{H} + \frac{3}{2}(1 - \zeta_1 - \nu)H^2 - \frac{1}{2}c_0 = 0. \quad (22)$$

The above equation with change of a variable from  $t$  to  $x = \ln a$  can be written as

$$\frac{dh^2}{dx} + 3(1 - \zeta_1 - \nu)h^2 = 3(\Omega_{\Lambda 0} - \nu), \quad (23)$$

where  $h = H/H_0$  is the dimensionless Hubble parameter and  $\Omega_{\Lambda 0} = \rho_{\Lambda 0}/3H_0^2$ . Assuming  $(\zeta_1 + \nu) < 1$  and using the normalized scale factor-redshift relation,  $a = (1+z)^{-1}$ , we can express the normalized Hubble function  $E(z) \equiv H(z)/H_0$  as

$$E(z) = \frac{1}{(1 - \zeta_1 - \nu)^{1/2}} \times \left[ (1 - \zeta_1 - \Omega_{\Lambda 0})(1+z)^{3(1-\zeta_1-\nu)} + \Omega_{\Lambda 0} - \nu \right]^{1/2}. \quad (24)$$

From the above equation, it is clear that for  $\nu = 0$  and  $\zeta_1 = 0$ , we recover exactly the  $\Lambda$ CDM expansion model whereas only  $\zeta_1 = 0$  gives the solution obtained in Ref.[40]. It is observed that at very late time we get an cosmological constant dominated era,  $H \approx H_0 \sqrt{\frac{\Omega_{\Lambda 0} - \nu}{(1 - \zeta_1 - \nu)}}$ , which implies a de Sitter phase of the scale factor. Using  $H = \dot{a}/a$ , the solution for the scale factor in terms of cosmic time  $t$  is given by

$$a = \left( \frac{(1 - \zeta_1 - \Omega_{\Lambda 0})}{\Omega_{\Lambda 0} - \nu} \right)^{\frac{1}{3(1-\zeta_1-\nu)}} \times \left[ \sinh\left(\frac{3}{2}\sqrt{(1 - \zeta_1 - \nu)(\Omega_{\Lambda 0} - \nu)} H_0 t\right) \right]^{\frac{2}{3(1-\zeta_1-\nu)}} \quad (25)$$

It can be observed that the scale factor evolves as power-law expansion, i.e.,  $a \propto t^{2/3(1-\zeta_1-\nu)}$  for small values of  $t$  whereas it expands exponentially, i.e.,  $a \propto \exp\sqrt{\frac{(\Omega_{\Lambda 0} - \nu)}{3(1-\zeta_1-\nu)}} H_0 t$  for large values of time  $t$ . In other words, the model expands with decelerated rate in early time of its evolution and expands with accelerated rate in late time of its evolution.

From Eq. (25), we can find the cosmic time in terms of the scale factor, which is given by

$$t(a) = \frac{2}{3H_0\sqrt{(1 - \zeta_1 - \nu)(\Omega_{\Lambda 0} - \nu)}} \sinh^{-1} \left[ \left( \frac{a}{a_I} \right)^{\frac{3(1-\zeta_1-\nu)}{2}} \right] \quad (26)$$

where  $a_I = \left( \frac{(1-\zeta_1-\Omega_{\Lambda 0})}{(\Omega_{\Lambda 0}-\nu)} \right)^{1/3(1-\zeta_1-\nu)}$ .

Using (24), the value of  $q$  in terms of redshift is calculated as

$$q(z) = -1 + \frac{3}{2} \frac{(1 - \zeta_1 - \Omega_{\Lambda 0})(1+z)^{3(1-\zeta_1-\nu)}}{\left[ \frac{(\Omega_{\Lambda 0} - \nu)}{(1 - \zeta_1 - \nu)} + \left( 1 - \frac{(\Omega_{\Lambda 0} - \nu)}{(1 - \zeta_1 - \nu)} \right) (1+z)^{3(1-\zeta_1-\nu)} \right]} \quad (27)$$

The above equation shows that the dynamics of  $q$  depends on the redshift which describes the transition of

the Universe from decelerated to accelerated phase. We observe that as  $z \rightarrow -1$ ,  $q(z)$  approaches to  $-1$ . However, the model decelerates or accelerates if  $\Omega_{\Lambda 0} = \nu$ , which gives  $q = -1 + 1.5(1 - \zeta_1 - \nu)$ . Thus, a cosmological constant is required for a transition phase. Also, for  $z = 0$ , we find the present value of  $q$  which is given by

$$q_0 = -1 + 1.5(1 - \zeta_1 - \Omega_{\Lambda 0}). \quad (28)$$

The transition redshift,  $z_{tr}$  of the Universe, which is defined as a zero point of the deceleration parameter,  $q = 0$ , can be calculated as

$$z_{tr} = -1 + \left( \frac{2(\Omega_{\Lambda 0} - \nu)}{(3(1 - \zeta_1 - \nu) - 2)(1 - \zeta_1 - \Omega_{\Lambda 0})} \right)^{\frac{1}{3(1-\zeta_1-\nu)}}. \quad (29)$$

In this case, the effective EoS parameter is defined by  $w_{eff} = -1 - \frac{1}{3} \frac{d \ln h^2}{dx}$ , where  $x = \ln a$  and  $h = H/H_0$ . Using Eq. (24), we get

$$w_{eff}(z) = -1 + \frac{(1 - \zeta_1 - \Omega_{\Lambda 0})(1+z)^{3(1-\zeta_1-\nu)}}{\left[ \frac{(\Omega_{\Lambda 0} - \nu)}{(1 - \zeta_1 - \nu)} + \left( 1 - \frac{(\Omega_{\Lambda 0} - \nu)}{(1 - \zeta_1 - \nu)} \right) (1+z)^{3(1-\zeta_1-\nu)} \right]} \quad (30)$$

The present value of  $w_{eff}$  at  $z = 0$  is given by

$$w_{eff}(z = 0) = -1 + (1 - \zeta_1 - \Omega_{\Lambda 0}). \quad (31)$$

We can observe that the model will accelerate provided  $3w_{eff}(z = 0) + 1 = -2 + 3(1 - \zeta_1 - \Omega_{\Lambda 0}) < 0$ .

In Section IV, we will perform the observational analysis to estimate the parameters of the model and analyse the evolution and dynamics of the model in detail.

### C. Cosmology with $\zeta = \zeta_0 + \zeta_1 H$

In this subsection, we assume that the bulk viscous coefficient is a linear combination of two terms:  $\zeta_0$  and  $\zeta_1 H$ , i.e.,  $\zeta = \zeta_0 + \zeta_1 H$ . In literature, many authors [72, 78, 79] have assumed such a form of  $\zeta$  to study the dynamics of Universe. Using (12), Eq. (9) takes the form

$$\dot{H} + \frac{3}{2}(1 - \zeta_1 - \nu)H^2 - \frac{3}{2}\zeta_0 H = \frac{1}{2}c_0. \quad (32)$$

Assuming  $(\zeta_1 + \nu) < 1$ , we integrate (32) to obtain the solution for Hubble parameter which is given by

$$H = \frac{\zeta_0}{2(1 - \zeta_1 - \nu)} + \sigma_1 \left( \frac{1 + e^{-3(1-\zeta_1-\nu)\sigma_1 t}}{1 - e^{-3(1-\zeta_1-\nu)\sigma_1 t}} \right), \quad (33)$$

where

$$\sigma_1 = \sqrt{\left( \frac{\zeta_0}{2(1 - \zeta_1 - \nu)} \right)^2 + \frac{H_0^2(\Omega_{\Lambda 0} - \nu)}{(1 - \zeta_1 - \nu)}}.$$

On simplification, the above equation can be written as

$$H = \frac{\zeta_0}{2(1 - \zeta_1 - \nu)} + \sigma_1 \coth \left( \frac{3}{2}(1 - \zeta_1 - \nu)\sigma_1 t \right). \quad (34)$$



The corresponding expression for the scale factor in normalized unit has the form

$$a = e^{\frac{\zeta_0}{2(1-\zeta_1-\nu)}t} \left[ \sinh \left( \frac{3}{2}(1-\zeta_1-\nu)\sigma_1 t \right) \right]^{\frac{2}{3(1-\zeta_1-\nu)}}. \quad (35)$$

The respective deceleration parameter and effective EoS parameter are calculated as

$$q = -1 + \frac{3(1-\zeta_1-\nu)\sigma_1^2 \csc^2 h(\frac{3}{2}(1-\zeta_1-\nu)\sigma_1 t)}{2 \left( \frac{\zeta_0}{2(1-\zeta_1-\nu)} + \sigma_1 \coth(\frac{3}{2}(1-\zeta_1-\nu)\sigma_1 t) \right)^2} \quad (36)$$

and

$$w_{eff} = -1 + \frac{(1-\zeta_1-\nu)\sigma_1^2 \csc^2 h(\frac{3}{2}(1-\zeta_1-\nu)\sigma_1 t)}{\left( \frac{\zeta_0}{2(1-\zeta_1-\nu)} + \sigma_1 \coth(\frac{3}{2}(1-\zeta_1-\nu)\sigma_1 t) \right)^2} \quad (37)$$

#### D. Cosmology with $\zeta = \zeta_0 + \zeta_1 H + \zeta_2 (\ddot{a}/aH)$

Lastly, we assume a more general form of bulk viscous coefficient which is a combination of three terms:  $\zeta_0$ ,  $\zeta_1 H$  and  $\zeta_2 \ddot{a}/aH$ . This generalized form of  $\zeta$  is well motivated as discussed earlier and has been studied by many authors [66, 71, 72, 81, 99, 100]. This form of  $\zeta$  can be rewritten as

$$\zeta = \zeta_0 + \zeta_1 H + \zeta_2 \left( \frac{\dot{H}}{H} + H \right). \quad (38)$$

Using Eqs.(38) and (12), Eq.(9) reduces to

$$(1 - \frac{3}{2}\zeta_2)\dot{H} + \frac{3}{2}(1-\zeta_1-\zeta_2-\nu)H^2 - \frac{3}{2}\zeta_0 H - \frac{1}{2}c_0 = 0, \quad (39)$$

which on integration, it gives

$$H = \frac{\zeta_0}{2(1-\zeta_1-\zeta_2-\nu)} + \sigma_2 \coth \left( \frac{3}{2} \frac{(1-\zeta_1-\zeta_2-\nu)\sigma_2}{(1-\frac{3}{2}\zeta_2)} t \right), \quad (40)$$

where

$$\sigma_2 = \sqrt{\left( \frac{\zeta_0}{2(1-\zeta_1-\zeta_2-\nu)} \right)^2 + \frac{(1-\frac{3}{2}\zeta_1)H_0^2(\Omega_{\Lambda 0}-\nu)}{(1-\zeta_1-\zeta_2-\nu)}}. \quad \text{The solution for the scale factor can be obtained as}$$

$$a = e^{\frac{\zeta_0}{2(1-\zeta_1-\zeta_2-\nu)}t} \left[ \sinh \left( \frac{3}{2} \frac{(1-\zeta_1-\zeta_2-\nu)\sigma_2}{(1-\frac{3}{2}\zeta_2)} t \right) \right]^{\frac{2(1-\frac{3}{2}\zeta_2)}{3(1-\zeta_1-\zeta_2-\nu)}} \quad (41)$$

The deceleration parameter and effective EoS parameter are calculated as

$$q = -1 + \frac{\frac{(1-\zeta_1-\zeta_2-\nu)\sigma_2^2 \csc^2 h(\frac{3}{2} \frac{(1-\zeta_1-\zeta_2-\nu)\sigma_2}{(1-\frac{3}{2}\zeta_2)} t)}{\left( \frac{\zeta_0}{2(1-\zeta_1-\zeta_2-\nu)} + \sigma_2 \coth(\frac{3}{2} \frac{(1-\zeta_1-\zeta_2-\nu)\sigma_2}{(1-\frac{3}{2}\zeta_2)} t) \right)^2}} \quad (42)$$

and

$$w_{eff} = -1 + \frac{\frac{(1-\zeta_1-\zeta_2-\nu)\sigma_2^2 \csc^2 h(\frac{3}{2} \frac{(1-\zeta_1-\zeta_2-\nu)\sigma_2}{(1-\frac{3}{2}\zeta_2)} t)}{\left( \frac{\zeta_0}{2(1-\zeta_1-\zeta_2-\nu)} + \sigma_2 \coth(\frac{3}{2} \frac{(1-\zeta_1-\zeta_2-\nu)\sigma_2}{(1-\frac{3}{2}\zeta_2)} t) \right)^2}} \quad (43)$$

#### IV. GROWTH OF PERTURBATIONS

In cosmic structure formation it is assumed that the present abundant structure of the Universe is developed through gravitational amplification of small density perturbations generated in its early evolution. In this section, we briefly discuss the linear perturbation within the framework of viscous fluid with varying  $\Lambda(t)$ . We refer the reader to Refs. [106, 107] for the detailed perturbation equations since here we have discussed some basic equations only. The differential equation for the matter density contrast  $\delta_m \equiv \delta\rho_m/\rho_m$  for our model considered here can be approximated as follows [108]:

$$\delta_m'' + \left( \frac{3}{a} + \frac{H'(a)}{H(a)} \right) \delta_m' - \frac{4\pi G \rho_m}{H^2(a)} \frac{\delta_m}{a^2} = 0 \quad (44)$$

where prime represents derivative with respect to the scale factor  $a$ . The above second-order differential equation turns out to be accurate since the main effects come from the different expression of the Hubble function. We consider the Hubble function as obtained in Part B of Sect. III. Equation (44) describes the smoothness of the matter perturbation in extended viscous  $\Lambda(t)$  model.

The linear growth rate of the density contrast,  $f$ , which is related to the peculiar velocity in the linear theory [109] is defined as

$$f(a) = \frac{d \ln D_m(a)}{d \ln a}, \quad (45)$$

where  $D_m(a) = \delta_m(a)/\delta_m(a=1)$  is the linear growth function. The weighted linear growth rate, denoted by  $f\sigma_8$ , is the product of the growth rate  $f(z)$ , defined in (45), and  $\sigma_8(z)$ . Here,  $\sigma_8$  is the root-mean-square fluctuation in spheres with radius  $8h^{-1}$  Mpc scales [110, 111], and it is given by [112]

$$\sigma_8(z) = \frac{\delta_m(z)}{\delta_m(z=0)} \sigma_8(z=0). \quad (46)$$

Using (45) and (46), the weighted linear growth rate is given by

$$f\sigma_8(z) = -(1+z) \frac{\sigma_8(z=0)}{\delta_m(z=0)} \frac{d\delta_m}{dz}. \quad (47)$$

#### V. DATA AND METHODOLOGY

In this section, we present the data and methodology used in this work. We constrain the parameters of the

$GR - \Lambda$ CDM and  $\zeta = \zeta_1 H$  with varying  $\Lambda$  models using a large, robust and latest set of observational data which involve observations from: (i) distant type Ia supernovae (SNe Ia); (ii) a compilation of cosmic chronometer measurements of Hubble parameter  $H(z)$  at different redshifts; (iii) baryonic acoustic oscillations (BAO); and (iv)  $f(z)\sigma_8(z)$  data. A brief description of each of datasets are as follows:

#### A. Pantheon SNe Ia sample

The most known and frequently used cosmological probe are distant type Ia supernovae (SNe Ia) which are used to understand the actual evolution of the Universe. A supernova explosion is an extremely luminous event, with its brightness being comparable with the brightness of its host galaxy [113]. We use the recent SNe Ia data points, the so-called Pantheon sample which includes 1048 data points of luminosity distance in the redshift range  $0.01 < z < 2.26$ . Specifically, one could use the observed distance modulo,  $\mu_{obs}$ , to constrain cosmological models. The Chi-squared function for SNe Ia is given by

$$\chi_{SNe\,Ia}^2 = \sum_{i=1}^{1048} \Delta\mu^T C^{-1} \Delta\mu, \quad (48)$$

where  $\Delta\mu = \mu_{obs} - \mu_{th}$ . Here,  $\mu_{obs}$  is the observational distance modulus of SNe Ia and is given as  $\mu_{obs} = m_B - \mathcal{M}$ , where  $m_B$  is the observed peak magnitude in the rest frame of the  $B$  band,  $\mathcal{M}$  is the absolute B-band magnitude of a fiducial SNe Ia, which is taken as  $-19.38$ . The theoretical distance modulus  $\mu_{th}$  is defined by

$$\mu_{th}(z, \mathbf{p}) = 5 \log_{10} \left( \frac{D_L(z_{hel}, z_{cmb})}{1 \text{Mpc}} \right) + 25, \quad (49)$$

where  $\mathbf{p}$  is the parameter space and  $D_L$  is the luminosity distance, which is given as  $D_L(z_{hel}, z_{cmb}) = (1 + z_{hel})r(z_{cmb})$ . Here,  $r(z_{cmb})$  is given by

$$r(z) = cH_0^{-1} \int_0^z \frac{dz'}{E(z', \mathbf{p})}, \quad (50)$$

where  $c$  is the speed of light,  $E(z) \equiv H(z)/H_0$  is the dimensionless Hubble parameter,  $z_{hel}$  and  $z_{cmb}$  are heliocentric and CMB frame redshifts, respectively. Here,  $C$  is the total covariance matrix which takes the form  $C = D_{stat} + C_{sys}$ , where the diagonal matrix  $D_{stat}$  and covariant matrix  $C_{sys}$  denote the statistical uncertainties and the systematic uncertainties.

#### B. BAO measurements

In this work, we have used six points of BAO datasets from several surveys, which includes the Six Degree

Field Galaxy Survey (6dFGS), the Sloan Digital Sky Survey (SDSS), and the LOWZ samples of the Baryon Oscillation Spectroscopic Survey (BOSS) [114–116].

The dilation scale  $D_v(z)$  introduced in [117] is given by

$$D_v(z) = \left( \frac{d_A^2(z)z}{H(z)} \right)^{1/3} \quad (51)$$

Here,  $d_A(z)$  is the comoving angular diameter distance and is defined as

$$d_A(z) = \int_0^z \frac{dy}{H(y)}, \quad (52)$$

Now, the corresponding Chi-squared function for the BAO analysis is given by

$$\chi_{BAO}^2 = A^T C_{BAO}^{-1} A, \quad (53)$$

where  $A$  depend on the considered survey and  $C_{BAO}^{-1}$  is the inverse of the covariance matrix [116].

#### C. $H(z)$ data

The cosmic chronometer (CC) data, which is determined by using the most massive and passively evolving galaxies based on the ‘galaxy differential age’ method, are model independent (see, Ref.[118] for detail). In our analysis, we use 32 CC data points of the Hubble parameter measured by differential age technique [118] between the redshift range  $0.07 \leq z \leq 1.965$ . The Chi-squared function for  $H(z)$  is given by

$$\chi_{H(z)}^2 = \sum_{i=1}^{32} \frac{[H(z_i, \mathbf{p}) - H_{obs}(z_i)]^2}{\sigma_{H(z_i)}^2} \quad (54)$$

where  $H(z_i, \mathbf{p})$  represents the theoretical values of Hubble parameter with model parameters,  $H_{obs}(z_i)$  is the observed values of Hubble parameter and  $\sigma_i$  represents the standard deviation measurement uncertainty in  $H_{obs}(z_i)$ .

#### D. $f(z)\sigma_8(z)$ data

In Section IV, we have mainly discussed the background evolution of the growth perturbations and defined the weighted linear growth rate by Eq. (47). To make more complete discussion on viscous  $\Lambda(t)$  model in perturbation evolution, we focus on an observable quantity of  $f(z)\sigma_8(z)$ . We use 18 data points of ‘Gold -17’ compilation of robust and independent measurements of weighted linear growth  $f(z)\sigma_8(z)$  obtained by various galaxy surveys as compiled in Table III of Ref. [119]. In order to compare the observational data set with that

TABLE I. Constraints on parameters of  $\Lambda$ CDM for different set of observation data. Here “BASE” denotes “SNe Ia+BAO”

$\Lambda$ CDM			
Parameter	BASE	+CC	+ $f\sigma_8$
$H_0$	$68.987^{+0.263}_{-0.276}$	$69.001^{+0.238}_{-0.223}$	$68.793^{+0.193}_{-0.221}$
$\Omega_\Lambda$	$0.701^{+0.013}_{-0.020}$	$0.699^{+0.016}_{-0.015}$	$0.684^{+0.015}_{-0.014}$
$\sigma_8$	—	—	$0.794^{+0.014}_{-0.015}$
$S_8$	—	—	$0.811^{+0.022}_{-0.022}$
$z_{tr}$	$0.670^{+0.038}_{-0.038}$	$0.674^{+0.035}_{-0.035}$	$0.625^{+0.041}_{-0.041}$
$q_0$	$-0.549^{+0.020}_{-0.023}$	$-0.551^{+0.020}_{-0.020}$	$-0.523^{+0.025}_{-0.025}$
$w_0$	$-0.699^{+0.013}_{-0.015}$	$-0.701^{+0.013}_{-0.013}$	$-0.682^{+0.017}_{-0.017}$
$t_0(Gyr)$	$13.73^{+0.017}_{-0.017}$	$13.69^{+0.015}_{-0.015}$	$13.54^{+0.013}_{-0.013}$

of the predicted by our model, we define the Chi-square function as

$$\chi^2_{(f\sigma_8)} = \sum_{i=1}^{18} \frac{[f\sigma_8^{the}(z_i, \mathbf{p}) - f\sigma_8^{obs}(z_i)]^2}{\sigma_{f\sigma_8(z_i)}^2}, \quad (55)$$

where  $f\sigma_8^{the}(z_i, \mathbf{p})$  is the theoretical value computed by Eq.(47) and  $f\sigma_8^{obs}(z_i)$  is the observed data [119].

Using the observational data as discussed above, we use the Markov Chain Monte Carlo (MCMC) method by employing EMCEE python package [120] to explore the parameter spaces of viscous model with decaying vacuum density as discussed in part B of Sect.III by utilizing different combinations of data sets. The combinations are as follows:

- **BASE:** The combination of two datasets *SNe Ia* + *BAO* is termed as “BASE”, whose the joint  $\chi^2$  function is defined as  $\chi^2_{tot} = \chi^2_{SNe Ia} + \chi^2_{BAO}$ .
- **+CC:** We combine *CC* data to the BASE, where  $\chi^2_{tot} = \chi^2_{SNe Ia} + \chi^2_{BAO} + \chi^2_{H(z)}$
- **+ $f\sigma_8(z)$ :** The BASE data is complemented with *CC* and  $f\sigma_8$ , where  $\chi^2_{tot} = \chi^2_{SNe Ia} + \chi^2_{BAO} + \chi^2_{H(z)} + \chi^2_{f\sigma_8}$ .

We consider the  $\Lambda$ CDM model as a reference model and its parameters are also constrained with the above sets of data.

## VI. RESULTS AND DISCUSSION

In this section, we present the main results obtained through the observational data on the viscous  $\Lambda(t)$  model of the form  $\zeta = \zeta_1 H$  with  $\Lambda = c_0 + 3\nu H^2$  (Refers to part B of Sect.III). We also present the cosmological observation for  $\Lambda$ CDM model using the three combination of datasets. The viscous  $\Lambda(t)$  model has 4 free parameter spaces  $\{H_0, \Omega_\Lambda, \zeta_1, \nu\}$ , where as  $\Lambda$ CDM has 2 free parameters  $\{H_0, \Omega_\Lambda\}$ . We calculate the best-fit values by minimizing the combination of  $\chi^2$  function for above defined data sets. We also provide the fitting values of the

TABLE II. Constraints on parameters of viscous  $\Lambda(t)$  model using different set of observation data.

Viscous $\Lambda(t)$			
Parameter	BASE	+CC	+ $f\sigma_8$
$H_0$	$68.843^{+0.274}_{-0.238}$	$68.913^{+0.262}_{-0.261}$	$68.684^{+0.259}_{-0.241}$
$\Omega_\Lambda$	$0.680^{+0.018}_{-0.020}$	$0.684^{+0.013}_{-0.020}$	$0.674^{+0.012}_{-0.016}$
$\zeta_1$	$0.006^{+0.007}_{-0.004}$	$0.006^{+0.008}_{-0.004}$	$0.003^{+0.005}_{-0.002}$
$\nu$	$0.004^{+0.003}_{-0.003}$	$0.003^{+0.004}_{-0.002}$	$0.003^{+0.004}_{-0.002}$
$\sigma_8$	—	—	$0.790^{+0.008}_{-0.010}$
$S_8$	—	—	$0.822^{+0.019}_{-0.019}$
$z_{tr}$	$0.664^{+0.031}_{-0.042}$	$0.665^{+0.031}_{-0.037}$	$0.626^{+0.028}_{-0.038}$
$q_0$	$-0.533^{+0.025}_{-0.020}$	$-0.535^{+0.023}_{-0.020}$	$-0.516^{+0.022}_{-0.017}$
$w_0$	$-0.689^{+0.017}_{-0.013}$	$-0.690^{+0.015}_{-0.013}$	$-0.677^{+0.014}_{-0.011}$
$t_0(Gyr)$	$13.52^{+0.019}_{-0.019}$	$13.48^{+0.017}_{-0.017}$	$13.47^{+0.013}_{-0.015}$

$\Lambda$ CDM for comparison with the viscous  $\Lambda(t)$  model. The constraints of the statistical study are presented in Tables I and II. Figures 1-3 show the  $1\sigma(68.3\%)$  and  $2\sigma(95.4\%)$  confidence level (CL) contours with marginalized likelihood distributions for the cosmological parameters of  $\Lambda$ CDM and viscous  $\Lambda(t)$  models considering combination of different datasets, respectively. It is observed from Tables I and II that the constraints on the parameter spaces of  $\Lambda$ CDM and viscous with  $\Lambda(t)$  are nearly the same.

Using best-fit values of parameters obtained from *BASE*, *+CC* and *+ $f\sigma_8$*  data into Eq.(27), the evolutions of the deceleration parameter with respect to the redshift are shown in Figs.4-6 for viscous  $\Lambda(t)$  model along with the  $\Lambda$ CDM model. It is observed that with each data set  $q(z)$  varies from positive to negative and show the similar trajectory that is comparable to the  $\Lambda$ CDM model. Thus, both the model depict a transition from the early decelerated phase to the late-time accelerated phase. Further,  $q(z)$  approaches to  $-1$  in late-time of evolution. Thus, the models successfully generate late-time cosmic acceleration along with a decelerated expansion in the past. Figures 4-6 also show that the transition from decelerated to accelerated phase take place at redshift  $z_{tr} = 0.664^{+0.031}_{-0.042}$  with *BASE* data,  $z_{tr} = 0.665^{+0.031}_{-0.037}$  with *+CC* data and  $z_{tr} = 0.626^{+0.028}_{-0.037}$  with *+ $f\sigma_8$*  data. The datasets *BASE*, *+CC* and *+ $f\sigma_8$*  yield the present deceleration parameter  $q_0$  as  $-0.533^{+0.025}_{-0.020}$ ,  $-0.535^{+0.023}_{-0.020}$  and  $-0.516^{+0.022}_{-0.017}$  respectively (cf. Table II). The present values of  $z_{tr}$  and  $q_0$  are very close and thus are in good agreement to  $\Lambda$ CDM as presented in Table I.

The evolutions of the Hubble parameter  $H(z)$  of viscous  $\Lambda(t)$  model with respect to the redshift are shown in Figs. 7-9. Throughout the expansion, viscous  $\Lambda(t)$  is coinciding with the  $\Lambda$ CDM model and the model paths cover majority of the dataset with the error bar of Hubble parameter, indicating that the viscous  $\Lambda(t)$  agrees well with the  $\Lambda$ CDM model for all the three combination of datasets. In the considered cosmological scenario, the present age of the Universe are found to be  $t_0 \approx 13.52 Gyr$ ,  $t_0 \approx 13.48 Gyr$  and  $t_0 \approx 13.47 Gyr$  respectively as presented in Table II. The ages thus obtained are very much compatible with that obtained from

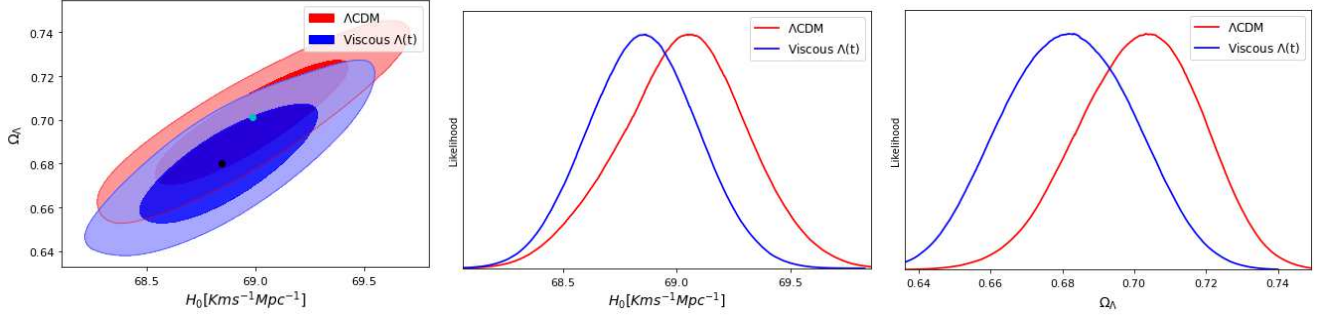


FIG. 1. Two-dimensional confidence contours of the  $H_0 - \Omega_\Lambda$  and one dimensional posterior distributions of  $H_0$ ,  $\Omega_\Lambda$  for the  $\Lambda$ CDM and viscous  $\Lambda(t)$  models using “*BASE*” data. The green and black dot on the contour represents the best fit value of  $\Lambda$ CDM and viscous  $\Lambda(t)$  models respectively.

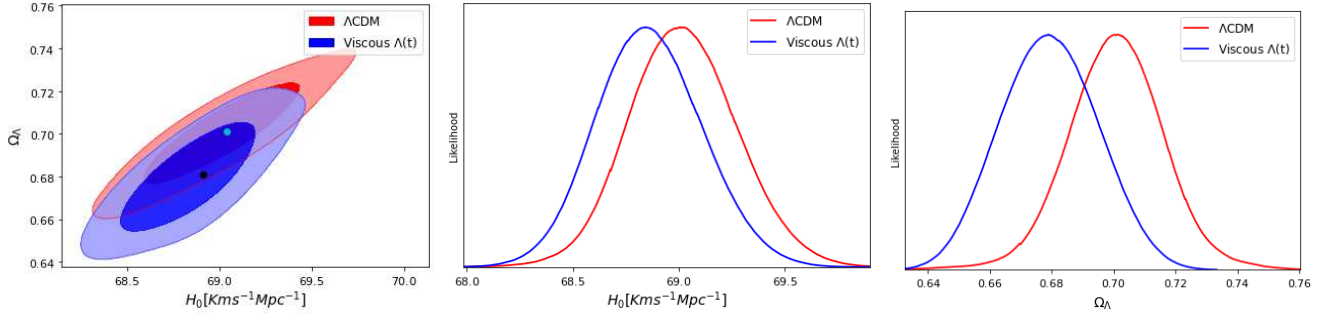


FIG. 2. Two-dimensional confidence contours of the  $H_0 - \Omega_\Lambda$  and one dimensional posterior distributions of  $H_0$ ,  $\Omega_\Lambda$  for the  $\Lambda$ CDM and viscous  $\Lambda(t)$  models using “*+CC*” data. The green and black dot on the contour represents the best fit value of  $\Lambda$ CDM and viscous  $\Lambda(t)$  models respectively.

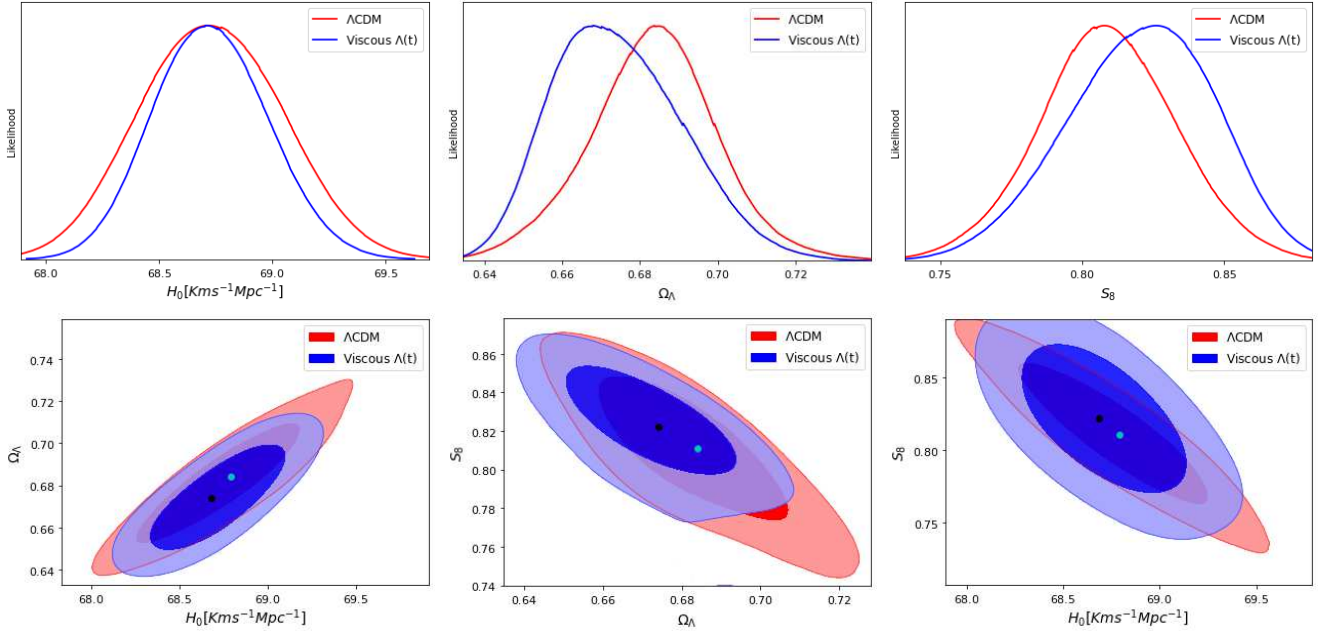


FIG. 3. Two-dimensional confidence contours of  $H_0 - \Omega_\Lambda$ ,  $\Omega_\Lambda - S_8$  and  $H_0 - S_8$  and one-dimensional posterior distributions of  $H_0$ ,  $\Omega_\Lambda$  and  $S_8$  for the  $\Lambda$ CDM and viscous  $\Lambda(t)$  models using “*+f\sigma\_8*” data. The green and black dot on the contour represents the best fit value of  $\Lambda$ CDM and viscous  $\Lambda(t)$  models respectively.

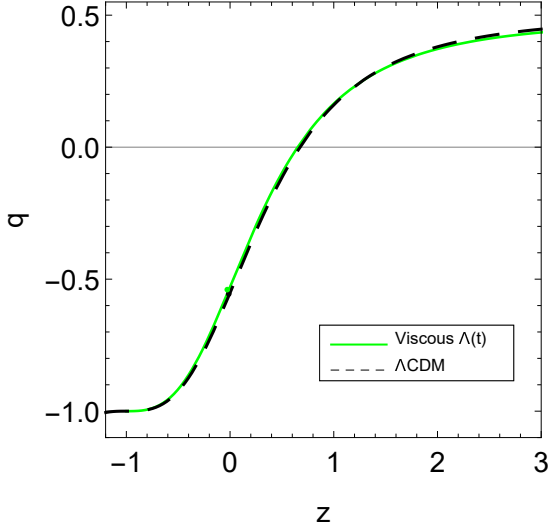


FIG. 4. The redshift evolution of the deceleration parameter for viscous  $\Lambda(t)$  using “*BASE*” dataset. The evolution of deceleration parameter in the standard  $\Lambda$ CDM model is also shown as the dashed curve. A dot denotes the current value of  $q$  (hence  $q_0$ ).

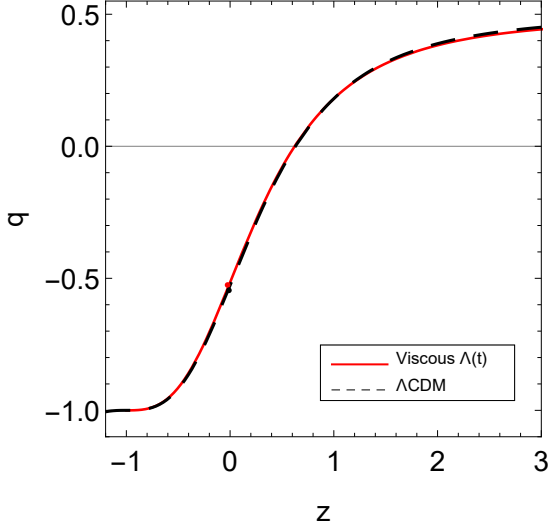


FIG. 6. The redshift evolution of the deceleration parameter for viscous  $\Lambda(t)$  using “ $+f\sigma_8$ ” dataset. The evolution of deceleration parameter in the standard  $\Lambda$ CDM model is also shown as the dashed curve. A dot denotes the current value of  $q$  (hence  $q_0$ ).

the  $\Lambda$ CDM model with the same datasets (cf. Table I).

Using the best-fit values of parameters in Eq. (30), the evolutions of the effective EoS parameter  $w_{eff}$  are shown in Figs.10-12. We conclude that for large redshifts,  $w_{eff}$  has small negative value  $w_{eff} > -1/3$  and in future the model asymptotically approaches to  $w_{eff} = -1$ . The trajectory of  $w_{eff}$  for *BASE* and *+CC* datasets coincides with the evolution of  $\Lambda$ CDM model. However, it

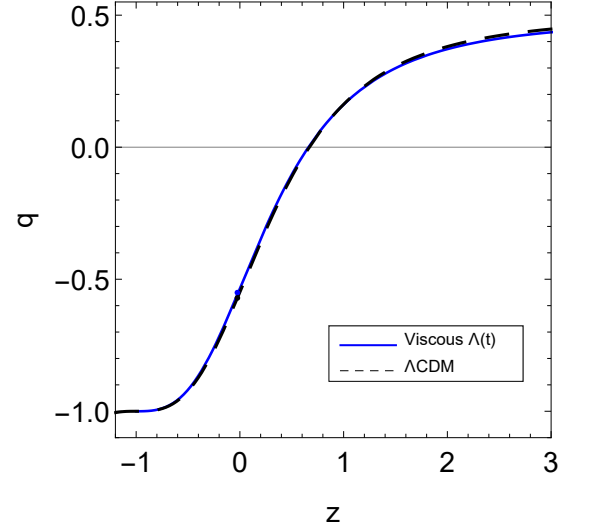


FIG. 5. The redshift evolution of the deceleration parameter for viscous  $\Lambda(t)$  using “ $+CC$ ” dataset. The evolution of deceleration parameter in the standard  $\Lambda$ CDM model is also shown as the dashed curve. A dot denotes the current value of  $q$  (hence  $q_0$ ).

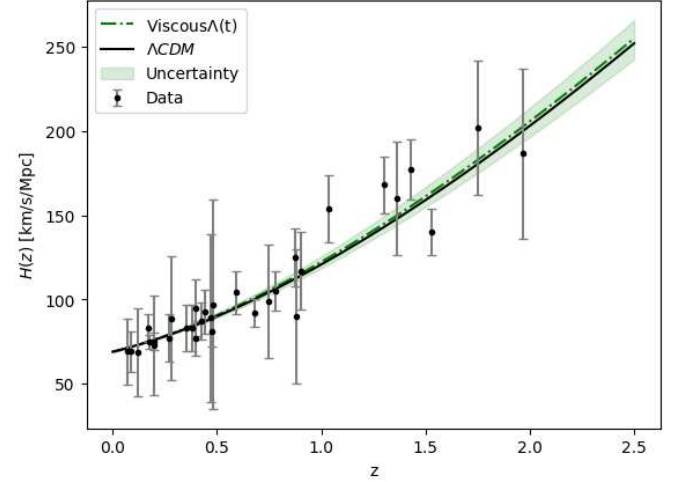


FIG. 7. Best fits using “*BASE*” data set over  $H(z)$  data for viscous  $\Lambda(t)$  (green dot-dashed line) and  $\Lambda$ CDM (black solid line) are shown. The grey points with uncertainty bars correspond to the 32 *CC* sample.

slightly varies with the best-fit values obtained through  $+f\sigma_8(z)$  data points. It can be observed that the viscous  $\Lambda(t)$  model behaves like a quintessence in early time and cosmological constant in late-time. The present values of  $w_{eff}$  are found to be  $-0.689^{+0.017}_{-0.013}$ ,  $-0.690^{+0.015}_{-0.013}$  and  $-0.677^{+0.014}_{-0.011}$  with *BASE*, *+CC* and *+f $\sigma_8$*  datasets respectively, which are very close to the current value of  $\Lambda$ CDM model as presented in Table I.

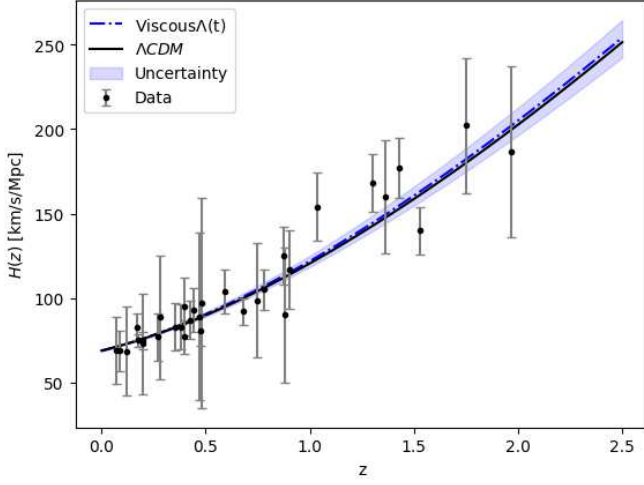


FIG. 8. Best fits using “+CC” data set over  $H(z)$  data for viscous  $\Lambda(t)$  (blue dot-dashed line) and  $\Lambda$ CDM (black solid line) are shown. The grey points with uncertainty bars correspond to the 32 CC sample.

From Tables I and II, let us discuss the present value  $H_0$  of Hubble parameter in case of viscous  $\Lambda(t)$  and  $\Lambda$ CDM models. The viscous  $\Lambda(t)$  model gives  $H_0 = 68.843^{+0.274}_{-0.238}$  km/s/Mpc with BASE data, the +CC data gives  $H_0 = 68.913^{+0.262}_{-0.261}$  km/s/Mpc and, finally, the  $+f\sigma_8$  renders the present value:  $H_0 = 68.684^{+0.259}_{-0.241}$  km/s/Mpc. Recently, the local measurement  $H_0 = 73.04 \pm 1.04$  km/s/Mpc from Riess et al.[121] exhibits a strong tension with the Planck 2018 release  $H_0 = 67.4 \pm 0.5$  km/s/Mpc [7] at the  $4.89\sigma$  confidence level. The residual tensions of our fitting results with respect to the latest local measurement  $H_0 = 73.04 \pm 1.04$  km/s/Mpc [121] are  $3.92\sigma$ ,  $3.85\sigma$  and  $4.07\sigma$  respectively.

Let us focus on  $\sigma_8$  and  $S_8$  which play very relevant role in structure formation. The best-fit values of these parameters for  $\Lambda$ CDM and viscous  $\Lambda(t)$  models using BASE + CC +  $f\sigma_8$  data are reported in Tables I and II, respectively. We can read off  $\sigma_8 = 0.794^{+0.014}_{-0.015}$  for  $\Lambda$ CDM model (cf. Table I), whereas the viscous  $\Lambda(t)$  model prediction is  $\sigma_8 = 0.790^{+0.008}_{-0.010}$  (cf. Table II). This is a very good result, which can be rephrased in terms of the fitting value of the related LSS observable  $S_8 = \sigma_8 \sqrt{(1 - \Omega_\Lambda)/0.3}$  quoted in the Tables I and II:  $S_8 = 0.811 \pm 0.022$  for  $\Lambda$ CDM and  $S_8 = 0.822 \pm 0.019$  for viscous  $\Lambda(t)$  model. The values of  $\sigma_8$  and  $S_8$  for viscous  $\Lambda(t)$  model is compatible for  $1\sigma$  confidence level with  $\Lambda$ CDM. Our result predicts that the tensions in  $\sigma_8$  and  $S_8$  are reduced to  $0.23\sigma$  and  $-0.38\sigma$ , respectively. The behavior of  $f(z)\sigma_8(z)$  as a function of redshift is plotted in Fig.14. We can see that the evolution of  $f\sigma_8$  for both viscous  $\Lambda(t)$  and  $\Lambda$ CDM models are consistent with the observational data points.

Table III presents the  $\chi^2$  and reduced  $\chi^2$  of  $\Lambda$ CDM and viscous  $\Lambda(t)$  models, respectively for the used datasets. To compute reduced  $\chi^2$ , denoted as  $\chi^2_{red}$ , we use  $\chi^2_{red}$

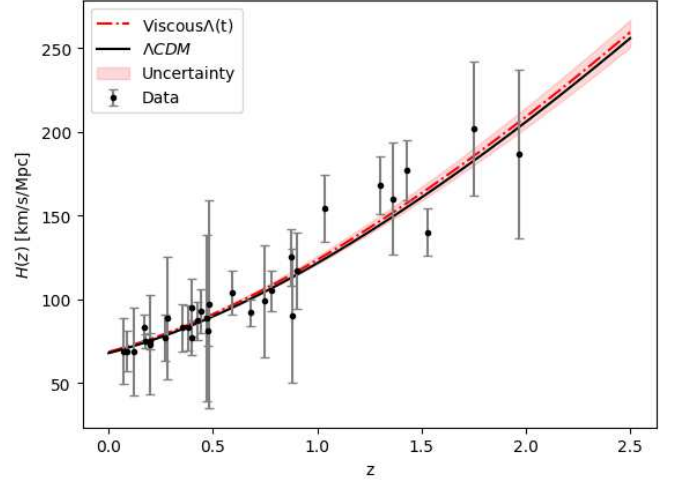


FIG. 9. Best fits using “+  $f\sigma_8$ ” data set over  $H(z)$  data for viscous  $\Lambda(t)$  (red dot-dashed line) and  $\Lambda$ CDM (black solid line) are shown. The grey points with uncertainty bars correspond to the 32 CC sample.

$= \chi^2_{min}/(N - d)$ , where  $N$  is the total number of data points and  $d$  is the total number of fitted parameters, which differs for the various models. It should be noted that when a model is fitted to data, a value of  $\chi^2_{red} < 1$  is regarded as the best fit, whereas a value of  $\chi^2_{red} > 1$  is regarded as a poor fit. In our observations, we have used  $N = 1054$  data points for BASE (SNIa and BAO),  $N = 1086$  data points for BASE+CC and  $N = 1104$  data points for BASE+CC+ $f\sigma_8$ . The number of free parameters of viscous  $\Lambda(t)$  is  $d = 4$  where as for  $\Lambda$ CDM it is  $d = 2$ . Using these information, the  $\chi^2_{red}$  for both the models are given in Table III. It can be observed that the value of  $\chi^2_{red}$  is less than unity with every data sets for both the models which show that the both models are in a very good fit with these observational data sets and the observed data are consistent with the considered models.

Using the three combination of data sets, we are also interested in investigating the cosmographical aspects of the models, such as jerk parameter, which is defined as

$$j = \frac{\ddot{a}(t)}{aH^3} = q(2q + 1) + (1 + z)\frac{dq}{dz}. \quad (56)$$

The jerk parameter which is a dimensionless third derivative of the scale factor, can provide us the simplest approach to search for departures from the  $\Lambda$ CDM model. It is noted that for  $\Lambda$ CDM model,  $j = 1$ (const.) always. Thus, any deviation from  $j = 1$  would favor a non- $\Lambda$ CDM model. In contrast to deceleration parameter which has negative values indicating accelerating Universe, the positive values of the jerk parameter show an accelerating rate of expansion. In Fig. 13, the evolutions of jerk parameter are shown for  $\Lambda$ CDM and viscous  $\Lambda(t)$  models using the best-fit values of parameters obtained from three combination of datasets. It is obvious from



the figure that this parameter remains positive and less than unity in past, and eventually tends to unity in late-time. Thus, the jerk parameter deviates in early time but it attains the same value as  $\Lambda$ CDM in late-time.

## VII. SELECTION CRITERION

There are two widely used selection criterion, namely, Akaike information criteria (AIC) and Bayesian information criteria (BIC) to measure the goodness of the fitted models compared to a base model. AIC is an essentially selection criteria based on the information theory where as the BIC is based on the bayesian evidence valid for large sample size. In cosmology, AIC and BIC are used to discriminate cosmological models based on the penalization associated with the number of free parameters of the considered models. The AIC parameter is defined through the relation [122]

$$AIC = \chi_{min}^2 + \frac{2dN}{N - d - 1}, \quad (57)$$

where  $d$  is the free parameters in a model,  $N$  the observational data points and  $\chi_{min}^2$  is the minimum value of the  $\chi^2$  function. AIC penalizes according to the number of free parameters of that model. To discriminate the proposed model  $m_1$  with the reference model  $m_2$ , we calculate  $\Delta AIC_{m_1 m_2} = AIC_{m_1} - AIC_{m_2}$ , which can be explained as “evidence in favor” of model  $m_1$  as compared to model  $m_2$ . In this paper, we consider  $\Lambda$ CDM model as a reference model ( $m_2$ ).

The value  $0 \leq \Delta AIC_{m_1 m_2} < 2$  refers to “strong evidence in favor” of the model  $m_1$ , for  $2 \leq \Delta AIC_{m_1 m_2} \leq 4$ , there is “average strong evidence in favor” of the model  $m_1$ , for  $4 < \Delta AIC_{m_1 m_2} \leq 7$ , there is “little evidence in favor” of the model  $m_1$ , and for  $\Delta AIC_{m_1 m_2} > 8$  there is “no evidence in favor” of the model  $m_1$ .

On the other hand, the Bayesian information criteria (BIC) can be defined as [123]

$$BIC = \chi_{min}^2 + d \ln N. \quad (58)$$

Similar to  $\Delta AIC$ ,  $\Delta BIC_{m_1 m_2} = BIC_{m_1} - BIC_{m_2}$  gives as “evidence against” the model  $m_1$  with reference to model  $m_2$ . For  $0 \leq \Delta BIC_{m_1 m_2} < 2$  gives “not enough evidence” of the model  $m_1$ , for  $2 \leq \Delta BIC_{m_1 m_2} < 6$ , we have “evidence against” the model  $m_1$ , and for  $6 \leq \Delta BIC_{m_1 m_2} < 10$ , there is “strong evidence against” the model  $m_1$ . Finally, if  $\Delta BIC > 10$  then there is strong evidence against the model and it is probably not the best model.

The values of  $\Delta AIC$  and  $\Delta BIC$  with respect to  $\Lambda$ CDM as the referring model are shown in Table III. According to our results,  $\Delta AIC(\Delta BIC) = 1.026(10.977)$  with respect to the *BASE* dataset,  $\Delta AIC(\Delta BIC) = 0.959(10.913)$  with *+CC* dataset, and for *+f $\sigma_8$*  dataset, we have  $\Delta AIC(\Delta BIC) = -7.492(2.416)$ . Thus, under AIC there is “strong evidence in favor” of the viscous

$\Lambda(t)$  model where as under BIC, there is “strong evidence against” the viscous  $\Lambda(t)$  model with *BASE* and *+CC* dataset and “positive evidence against” the model with *+f $\sigma_8$*  dataset.

## VIII. CONCLUSION

In this work, we have studied the analytical and observational consequences of cosmology inspired by dissipative phenomena in fluids according to Eckart theory with varying VED scenarios for spatially flat homogeneous and isotropic FLRW geometry. We have assumed the interaction of two components: viscous dark matter and vacuum energy density satisfying the conservation equation (8). To solve the field equations (9), we have considered various functional forms of bulk viscous coefficient, in particular (1)  $\zeta = \zeta_0$ ; (2)  $\zeta = \zeta_1 H$ ; (3)  $\zeta = \zeta_0 + \zeta_1 H$ ; and  $\zeta = \zeta_0 + \zeta_1 H + \zeta_2 (\ddot{a}/aH)$ . These viscous models have different theoretical motivations, but not all of them are able to constraint observationally. We have constrained only the viscous model  $\zeta = \zeta_1 H$  with varying VED. The motivation of the present work is to study the dynamics and evolutions of a wide class of viscous models with time varying vacuum energy density in the light of the most recent observational data. Current observations do not rule out the possibility of varying DE. It has been observed that the dynamical  $\Lambda$  could be useful to solve the coincidence problem. Although the functional form of  $\Lambda(t)$  is still unknown, a quantum field theory (QFT) approach has been proposed within the context of the renormalization group (RG). Thus, we have used the varying VED of the functional form  $\Lambda = c_0 + 3\nu H^2$  in all of viscous models presented in this paper. The motivation for this functional form stems from the general covariance of the effective action in QFT in curved geometry. It has been shown that the  $\Lambda(t)$  provides either a particle production processes or increasing the mass of the viscous dark matter particles. In what follows, we summarize the main results of the four different viscous  $\Lambda(t)$  models.

In case of the viscous  $\Lambda(t)$  models with  $\zeta = \zeta_0$ ,  $\zeta = \zeta_0 + \zeta_1 H$  and  $\zeta = \zeta_0 + \zeta_1 H + \zeta_2 (\ddot{a}/aH)$ , we have found the analytical solution of the cosmological parameters, like  $H(t)$ ,  $a(t)$ ,  $q(t)$  and  $w_{eff}(t)$ . It has been observed that these viscous  $\Lambda(t)$  models expand exponentially with cosmic time  $t$ . The models show the transition from decelerated phase to accelerated phase in late time. It is important to note that it is  $H(z)$  that is actually the observable quantity in cosmology which can be examined with current observations. However, assuming suitable choice of model parameters, the evolutions and dynamics of these models can be interpreted.

In case of viscous  $\Lambda(t)$  model with  $\zeta = \zeta_1 H$ , we have obtained the various cosmological parameters. We have performed a joint likelihood analysis in order to put the constrain on the main parameters by using the three different combinations of observational data: *BASE*, *+CC*

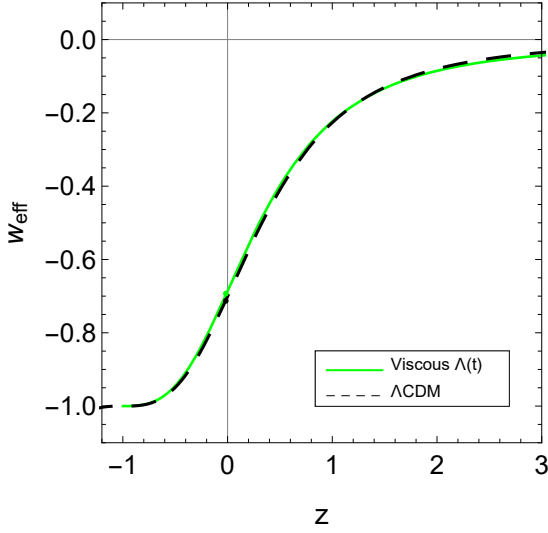


FIG. 10. Effective EoS parameter as a function of redshift  $z$  for viscous  $\Lambda(t)$  using “*BASE*” dataset. The evolution of EoS parameter in the standard  $\Lambda$ CDM model is also represented as the dashed curve. A dot denotes the present value of the EoS parameter.

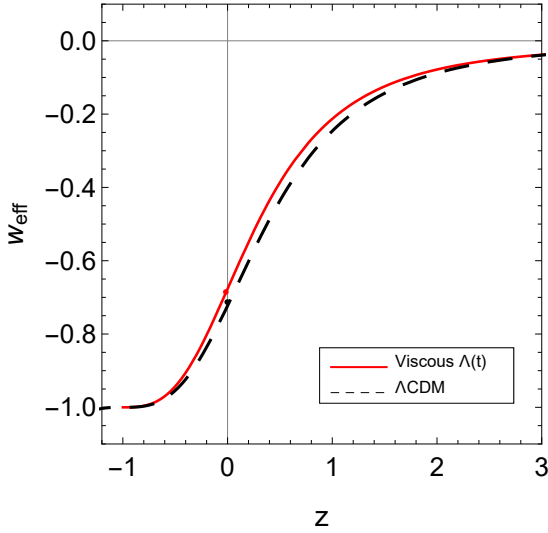


FIG. 12. Effective EoS parameter as a function of redshift  $z$  for viscous  $\Lambda(t)$  using “ $+f\sigma_8$ ” dataset. The evolution of EoS parameter in the standard  $\Lambda$ CDM model is also represented as the dashed curve. A dot denotes the present value of the EoS parameter.

and  $+f\sigma_8$ . To discriminate our model with the concordance  $\Lambda$ CDM model, we have also performed the statistical analysis for  $\Lambda$ CDM by using the same observational datasets. Our finding shows that this viscous  $\Lambda(t)$  model can accommodate a late time accelerated expansion. It has been observed that we can improve significantly the performance of the model by using *BASE* + *CC* +  $f\sigma_8$ . From observational consistency points of view, we have

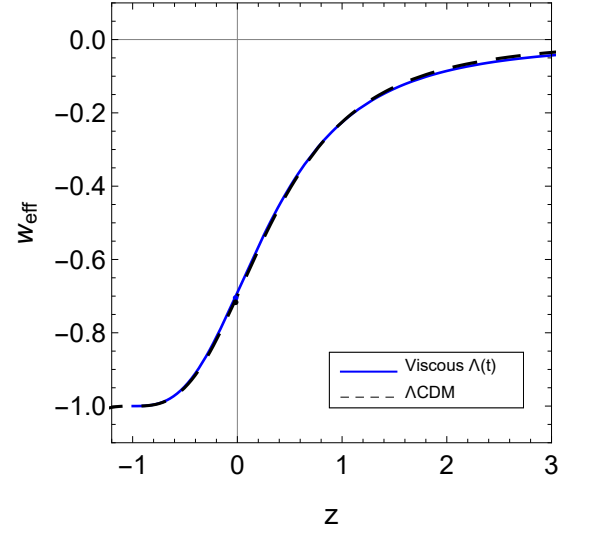


FIG. 11. Effective EoS parameter as a function of redshift  $z$  for viscous  $\Lambda(t)$  using “ $+CC$ ” dataset. The evolution of EoS parameter in the standard  $\Lambda$ CDM model is also represented as the dashed curve. A dot denotes the present value of the EoS parameter.

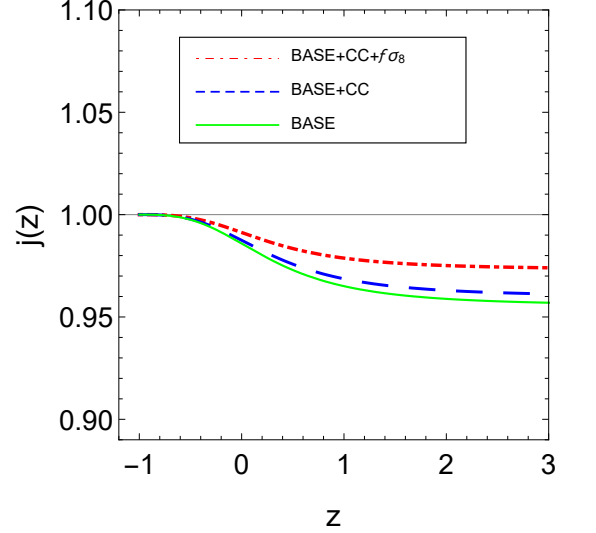


FIG. 13. Jerk parameter  $j(z)$  with redshift  $z$  using best-fit values of parameters for viscous  $\Lambda(t)$  model. The horizontal line represents the  $\Lambda$ CDM model.

examined the evolution of the viscous  $\Lambda(t)$  model on Hubble parameter, deceleration parameter and equation of state parameter by using the best-fit values of parameters. It has been observed that the model depicts transition from an early decelerated phase to late-time accelerated phase and the transition takes place at  $z_{tr} = 0.664^{+0.031}_{-0.042}$  with *BASE* data,  $z_{tr} = 0.665^{+0.031}_{-0.037}$

TABLE III. Values of Chi-squared, reduced Chi-squared, AIC and BIC of  $\Lambda$ CDM and viscous  $\Lambda(t)$  models. The  $\Lambda$ CDM model is considered as reference model to calculate the  $\Delta$ AIC and  $\Delta$ BIC.

Values	BASE		+CC		+ $f\sigma_8$	
	$\Lambda$ CDM	viscous $\Lambda(t)$	$\Lambda$ CDM	viscous $\Lambda(t)$	$\Lambda$ CDM	viscous $\Lambda(t)$
$\chi^2$	518.017	515.074	525.457	522.390	842.630	831.112
$d$	2	4	2	4	2	4
$N$	1054	1054	1086	1086	1104	1104
$\chi^2_{red}$	0.492	0.498	0.484	0.481	0.764	0.755
AIC	522.028	523.055	529.468	530.427	846.641	839.112
BIC	531.938	542.915	539.438	550.351	856.643	859.139
$\Delta$ AIC	—	1.026	—	0.959	—	-7.492
$\Delta$ BIC	—	10.977	—	10.913	—	2.496

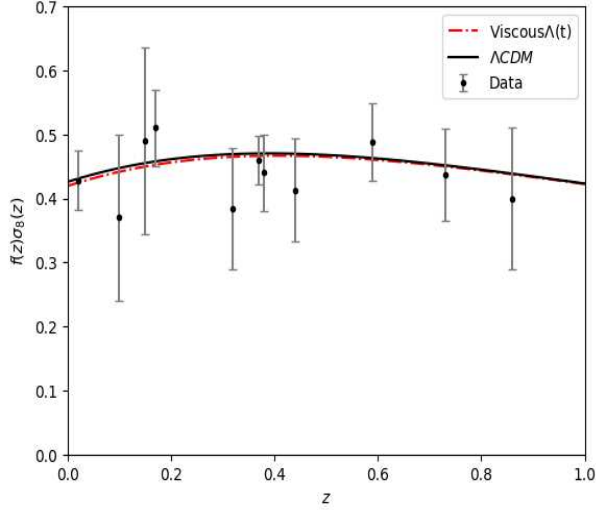


FIG. 14. Theoretical curves for the  $f(z)\sigma_8(z)$  corresponding to  $\Lambda$ CDM and viscous  $\Lambda(t)$  model along with some of the data points employed in our analysis. To generate this plot we have used the best-fit values of the cosmological parameters listed in Tables I and II for “+  $f\sigma_8$ ” data.

with + $CC$  data and  $z_{tr} = 0.626^{+0.028}_{-0.037}$  with + $f\sigma_8$  data. The present viscous  $\Lambda(t)$  model has  $q_0 = -0.533^{+0.025}_{-0.020}$ ,  $q_0 = -0.535^{+0.023}_{-0.020}$  and  $q_0 = -0.516^{+0.022}_{-0.017}$  respectively. Thus, both  $z_{tr}$  and  $q_0$  values are in good agreement with that of  $\Lambda$ CDM model. The ages of the Universe obtained for this model with each dataset are very much compatible with the  $\Lambda$ CDM model. The proposed model has small negative value of EoS parameter for large redshifts and asymptotically approaches to cosmological constant for small redshifts. Thus, the viscous  $\Lambda(t)$  model behaves like quintessence in early time and cosmological constant in late-time. The residual tensions of our fitting results with respect to the latest local measurement  $H_0 = 73.04 \pm 1.04$  km/s/Mpc [121] are  $3.92\sigma$ ,  $3.85\sigma$  and  $4.07\sigma$ , respectively. In Ref. [124], the authors found  $H_0 = 69.13 \pm 2.34$  km/s/Mpc assuming the  $\Lambda$ CDM. Such result almost coincides with  $H_0$  that we obtained in Tables I and II for  $\Lambda$ CDM and viscous  $\Lambda(t)$  models. We have

explored the  $\sigma_8$  and  $S_8$  parameters using the combined datasets of  $BASE+CC+f\sigma_8$ . The constraints on  $\sigma_8$  and  $S_8$  from this combined analysis are  $\sigma_8 = 0.790^{+0.008}_{-0.010}$  and  $S_8 = 0.822^{+0.019}_{-0.019}$ , respectively which are very close to the values of  $\Lambda$ CDM. The tension of our fitting results in  $\sigma_8$  and  $S_8$  for viscous  $\Lambda(t)$  model with respect to respective  $\sigma_8$  and  $S_8$  of  $\Lambda$ CDM are  $0.23\sigma$  and  $-0.38\sigma$ , respectively. The evolution of  $f\sigma_8$  as displayed in Fig.14 shows that the behaviour of  $f\sigma_8$  is consistent with the observational data points. It has been noticed that the best-fit results are consistent in the vicinity of Planck data [7].

It has been observed that the value of  $\chi^2_{red}$  is less than unity with every data sets which show that the model is in a very good fit with these observational data sets and the observed data are consistent with the considered model. The jerk parameter remains positive and less than unity in past, and eventually tends to unity in late-time. Thus, the jerk parameter deviates in early time but it attains the same value as  $\Lambda$ CDM in late-time.

To discriminate the viscous  $\Lambda(t)$  with the  $\Lambda$ CDM, we have examined the selection criterion, namely, AIC and BIC. According to the selection criteria  $\Delta$ AIC, we have found that the viscous  $\Lambda(t)$  model is “positively favored” over the  $\Lambda$ CDM model for  $BASE$ , + $CC$  and + $f\sigma_8$  datasets. Similarly, with respect to  $\Delta$ BIC our model has a “very strong evidence against” the model for  $BASE$  and + $CC$  datasets whereas when we add + $f\sigma_8$  dataset, there is “no significant evidence against” the model. As a concluding remark we must point out that the viscous models with decaying VED may be preferred as potential models to examine the dark energy models beyond the concordance cosmological constant. The viscous effects with decaying VED can drive an accelerated expansion of the Universe. Thus, a viable cosmology can be constructed with viscous fluids and decaying VED. With new and more accurate observations, and with more detailed analyses, it would be possible to conclusively answer the compatibility of viscous model with dynamical vacuum energy.

## ACKNOWLEDGMENTS

One of the author, VK would like to thank Delhi Technological University, India for providing Research Fellowship to carry out this work.

- 
- [1] A. G. Riess et al., *Astron. J.* **116**, 1009 (1998)
  - [2] S. Perlmutter et al., *Astrophys. J.* **517**, 565 (1999)
  - [3] C.L. Bennet et al., *Astrophys. J. Suppl.* **148**, 1 (2003)
  - [4] M. Tegmark et al., *Phys. Rev. D* **69**, 103501 (2004)
  - [5] S. Alam et al., *Mon. Not. R. Astron. Soc.* **470**, 2617 (2017)
  - [6] M.H. Amante et al., *Mon. Not. R. Astron. Soc.* **498**, 6013 (2020), arXiv:1906.04107
  - [7] N. Aghanim et al. [Planck Collaboration], *Astrophys. Astron.* **641**, A6 (2020), arXiv:1807.06209 [astro-ph.CO].
  - [8] R.R. Caldwell, R. Dave and P.J. Steinhardt, *Astrophys. Space Sci.* **261**, 303 (1998)
  - [9] L.-M. Wang, R. R. Cardwell, J.P. Ostriker and P.J. Steinhardt, *Astrophys. J.* **530**, 17 (2000)
  - [10] P.J. Steinhardt, *Phil. Trans. Roy. Soc. Lond. A* **361**, 2497 (2003)
  - [11] P.J.E. Peebles and B. Ratra, *Rev. Mod. Phys.* **75**, 559 (2003)
  - [12] S. Weinberg, *Rev. Mod. Phys.* **61**, 1 (1989)
  - [13] S.M. Carroll, *Living. Rev. Rel.* **4**, 1 (2001)
  - [14] T. Padmanabhan, *Phys. Rept.* **380**, 235 (2003)
  - [15] I. Zlatev, L.-M. Wang and P.J. Steinhardt, *Phys. Rev. Lett.* **82**, 896 (1999)
  - [16] K. Freese, F.C. Adams, J.A. Frieman and E. Mottola, *Nucl. Phys. A* **287**, 797 (1987)
  - [17] J.C. Carvalho, J.A. Lima and I. Waga, *Phys. Rev. D* **46**, 2404 (1992)
  - [18] J. A. Lima and J.M. Maia, *Phys. Rev. D* **49**, 5597 (1994)
  - [19] J.A. Lima, *Phys. Rev. D* **54**, 2571 (1996)
  - [20] P. Wang and X. Meng, *Class. Quantum. Grav.* **22**, 283 (2005)
  - [21] E. Elizalde, S. Nojiri, S.D. Odintsov and P. Wang, *Phys. Rev. D* **71**, 103504 (2005)
  - [22] J.S. Alcaniz and J.A.S. Lima, *Phys. Rev. D* **72**, 063516 (2005)
  - [23] H.A. Borges and S. Carneiro, *Gen. Relativ. Grav.* **37**, 1385 (2005)
  - [24] J. Solà and H. Stefancic, *Mod. Phys. Lett. A* **21**, 479 (2006)
  - [25] S. Carneiro, C. Pigozzo, H.A. Borges and J.S. Alcaniz, *Phys. Rev. D* **74**, 23532 (2006)
  - [26] H.A. Borges, S. Carneiro, J.C. Fabris and C. Pigozzo, *Phys. Rev. D* **77**, 043513 (2008)
  - [27] S. Carneiro, M.A. Dantas, C. Pigozzo and J.S. Alcaniz, *Phys. Rev. D* **77**, 083504 (2008)
  - [28] S. Basilakos, *Mon. Not. R. Astron. Soc.* **395**, 2374 (2009)
  - [29] S. Basilakos, M. Plionis and J. Solà, *Phys. Rev. D* **80**, 083511 (2009)
  - [30] F.E.M. Costa and J.S. Alcaniz, *Phys. Rev. D* **81**, 043506 (2010)
  - [31] C. Pigozzo, M.A. Dantas, S. Carneiro and J.S. Alcaniz, *JCAP*, **08**, 022 (2011)
  - [32] J. Solà, *J. Phys. Conf. Ser.* **283**, 012033 (2011)
  - [33] J. Grande, J. Solà, S. Basilakos and M. Plionis, *J. Cosmol. Astropart. Phys.* **11** 007 (2011)
  - [34] D. Bessada and O.D. Miranda, *Phys. Rev. D* **88**, 083530 (2013)
  - [35] J. Solà, *J. Phys. Conf. Ser.* **453**, 012015 (2013)
  - [36] E.L.D. Perico, J.A.S. Lima, S. Basilakos and J. Solà, *Phys. Rev. D* **88**, 063531 (2013)
  - [37] M. Szydlowski and A. Stachowski, *J. Cosmol. Astropart. Phys.* **066**, 1510 (2015)
  - [38] A. Gomez-Valent and J. Solà, *Mon. Not. R. Astron. Soc.* **448**, 2810 (2015)
  - [39] J. Solà and A. Gomez-Valent, *Int. J. Mod. Phys. D* **24**, 1541003 (2015)
  - [40] A.P. Jayadevan, M. Mukesh, A. Shaima and T.K. Mathew, *Astrophys. Space Sci.* **364**, 67 (2019)
  - [41] I. Shapiro and J. Solà, *JHEP* **202**, 006 (2002)
  - [42] J. Solà, *J. Phys. A* **41**, 164066 (2008)
  - [43] J. Grande, J. Solà and H. Stefancic, *J. Cosmol. Astropart. Phys.* **8**, 11 (2006)
  - [44] C.P. Singh and J. Solà, *Eur. Phys. J. C* **81**, 960 (2021)
  - [45] M. Rezaei, J. Solà and M. Malekjani, *Mon. Not. R. Astron. Soc.* **509**, 2593 (2022)
  - [46] W. Zimdahl, D.J. Schwarz, A.B. Balakin and D. Pavón, *Phys. Rev. D* **64**, 063501 (2001)
  - [47] A.B. Balakin, D. Pavón, D.J. Schwarz and W. Zimdahl, *New J. Phys.* **5**, 85 (2003)
  - [48] C. Eckart, *Phys. Rev. D* **58**, 919 (1940)
  - [49] L.D. Landau, E.M. Lifshitz, *Fluid Mechanics*, Vol. 6, Butterworth Heinemann Ltd. Oxford (1987)
  - [50] W.A. Hiscock and L. Lindblom, *Phys. Rev. D* **31**, 725 (1985)
  - [51] I. Müller, *Z. Phys.* **198**, 329 (1967)
  - [52] W. Israel and J.M. Stewart, *Phys. Lett. A* **58**, 213 (1976)
  - [53] Ø. Grøn, *Astrophys. Space Sci.* **173**, 191 (1990)
  - [54] R. Maartens, *Class. Quantum Grav.* **12**, 1455 (1995)
  - [55] A.A. Coley, R.J. Van den Hoogen and R. Maartens, *Phys. Rev. D* **54**, 1393 (1996)
  - [56] W. Zimdahl, *Phys. Rev. D* **53**, 5483 (1996)
  - [57] I. Brevik and S.D. Odintsov, *Phys. Rev. D* **65**, 067302 (2002)
  - [58] I. Brevik and A. Hallanger, *Phys. Rev. D* **69**, 024009 (2004)
  - [59] C.P. Singh, S. Kumar and A. Pradhan, *Class. Quantum Grav.* **24**, 455 (2007)
  - [60] C.P. Singh, *Pramana J. Phys.* **71**, 33 (2008)
  - [61] S. Nojiri and S.D. Odintsov, *Phys. Lett. B* **649**, 440 (2007)
  - [62] I. Brevik, O. Gorbunova, and D. Saez-Gomez, *Gen. Relativ. Grav.* **42**, 1513 (2010)
  - [63] H. Velten, J. Wang and X. Meng, *Phys. Rev. D* **88**, 123504 (2013)

- [64] J.X. Wang and X.H. Meng, *Mod. Phys. Lett. A* **29**, 1450009 (2014)
- [65] K. Bamba and S.D. Odintsov, *Eur. Phys. J. C* **76**, 18 (2016)
- [66] I. Brevik, Ø. Grøn, J. de Haro, S.D. Odintsov and E.N. Saridakis, *Int. J. Mod. Phys. D* **26**, 1730024 (2017)
- [67] G.M. Kremer and F.P. Devecchi, *Phys. Rev.D* **67**, 047301 (2003)
- [68] J.C. Fabris, S.V.B. Goncalves and R. de Sá Ribeiro, *Gen. Relativ. Grav.* **38**, 495 (2006)
- [69] I. Brevik and O. Gorbunova, *Gen. Relativ. Grav.* **37**, 2039 (2005)
- [70] M.-G. Hu and X.-H. Meng, *Phys. Lett. B* **635**, 186 (2006)
- [71] J. Ren and X.-H. Meng, *Phys. Lett. B* **633**, 1 (2006)
- [72] X.-H. Meng, J. Ren and M.-G. Hu, *Commun. Theor. Phys.* **47**, 379 (2007)
- [73] J.R. Wilson, G.J. Mathews and G.M. Muller, *Phys. Rev.D* **75**, 043521 (2007)
- [74] G.J. Mathews, N.Q. Lan and C. Kolda, *Phys. Rev.D* **78**, 043525 (2008)
- [75] A. Avelino and U. Nucamendi, *AIP Conf. Proc.* **1083**, 1 (2008)
- [76] A. Avelino, U. Nucamendi and F.S. Guzman, *AIP Conf. Proc.* **1026**, 300 (2008)
- [77] A. Avelino and U. Nucamendi, *J. Cosmol. Astropart. Phys.* **04**, 006 (2009)
- [78] X.H. Meng and X. Dou, *Commun. Theor. Phys.* **52**, 377 (2009)
- [79] A. Avelino and U. Nucamendi, *J. Cosmol. Astropart. Phys.* **08**, 009 (2010)
- [80] C.P. Singh and P. Kumar, *Eur. Phys. J. C* **74**, 3070 (2014)
- [81] A. Sasidharan and T.K. Mathew, *Eur. Phys. J. C* **75**, 348 (2015)
- [82] B.D. Normann and I. Brevik, *Mod. Phys. Lett. A* **32**, 1750026 (2017)
- [83] D. Wang, Y.-J. Yan and X.-H. Meng, *Eur. Phys. J. C* **77**, 660 (2017)
- [84] C.P. Singh and A. Kumar, *Eur. Phys. J. Plus* **133**, 312 (2018)
- [85] C.P. Singh and A. Kumar, *Mod. Phys. Lett. A* **33**, 1850225 (2018)
- [86] C.P. Singh and M. Srivastava, *Eur. Phys. J. C* **78**, 190 (2018)
- [87] C.P. Singh and A. Kumar, *Astrophys. Space Sci.* **364**, 94 (2019)
- [88] C.P. Singh and S. Kaur, *Astrophys. Space Sci.* **365**, 2 (2020)
- [89] C.P. Singh, *Nouvo Cimento B* **122**, 89 (2007)
- [90] C.P. Singh and S. Kumar, *Astrophys. Space Sci.* **323**, 407 (2009)
- [91] N. Mostafapoor and Ø. Grøn, *Astrophys. Space Sci.* **333**, 357 (2011)
- [92] J. Hu and H. Hu, *Eur. Phys. J. Plus* **135**, 718 (2020)
- [93] L. Herrera-Zamorano, A. Hernández-Almada and Miguel A. García-Aspeitia, *Eur. Phys. J. C* **80**, 637 (2020)
- [94] S. Weinberg, *Gravitation and Cosmology: Principles and applications of the general theory of relativity*, John Wiley and Sons, Inc. New York U.S.A. (1972)
- [95] S.L. Adler, *Rev. Mod. Phys.* **54**, 729 (1982)
- [96] L. Parker and D.J. Toms, *Phys. Rev. D* **32**, 1409 (1985)
- [97] I.L. Shapiro and J. Sola, *Phys. Lett. B* **682**, 105 (2009)
- [98] F.E.M. Costa, J.A.S. Lima and F.A. Oliveira, *Class. Quantum Grav.* **31**, 045004 (2012)
- [99] A. Avelino, R.G. Salcedo, T. Gonzalez, U. Nucamendi and I. Quiros, *J. Cosm. Astropart. Phys.* **08** 12 (2013)
- [100] A. Sasidharan and T.K. Mathew, *J. High Energy Phys.* **06**, 138 (2016)
- [101] C.P. Singh and Ajay Kumar, *Grav. Cosmol.* **25**, 58 (2019)
- [102] T. Padmanabhan and S. M. Chitre, *Phys. Lett. A* **120** 433 (1987)
- [103] A. Montiel and N. Breton, *J. Cosmol. Astropart. Phys.* **08** 023 (2011)
- [104] I. Brevik, *Phys. Rev.D* **65**, 127302 (2002)
- [105] J. Hu and H. Hu, *Eur. Phys. J. Plus* **135**, 718 (2020)
- [106] J. Sola, et al., *Mon. Not. Roy. Astron. Soc.* **478**, 4357 (2018)
- [107] A. Gomez-Valent and J. Sola Peracaula, *Mon. Not. Roy. Astron. Soc.* **478**, 126 (2018), arXiv:1801.08501
- [108] J. de Cruz, J. Sola and C.P. Singh, arXiv: 2302.04807, (2023)
- [109] P.J.E. Peebles, *Principles of Physical Cosmology*, Princeton University Press (1993)
- [110] Y.-S. Song, W.J. Percival, *J. Cosmol. Astropart. Phys.* **2009** 004 (2009)
- [111] D. Huterer, D. Kirkby, et al. *Astropart. Phys.* **63**, 23 (2015)
- [112] S. Nesseris and L. Perivolaropoulos, *Phys. Rev. D* **77**, 023504 (2008)
- [113] D.M. Scolnic et al., *Astrophys. J.* **859**, 101 (2018)
- [114] C. Blake et al., *Month. Not. R. Astron. Soc.* **418**, 1707 (2011)
- [115] W. J. Percival et al., *Month. Not. R. Astron. Soc.* **401**, 2148.
- [116] R. Giotri et al., *J. Cosmol. Astropart. Phys.* **1203**, 027 (2012)
- [117] Eisenstein D, *Astrophys. J.* **633**, 560 (2005), [astro-ph/0501171]
- [118] M. Moresco et al., *Living Rev. Relativ.* **25**, 6 (2022)
- [119] S. Nesseris, G. Pantazis and L. Perivolaropoulos, *Phys. Red. D* **96**, 023542 (2017), arXiv: 1703.10538[astro-ph.CO]
- [120] Daniel Foreman-Mackey, et al., *PASP* **125**, 306 (2013)
- [121] A. G. Riess et al., *Astrophys. J. Lett.*, **934**, 1, L7, (2022), arXiv:2112.04510
- [122] H. Akaike, *IEEE Trans. Autom. Control* **19**, 716 (1974)
- [123] G. Schwarz, *Ann. Stat.* **6**, 461 (1978)
- [124] Y. Wang, Lixin, Xu and G.B. Zhao, *Astrophys. J.* **849**, 84 (2017)

RESEARCH ARTICLE | SEPTEMBER 05 2023

## Vision transformer based Devanagari character recognition



Shailendra Kumar; Abhinav Chopra ✉; Sambhav Jain; Sarthak Arora



AIP Conf. Proc. 2754, 160001 (2023)

<https://doi.org/10.1063/5.0169520>



CrossMark

### Articles You May Be Interested In

Analysis of stop consonants in Devanagari alphabet

*Proc. Mtgs. Acoust* (June 2013)

Analysis of stop consonants in Devanagari alphabet

*J Acoust Soc Am* (May 2013)

A preliminary ultrasound study of Nepali lingual articulations

*Proc. Mtgs. Acoust* (June 2013)

500 kHz or 8.5 GHz?  
And all the ranges in between.

Lock-in Amplifiers for your periodic signal measurements



Find out more





# Vision Transformer Based Devanagari Character Recognition

Shailendra Kumar<sup>1,a)</sup>, Abhinav Chopra<sup>1,b)</sup>, Sambhav Jain<sup>1,c)</sup>, Sarthak Arora<sup>1,d)</sup>

<sup>1</sup>Delhi Technological University, Delhi, India

<sup>a)</sup>shailenderkumar@dce.ac.in

<sup>b)</sup>Corresponding author: abhinavchopra\_2k18co011@dtu.ac.in

<sup>c)</sup>sambhavjain\_2k18co313@dtu.ac.in

<sup>d)</sup>sarthakarora\_2k18co325@dtu.ac.in

**Abstract.** Devanagari is an ancient script that is used to write Hindi, Nepali, Marathi, Maithili, Awadhi, Newari, and Bhojpuri, among other Indo-Aryan languages. Thousands of individuals in India use this script to write documents in Marathi and Hindi. Indian mythology is based on this script. Because of the script's prominence, handwritten Devanagari character identification has grown in popularity over time. Handwritten recognition of languages such as English has received a lot of attention, but Indian languages written in the Devanagari script are also a rich source of information. Most of the work on this problem statement has been done either using deep neural networks like CNN at its heart coupled with other machine learning techniques like SVM, Random Forest etc. In this paper we are utilising a recently introduced transformer model for computer vision known as Vision Transformer for the task of Devanagari Character Recognition. We have also compared our model with various pretrained CNN-based architectures like ResNet50, VGG16 and InceptionV3 and ViT has outperformed these models both on DHCD dataset and the modified slightly more complex version of it with accuracy scores of 99.68% on the original testing dataset of the DHCD dataset and accuracy score of 96.55% on the modified (blurred) slightly more complex version of the original testing dataset. The ViT model thus generalized better than standard CNN-based models on the problem of Devanagari Character recognition.

**Keywords:** ViT (Vision Transformer), CNN, ResNet50, VGG16, InceptionV3.

## INTRODUCTION

Humans have long been curious about whether computers can be programmed to understand and observe things in the same way that humans do with the help of their sense organs. This gave rise to advancement of technology in traditional machine learning techniques as well as development of Deep Learning techniques. Also with the advancement of hardware and introduction of GPUs led to more processing power which in turn resulted in reduction of training times of these models. Character recognition is an important task, if it can be done with the help of a computer, it can ease human efforts, in reading large texts, written not only in English but also in their native languages (Devanagari). Character Recognition is a challenging task, it means that computer has to identify the character through its image, which is not at all fixed, and can vary depending on the person who wrote that character, because every person has a different handwriting. When the researchers were thinking to perform this mammoth task, they were amazed that how human beings are able to do it so easily. People, on the other hand, have a built-in neural net, so whatever image they perceive passes through numerous layers, and this seemingly tough process is completed with ease. A Devanagari character recognition system is a very important tool for recognizing the records which are kept digitally (for e.g. the Indian Government keeps the records written in Devanagari script in a digital format). Now suppose a historian wants to extract information from this huge record, but it won't be feasible as that search tool can't work on images, so in order to do that we will need character recognition for our images.

Most of the work done on character recognition for images have used Convolutional Neural Networks as the go to choice [1]. These CNN-based models have been state of the art for a majority of these computer vision based problems.

Recently, though, [2] introduced the model called Vision Transformers to the world. Transformers were well known for their use in the field of Natural Language Processing such as sentiment classification, machine translation

etc. [3] They were able to solve the bottleneck problem and with hardware like GPUs being more readily available, were computationally efficient than the standard sequence models like LSTMs, RNNs because of the parallel processing [3]. Further, they captured the long term dependencies in a better way because in general every token interacted with every other token when passed through the “Self Attention” layer [3]. In order to give a sense of position to every token, a positional embedding is also added before computing self attention. ViT or Vision Transformers [2] extends this to images by making patches of images and linearly transforming them to tokens and adding positional representation to each token. All these tokens are then feeded into the standard transformer encoder network and the 0th token is passed through the head for classification task.

These large models like ViTs thrive on data. Training them on more and more data improves their performance by many folds. So, using these models directly on our downstream task of character recognition wouldn't be as fruitful due to limited availability of labelled data. In the near past, to solve this problem, a different approach, of pretraining these models on large volumes of data on some task which has these large datasets available and is somewhat related to the downstream task for which we want the model and then fine-tuning these pretrained models on the downstream task [4], has achieved very good results and has somewhat addressed this problem of small datasets.

In our work, we use a ViT model (provided by the Timm Library) [5], pretrained on big datasets like image net and then fine-tune it on the DHCD dataset. We then compare its performance with some standard CNN-based models like ResNet50, VGG16 and InceptionV3 (pre trained versions of these models available in PyTorch [6] were also fine tuned on the DHCD dataset [7]). The ViT achieved an accuracy score of 99.68% and outperformed all these models. Further we analysed the performance of these models on a modified, slightly more complex version of the DHCD test dataset and ViT again outperformed these CNN-based models with greater margins.

The further paper is divided as follows: Section II reviews some of the notable works done in this field. Section III explains the methodology we have used for developing our models with details about the dataset used, how the images are preprocessed for each model and what are the hyperparameters chosen for each model. Results are presented and analysed in section IV.

## BACKGROUND AND RELATED WORK

Since the foundation is necessary for conducting the research, the second section tries to give important details about the work already done on this topic. After developing a good background of the work already done, we can propose our solution to the problem. In many research studies, different ways of image processing have been proposed. Most of them used CNN for feature extraction but differed in their approaches for further classification.

CNN was utilised by Sonika Narang, Munish Kumar, and M.K. Jindal [8] for both feature extraction and classification, proposing a method achieving 93% accuracy. They produced and worked on a dataset from the 15th to 19th centuries containing 5484 images of 33 classes of Devanagari characters. The manuscript images were converted into binary data images, paragraphs split into characters, followed by standardisation of image sizes yet inconsistently sized characters. The dataset was then split into ratio of training to test set size as 3:1. In the suggested technique, 3 Convolution Layers, 3 Pooling Layers and a Fully Connected Layer were used in the method described. The 3 convolution layers contained 32, 64 and 128 filters respectively followed by which ReLu Activation function was applied, pooling technique used was Max-Pooling and the Fully Connected Layer used the Softmax Function for final classification. Besides achieving a good accuracy, the given methodology had some limitations like inability to classify similar looking characters like [pa] and [ya], [ha] and [da], [bha] and [ma] and many more, correctly at all times and misclassification of faded characters like [sha] and [ja], because of the antiquity of manuscripts.

Mamta Bisht, Richa Gupta [9] focussed majorly upon feature extraction so 2 CNN-based models, Histogram of Oriented Gradients (HOG) were used for feature extraction and a Support Vector Machine (SVM) - based Classifier for final text classification of the offline handwritten characters (both normal and modified). The datasets used were that of the Hindi consonants and the matras dataset. The two CNN models followed single and double architecture respectively. Both of them had undergone a preprocessing step for images conversion to black and white, followed by the Convolution part containing 7 layers with 8,16,32,64,128,256,256 filters respectively, activation functions and 7 down-sampling layers following each convolution layer. The single architecture model was trained on combination of consonants and matras dataset while other model was trained in two stages - firstly on consonants dataset then on the same combination as in the case of single architecture, after which descriptive features were extracted using HOG technique which were then sent to SVM classifier. Besides calculating the accuracy in the six-fold cross-validation method, the CNN model gave a better combined accuracy of 91% on the test set on random distribution of dataset for 11 experiments in each stage of double architecture.

Saptarshi Kattyayan, T. Kar, P. Kanungo [10] performed text-classification on the Devanagari Character Set. In the proposed method, CNN-based architecture was used for feature extraction and for final text classification, comparisons between 5 different types of Machine Learning Classifiers and CNN were made. Firstly, the images in

the dataset are made to undergo binarization, followed by conversion to black and white image using thresholding technique. The dataset was then split into a training and a test set in the ratio of 17:3. The CNN model used consists of two Convolution layers, ReLU activation function and two max-pooling layers following each convolution layer. For final text classification, five Machine Learning Algorithms - Gaussian Naive-Bayes, Decision Tree, Random Forest, K-Nearest Neighbour, Extra Tree Classifiers; and CNN-based Classifier were used. Following the completion of the classification operation, CNN achieved the best accuracy of 98% for kernel sizes of 3\*3 and 5\*5 in 1st and 2nd convolution layers respectively after conducting 4 experiments while among the Machine Learning Algorithms, KNN Classifier achieved the best accuracy of 72%.

Piyush Gupta, Saurabh Deshmukh, Siddhant Pandey, Kedar Tonge, Vrushabh Urkundesainath Kide [11] had proposed a CNN-based text classification method involving repetitively performing two convolution and max-pooling operations by changing kernel filters and checking for accuracy on different input images for effective feature extraction. The process of Convolution involved a filter matrix being slid over the image pixels matrix, and all the corresponding values in that matrix window were multiplied and added, and then assigning the resultant value to the centre of the matrix window. This was followed by max-pooling process and activation functions for training dataset optimization and loss reduction at each step followed by which the extracted features were sent to the Classification layer. The training of the CNN model was done on a dataset containing 10,000 images per consonant and 2000 images per vowel. The tech stack used to code the training step was Keras in TensorFlow. After the training step, a .h5 file was produced containing data about all the constraints and parameters used in the model. During testing, to record/analyze the input, tkinter and OpenCV applications were used. Tkinter helped in recognition of the input which was drawn on the screen on a python-based GUI created by it, whereas OpenCV helps in recognition of the input drawn in mid-air. Following the above procedure, the model was successful in pulling off an accuracy of 96.8% after 3 epochs.

Nagender Aneja and Sandhya Aneja [12] proposed an approach involving the use of pre-trained models for character classification by transfer learning. The dataset used for strategy implementation was Devanagari Character Set of grayscale images with training and test set in the ratio 17:3. Performance of different feature extractors - AlexNet, Vgg, DenseNet and Inception ConvNet on Devanagari Character Set, were noted individually. By retaining the original structure of the pre-trained models, the training of the fully connected layer was performed. After the whole process, AlexNet was the fastest as a feature extractor, pulling off an accuracy of 98%.

A.N. Holambe, R.C. Thool, and S. Jagade [13] came up with an approach involving feature extraction and recognition of visual images directly from dataset images with least amount of pre-processing. A self-created dataset of handwritten and online input characters of 42 classes of approximately 41k images being split into training and test sets of 33k images and 8k images respectively. Preprocessing part of the whole dataset involved processes like fragmentation of characters by removing the matra and shirorekha, normalisation on online input data followed by up-sampling and dilation. CNN with one input layer, two convolutional layers and two fully connected layers (hidden and output) was used to train the dataset and techniques like exponential delay and inverse scale annealing, helped in achieving a test set accuracy of 98.19% and 1.81% error rate on test set and 0.8% error rate on training set.

S. Acharya, A.K. Pant, and P.K. Gyawali [14] proposed an approach in which traditional dataset (DHCD) was trained using Deep CNN, along with Dropout and dataset increment approach to face off the challenge of similarity of character pairs in the dataset differing only by dots or size of shirorekha and for enhancement of test set accuracy. The whole approach involves cropping, scanning and labelling character images written by hand. This is followed by pre-processing of dataset and train-test split in the ratio 17:3 with fixed pixel count of images and their corresponding characters. For training of dataset, two convolutional layers (5 \* 5 kernels) followed by a subsampling layer each, were used whose features are then transferred to the fully connected layer for classification. To avoid overfitting of the large deep CNN, techniques like dataset increment, which involves creation of new images with re-positioned characters for better training, and dropout method, which involves dropping units at random resulting in a lighter network, thus helping in faster training. The above proposed method, after applying accuracy optimization methods, was successful in achieving test accuracy of 98.47

V.P. Agnihotri [15] proposed an approach that differed from other approaches in its feature extraction method. The initial stages of the deployed strategy included various steps like the acquisition of images, involving creation of scanned images to be considered as an input. This step was followed by the pre-processing step which involved images of dataset to undergo various operations like grayscale conversion, edge detection and dilation technique allowing it to be used in the segmentation procedure which involved breaking down of characters sequence into single characters and applying labelling on them in order to provide details about count of a particular character as well. This stage was followed by feature extraction stage using diagonal feature extraction which involved distribution of a 90\*60 pixels image into 54 zones of 10\*10 pixels each and a single averaged value was then

calculated from each of the zones along the zones' diagonals leading to capturing of 54 features for each image in the dataset. All the features extracted in the previous step were converted to chromosome bit strings and the classification process took place using the chromosome fitness method. Besides achieving a precision of 85.78% match, the model faced setbacks during recognition because of the similar shape of some characters, different handwriting, different fonts and different positions of the same character.

## METHODOLOGY USED

Our methodology consists of several steps. We first load the DHCD dataset [7]. The dataset is then split in the ratio 85:15 for training and testing respectively. For better performance comparison of our ViT model and the CNN-based models, we took a copy of the testing dataset and blurred images in it using Gaussian blurr. This blurred copy of the original testing dataset will be referred to as the modified (blurred) testing dataset further in the paper. So ultimately we have two sets of images to compare the performance of these models: the original testing dataset and the modified(blurred) testing dataset.

The images are then preprocessed to make them ready to be fed as input to the pretrained ViT model and the CNNbased models. The exact configurations according to which the images were preprocessed for each model are mentioned in the data preprocessing section. These pretrained models are then fine tuned on the DHCD dataset. After fine-tuning, to compare the performance of ViT to the CNN-based models, we calculate the accuracy of these models on the two sets of testing dataset we have i.e. the original testing dataset and the modified(blurred) testing dataset. This entire pipeline is shown in Figure 4. The hyperparameters used for finetuning each model have been specified in the training process section.

### Dataset Description

We have used a public Devanagari Handwritten character dataset (DHCD) which was developed by extracting and manually annotating thousands of characters from handwritten manuscripts [7]. The dataset size is 92000 grey-scale images with 46 classes (36 of characters and 10 for digits). Each class has 2000 images and each image is of size 32X32. The dataset is divided into two parts: a training set (85 percent) and a testing set (15 percent).

#### *Developing a modified(blurred) copy of the testing dataset*

In order to better analyse and compare how these models are performing relative to one another, we created a slightly modified version of the testing dataset by blurring images in the original testing dataset using Gaussian Blur. This was done so as to see how these models are performing on data examples that are slightly harder to classify. So we ultimately have 2 sets of images to evaluate our models: the original testing dataset and a slightly more complex version of it.

### Data Preprocessing

These pretrained models require images to be fed into them with a certain configuration. The ViT base model [2] we used required images to be of size (224X224) and having zero mean and standard deviation in each channel. Similarly, ResNet50 requires images of size (224X224) and mean and standard deviation to be (0.485, 0.456, 0.406) and (0.229, 0.224, 0.225) for each channel respectively. InceptionV3 requires images of size (299X299) and mean and standard deviation to be (0.485, 0.456, 0.406) and (0.229, 0.224, 0.225) for each channel respectively. VGG16 just like ResNet50 requires images of size (224X224) and mean and standard deviation to be (0.485, 0.456, 0.406) and (0.229, 0.224, 0.225) for each channel respectively. Originally, the DHCD dataset has greyscale images of size (32X32), so appropriate transforms were used to make the image match the configuration of these pretrained models so that we could fine tune and evaluate these models on the DHCD dataset.

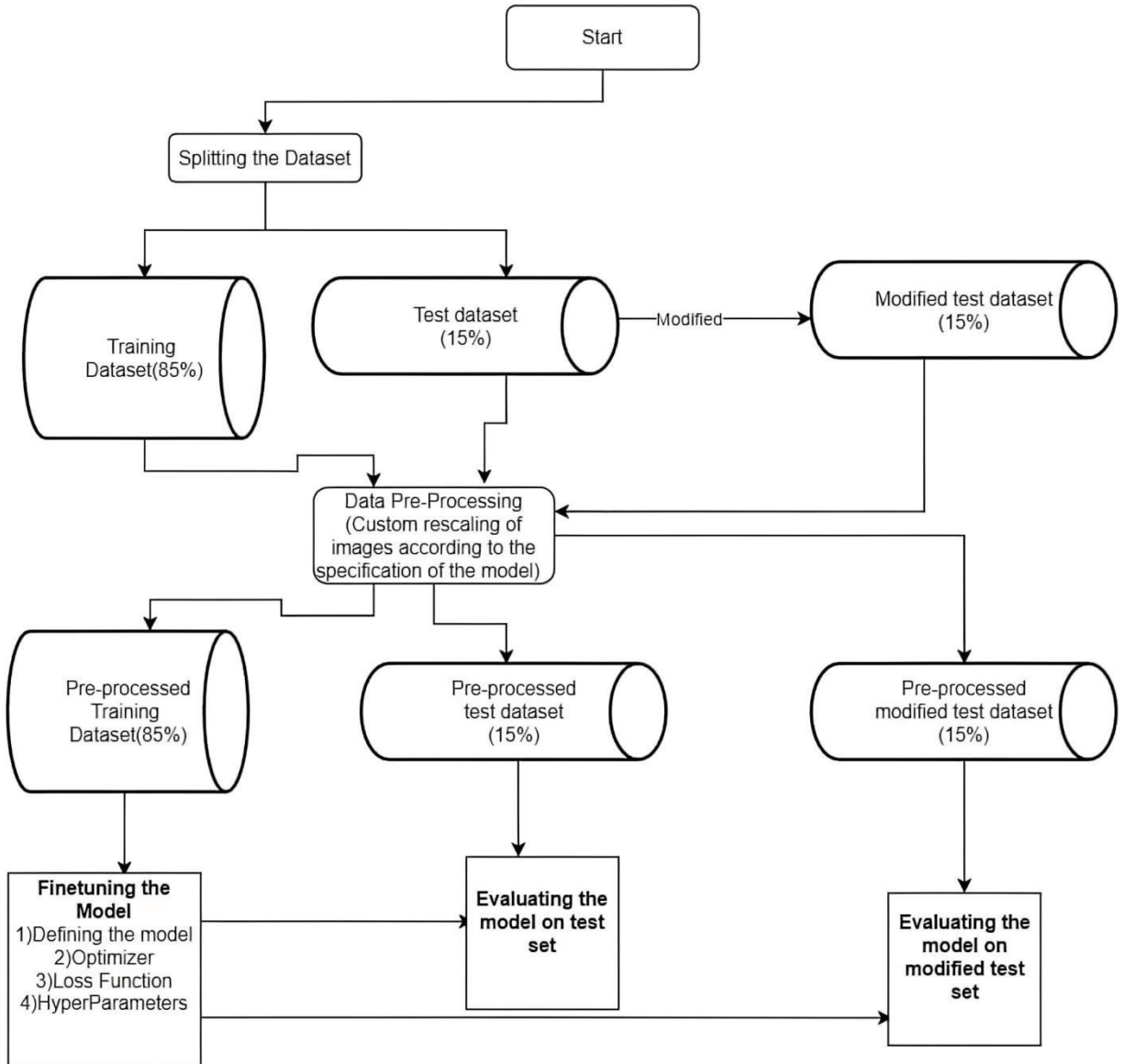


FIGURE 1. Entire methodology

### Vision Transformer (ViT)

Deep learning is a subset of the broader science of Machine Learning, which focuses on using artificial neural networks and other forms of neural nets to complete tasks based on their prior experience (training). This technique is mostly useful for recognizing the images and shapes among many other things [16]. Vision Transformers is an adaptation of Transformers to the world of computer vision [2]. In order to make an image ready to be fed into transformers, the image is first scaled to the size of 224X224 through Bi-linear interpolation and further transformed the pixel values, to have a mean of 0 and standard deviation as 1. Then it was divided into patches of size 16X16 and that patches were linearly transformed. Positional embeddings were added to the linearly transformed patches to prepare the token which is passed to the pretrained ViT base model [5] and further fine tuned to our task. The 0th token having embedding dimension of size 768 is passed through the output layer having 46 nodes and then probabilities are calculated using softmax classifier. Vision Transformer is a more generalized model than a CNN model as every token interacts with every other token because of a self attention which enables a patch to have an idea of every other patch from the very





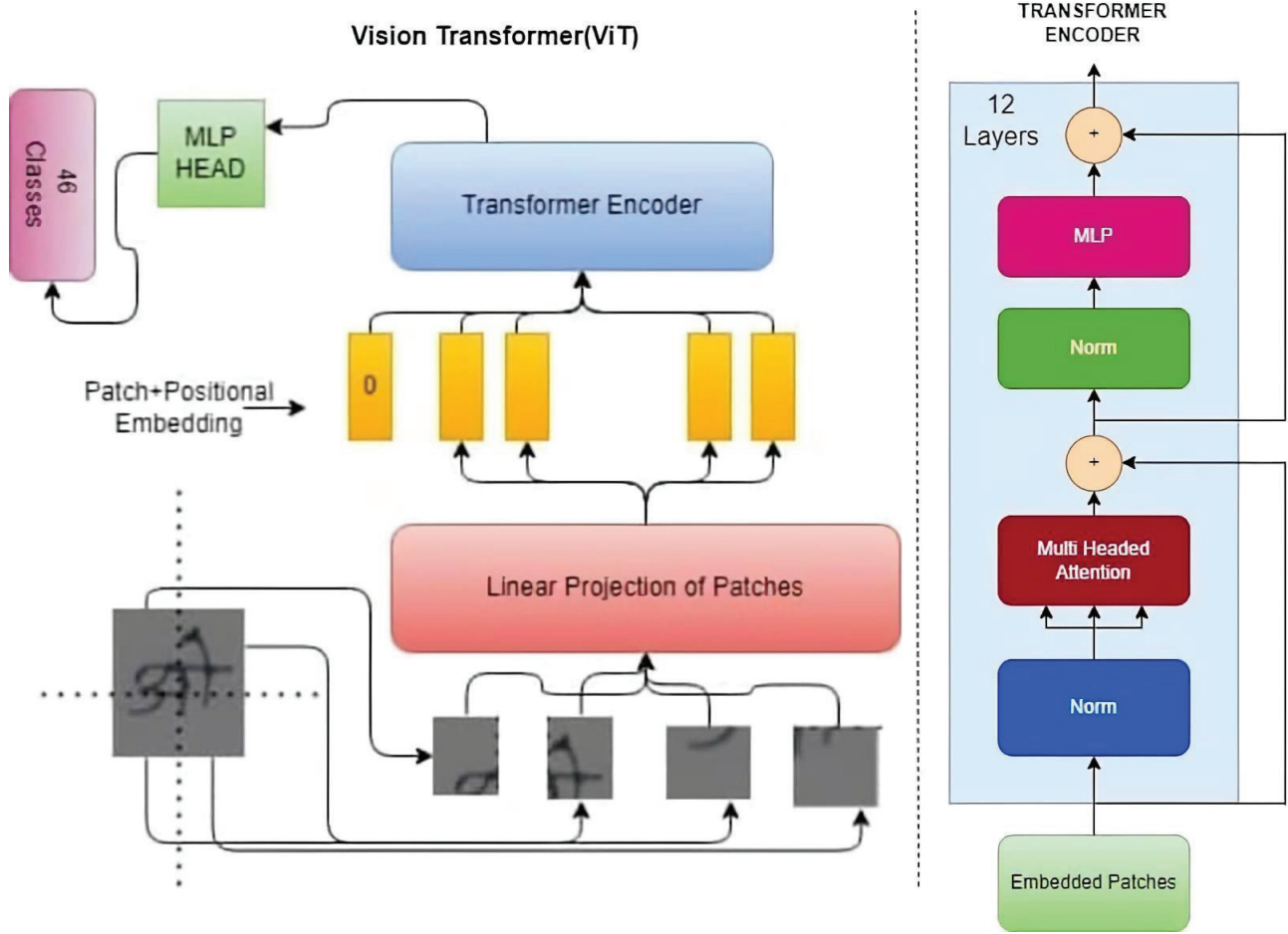
**FIGURE 2. Devanagari Character Set**

first layer itself. CNN uses filters of specific size which traverses over the image but with the bias that pixels close to each other are related and are taken together for feature computation, thus the initial layers predominantly capture more of intra patch features for overlapping patches and only after a few layers these features can interact further to form inter patch features. This depends on the size of the filter and of the image. So in the case of a few convolution layers there might be a case where a feature corresponding to a particular patch may not be aware of the faraway patch. This inter patch interaction is done in Vision Transformer from the very first layer itself and the model learns dynamic weights for interaction between these patches just making it more general than the CNN model. With the advent of Big Data, these models can outperform the current best models.

### **Transformer Encoder:**

The embedded patches which were obtained by linear transformation and adding of position embeddings are passed through Layer Norm in order to speed up training. These embeddings are then multiplied by Key, Query And Value Matrices [3] and we get key, query and value vectors on which multi headed attention is performed. For a particular query vector belonging to an embedded patch or token, a dot product is calculated with the key vectors of other patches and the values are softmaxed and a weighted average is taken with the value vectors. This description is of a single head which when extended to multiple heads (multiple key, query, value) vectors for a particular token is called multi headed attention. Residual Connection is made with embedded patches and so as to prevent the problem of vanishing gradient and facilitating faster convergence to a local optimum. This is again Layer Normalised and finally passed through Multi Layer Perceptron so as to introduce non linearity in the features extracted. This coupled with residual connection with the embedded patches gives the output of the layer. This layer is the fundamental unit of the transformer encoder and when stacked 12 times forming the ViT base model. This model is pretrained on a big dataset having 1000 classes and we have fine tuned it on our task by passing the 0th token (CLS token of transformer) to an output layer of 46 classes and training this modified model further on DHCD dataset.





**FIGURE 3. Depiction of Vision Transformer  
Standard CNN-based models**

In order to analyse the performance of ViT, we compared it with some standard CNN-based models namely: ResNet50, VGG16 and InceptionV3. All these models are known to perform very well in the field of computer vision. We loaded pre trained versions of these models through Pytorch's Torchvision [6] module and fine-tuned them on the DHCD dataset.

### Training Process

The ViT model was fine tuned using AdamW optimizer with the following hyperparameters: Epochs: 25 Learning rate: 0.00001 Betas: 0.9, 0.999 Weight decay: 0.01 The standard CNN models were fine-tuned using the SGD optimizer with: Learning Rate: 0.001 Momentum: 0.9 The loss function used in each model is Cross Entropy Loss. Figures 4 and 5 show the accuracy and loss plots obtained during the fine tuning process of Vision Transformer model over 25 epochs.

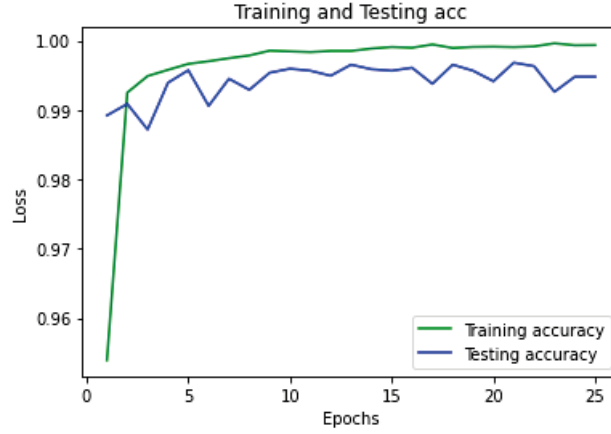


FIGURE 4. Analyzing Training and Testing Accuracy vs Epochs

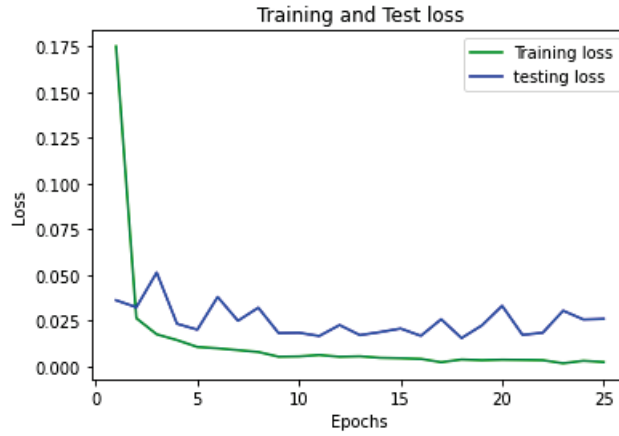


FIGURE 5. Analyzing Training and Testing Loss vs Epochs

## RESULTS AND ANALYSIS

Table I shows the results of the models on the two sets of testing datasets: i.e. the original testing dataset and the modified(blurred) testing dataset.

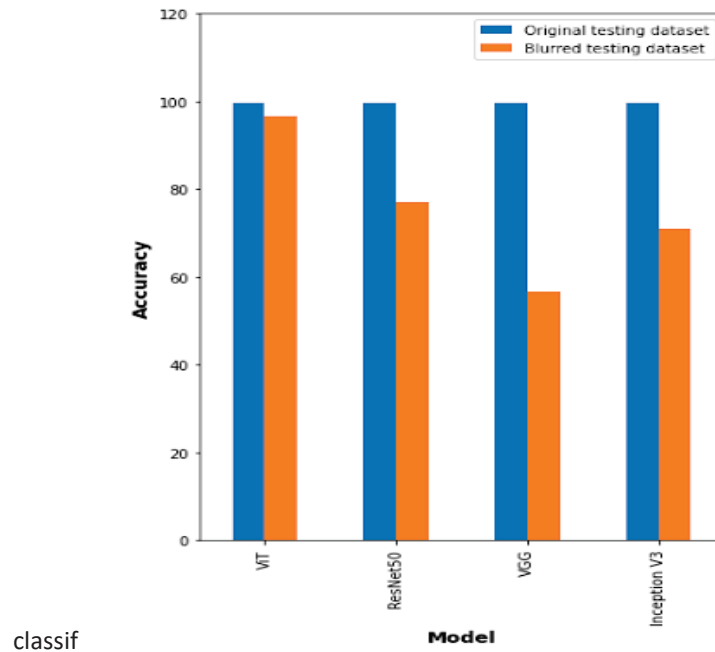
On the original testing dataset, The ViT model performed marginally better than the CNN-based models with an accuracy score of 99.68%. However, as it can be seen in Table I, when these models were evaluated on the modified(blurred) testing dataset, ViT outperforms CNN-based models by big margins. The accuracy scores of CNN based models drop off by huge margins with InceptionV3 having the highest accuracy of 77% and VGG16 dropping off to 57%. ViT, on the other hand, still gives an appreciable accuracy score of 96.55%

This gap in performance shows that the ViT model is better at generalizing. This could be due to the fact that ViT

TABLE I. This table illustrates accuracy of models on original and modified testing dataset

Model Used	Accuracy on Original Test Dataset	Accuracy on Modified Test Dataset
ViT	99.68%	96.55%
ResNet50	99.64%	77%
VGG16	99.64%	56.78%
InceptionV3	99.62%	70.92%

doesn't have any inherent bias and makes every token interact with each other just like a fully connected layer with dynamic weights and hence generalizes more. The CNN-based models on the other hand, have a bias that they focus more on local spatial features and start to generalize only after several layers. So ViT model has a higher chance to recognize an unseen image from another source due to this ability of better generalization on slightly complex data. Figure 6. also demonstrates this graphically as to how ViT, although marginally better than CNN-based models on the original testing dataset, generalizes much better when evaluated on blurred images which are slightly harder to



**FIGURE 6. Accuracy of different models on test and blurred dataset**

## CONCLUSION AND FUTURE WORK

- In this work, we successfully demonstrated how ViT could be applied to the field of Devanagari Character Recognition. The ViT model outperformed CNN-based models on slightly more complex data by big margins, proving that it could be very useful in recognizing characters in the real world.
- These days, almost every person owns a smartphone or any other technology that supports a camera. These are major sources of image data. But most of the people are not adept in capturing perfect images, so although all the models gave good performance on the original test dataset, the ViT model beat them with a big margin on more complex data, so in a real world where there are lots of imperfect images like blurred images, ViT is a better fit for such type of data.
- Also, as mentioned before, if a historian wants to extract information from digitally stored Devanagari scripts, where character recognition is an important step, here also there are chances of images being blurred or imperfect. So using ViT for this task would be a better choice.
- This current classifier can be extended to classify compound devanagari characters which are formed by combination of various half letter and full letter consonants and are very commonly found in Devanagari text.

## ACKNOWLEDGMENTS

The authors would like to thank Delhi Technological University for providing the resources and guidance for this work.

## REFERENCES

1. P. Hirugade, N. Suryavanshi, R. Bhagwat, S. Rajput, and R. Phadke, "A survey on optical character recognition for handwritten devanagari script using deep learning," Available at SSRN 4031738 (2022).
2. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929 (2020).
3. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems* 30 (2017).
4. K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data* 3, 1–40 (2016).
5. R. Wightman, "Pytorch image models," <https://github.com/rwightman/pytorch-image-models> (2019).
6. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems* 32, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019) pp. 8024–8035.
7. S. Acharya, A. K. Pant, and P. K. Gyawali, "Deep learning based large scale handwritten devanagari character recognition. skima 2015-9th international conference on software, knowledge," *Information Management and Applications* (2016).
8. S. R. Narang, M. Kumar, and M. K. Jindal, "Deepnetdevanagari: a deep learning model for devanagari ancient character recognition," *Multimedia Tools and Applications* 80, 20671–20686 (2021).
9. M. Bisht and R. Gupta, "Offline handwritten devanagari modified character recognition using convolutional neural network," *Sadhanā* 46, 1–4 (2021).
10. S. Kattyayan, T. Kar, and P. Kanungo, "Performance evaluation of learning based frameworks for devanagari character recognition," in *2020 IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)* (2020) pp. 1–6.
11. P. Gupta, S. Deshmukh, S. Pandey, K. Tonge, V. Urkunde, and S. Kide, "Convolutional neural network based handwritten devanagari character recognition," in *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)* (IEEE, 2020) pp. 322–326.
12. N. Aneja and S. Aneja, "Transfer learning using cnn for handwritten devanagari character recognition," in *2019 1st International Conference on Advances in Information Technology (ICAIT)* (IEEE, 2019) pp. 293–296.
13. A. N. Holambe, R. C. Thool, and S. Jagade, "Printed and handwritten character & number recognition of devanagari script using gradient features," *International Journal of Computer Applications* 2, 975–8887 (2010).
14. S. Acharya, A. K. Pant, and P. K. Gyawali, "Deep learning based large scale handwritten devanagari character recognition. skima 2015-9th international conference on software, knowledge," *Information Management and Applications* (2016).
15. V. P. Agnihotri, "Offline handwritten devanagari script recognition," *IJ Information Technology and Computer Science* 8, 37–42 (2012).
16. L. C. Yan, B. Yoshua, and H. Geoffrey, "Deep learning," *nature* 521, 436–444 (2015).

## Water quality management by enhancing assimilation capacity with flow augmentation: a case study for the Yamuna River, Delhi

Nibedita Verma <sup>a,\*</sup>, Geeta Singh<sup>a</sup> and Naved Ahsan<sup>b</sup>

<sup>a</sup> Department of Environmental Engineering, Delhi Technological University, New Delhi, Delhi 110042, India

<sup>b</sup> Civil Engineering Department, Jamia Millia Islamia University, Jamia Nagar, New Delhi 110025, India

\*Corresponding author. E-mail: nibedita\_2k19phd@dtu.ac.in

 NV, 0000-0002-1717-0570

### ABSTRACT

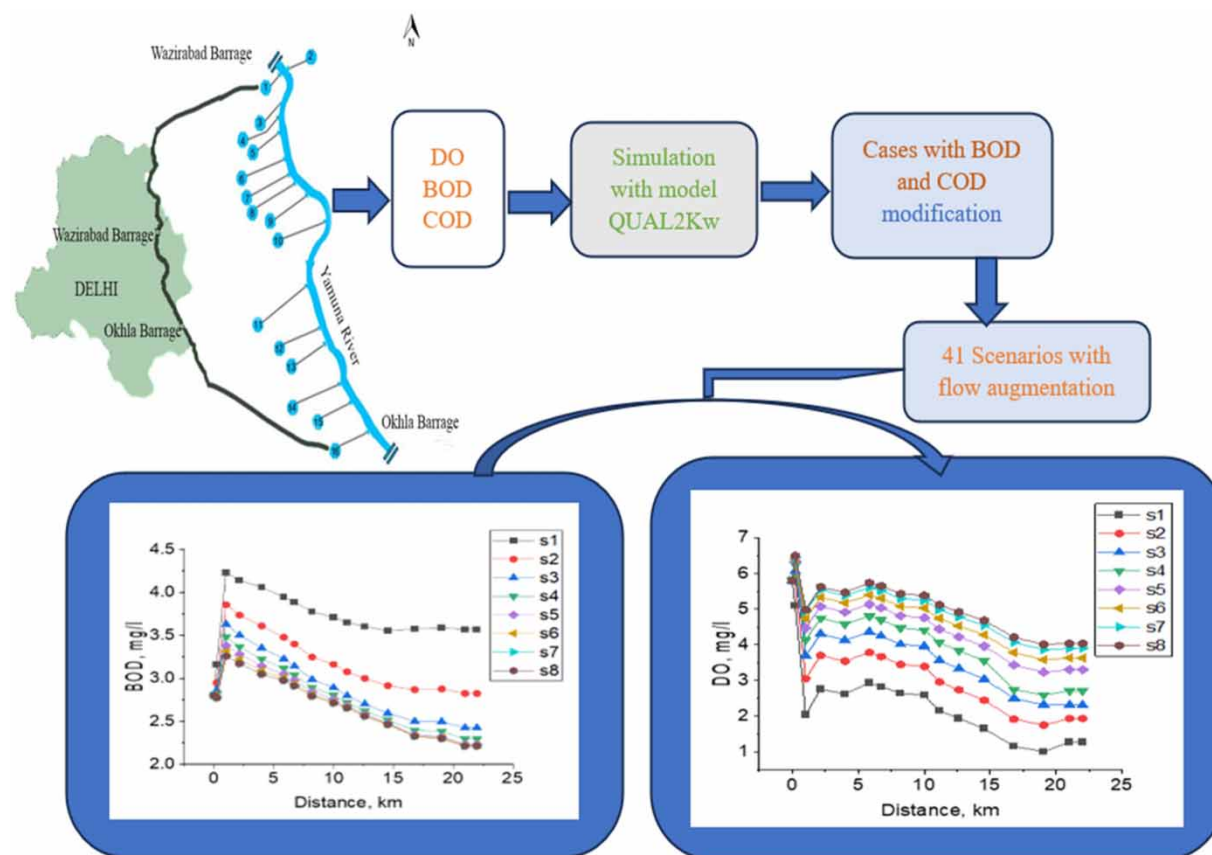
This paper aims to assess the requirement of load reductions and flow augmentation to enhance the assimilation yield of the Yamuna River, Delhi. The framework QUAL2Kw was used to predict river quality. The model was calibrated and confirmed in critical flow conditions of pre-monsoon periods. Three strategies were established for varying pollutant loads. The dissolved oxygen (DO) concentration was predicted with changing biochemical oxygen demand (BOD) and chemical oxygen demand (COD) loads. The 16 outfalling drains were considered pollutant sources between the 22 km stretch of the river. Four cases were studied with varying flow augmentation at upstream and varying load. It has been observed that with 80 cumecs of upstream flow, the reach can assimilate 31.33 TPD of BOD and 142.85 TPD of COD load, maintaining the desired level of DO ( $\geq 4$  mg/l) and BOD ( $\leq 3$  mg/l).

**Key words:** assimilation capacity, dissolved oxygen, QUAL2Kw, water quality model

### HIGHLIGHTS

- Managing the river quality by flow augmentation and load reduction.
- DO concentration prediction with varying BOD and COD.
- The Yamuna River, Delhi, reach can assimilate 31.33 TPD of BOD with 80 cumecs of upstream flow.
- This is a novel approach to assessing the self-purification capacity of the Yamuna River in Delhi by varying BOD, COD, and upstream flow with the model QUAL2Kw.

## GRAPHICAL ABSTRACT



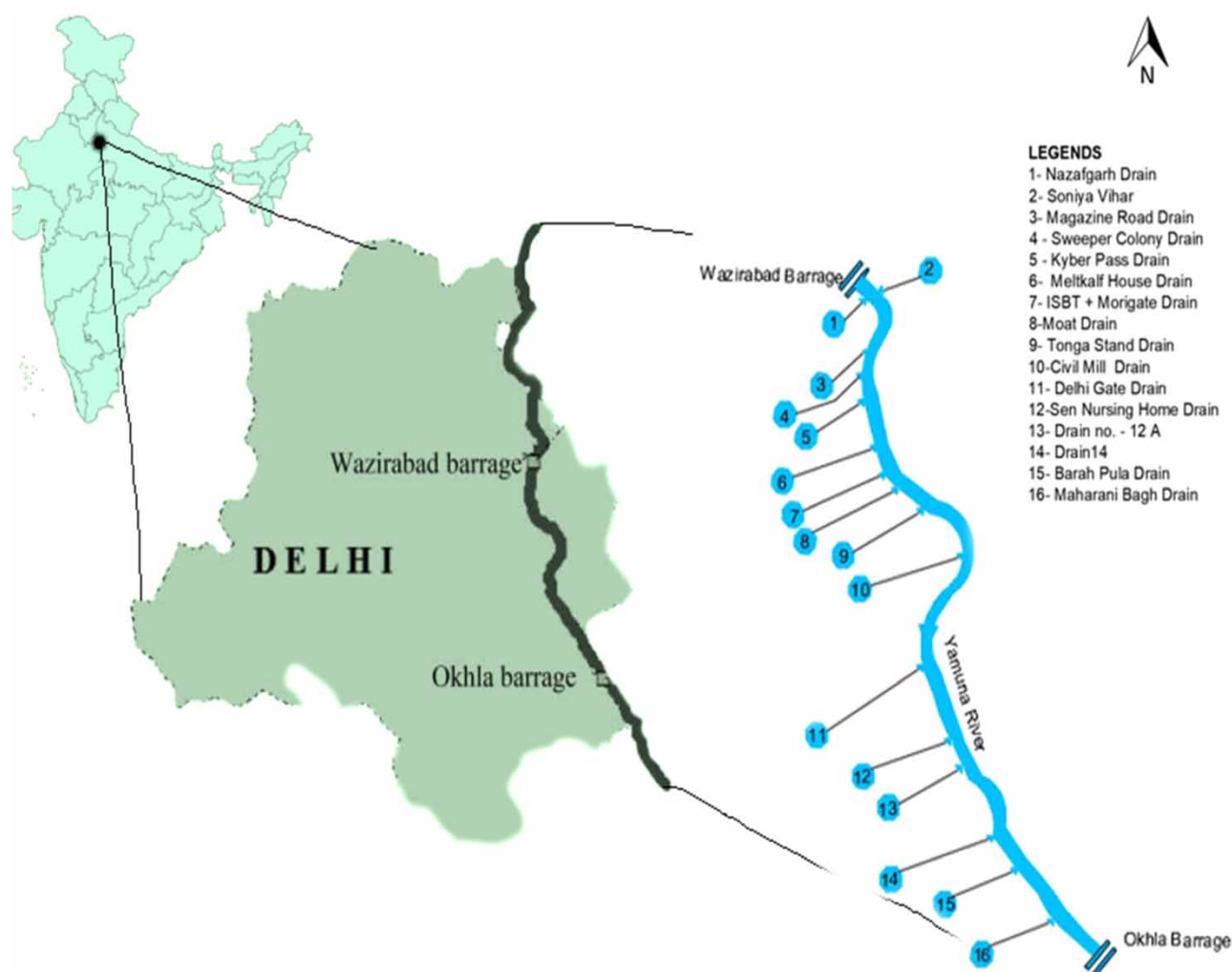
## 1. INTRODUCTION

Water is a crucial element of nature, and all civilizations have advanced near rivers. Nowadays, due to higher development activities with the massive population growth, the quality and quantity of water have become a sensitive issue, and freshwater will be scanty after a while (Pinto & Maheshwari 2011). Progressively worsening water quality results from rapid industrialization and urban sprawling, degrading the environment. Wastewater from the municipality and industry creates rivers' quality degradation and crucial global issues (González *et al.* 2014). The aquatic systems play an essential role in carrying out the pollutants accountable for water contamination (Shrestha & Kazama 2007). These contaminants stabilize with the system's physical, chemical, and biological processes. The pollution level rises when rivers' assimilation capacity is lower than the pollutants added to the water. The self-assimilation of aquatic systems is a complex phenomenon, including physiochemical and biological reactions. This phenomenon helps these systems regain water quality after flowing for a while if the water flow has sufficient substances to stabilize the waste input. Hence, self-purification and water quality enhancements are the primary criteria for natural water systems to sustain aquatic species (Wei *et al.* 2009). Self-purification is a way to partially or fully repair an aquatic system to a cleaner system after introducing foreign substances, causing a sufficient modification of the properties of water (Benoit 1971). The process is the recycling of substances with the assistance of physical, chemical, and biological processes. Dilution, adsorption, sedimentation, volatilization, acid-base reactions, precipitation reactions, coagulation, flocculation, bacterial degradation, and assimilation of materials by organisms (Vagnetti *et al.* 2003) are included in this process. When rivers flow, oxygen increases due to reaeration, and microorganisms present in sewage oxidize organic substances to inorganic materials and purify rivers (González *et al.* 2014). Thus, assimilation restores the conditions of the aquatic system before receiving wastewater (Ostroumov 2005). Nowadays, researchers focus on the self-assimilation of the contaminated river stretch as the water quality pattern is accountable and effortlessly modified with the environmental transformation (Wei *et al.* 2009). Rivers can be managed by improving the assimilation capacity. This capacity can be improved by



decreasing contaminants as well as increasing the freshwater flow. This study intends to manage the Delhi reach of the Yamuna River by enhancing assimilation capacity with flow augmentation.

In the waterbodies, dissolved oxygen (DO) depletes due to organic pollutants from wastewater. DO and biochemical oxygen demand (BOD) indicate the presence of organic substances (Basant *et al.* 2010). The oxygen-demanding contaminants' natural purification depends on the required oxygen to be concentrated, the oxygen necessary to secure the ecosystem qualities fit for species with designated standards, and the aquatic system's BOD purifying extent, which is assessed by reduction and replenishment of oxygen (Chapra *et al.* 2021). Rivers have their assimilation capacity, and it is necessary to acquire knowledge about the pollutant loads disposed of from diverse sources that rivers could receive without retrogression of their indigenous state (Oliveira *et al.* 2012). Mathematical modeling can validate waste load in a water body by establishing the cause–effect relationship between contaminant load and water quality. Hence, assimilation capacity could be evaluated by several simulation models (González *et al.* 2014). These frameworks are used as decision-making tools for wastewater management policies (McIntyre & Wheeler 2004). The simulation models correlate the water quality after being disposed of wastewater into a water body (Cox 2003). Water quality models can also predict the reciprocation of the aquatic system with different scenarios. The modeling outcomes are effective managing tools for assisting the river quality administrator in evaluating realistic water body conservation strategies and aspects of pollutant loading uncertainty. The present study aspires to assess the assimilation capacity of a severely polluted river stretch, Yamuna River, Delhi. The model QUA12Kw predicts the river quality and assesses the assimilating efficiency of the pollutant load. Kennel *et al.* (2007) appraised the



**Figure 1** | Study area showing 22-km river stretch with outfalling drains.

river conditions of the Bagmati River, Nepal, with QUAL2Kw, and the framework represents the observed data properly and is highly sensitive to water depth. Neilson *et al.* (2013) studied nutrient criteria and waste load analysis using the model QUAL2Kw and set nutrient criteria for the rivers in Utah. Zare Farjoudi *et al.* (2021) worked on the Zarjub River, Iran, to reduce the cost of waste load treatment and determination of self-purification capacity and found this framework as a suitable tool. The QUAL2Kw framework has been used for the Cetrina River, Portugal, enriched with nutrients. It was observed that the framework predicted the river quality parameter with limited data availability and evaluated the waterbody's conditions in modifications of the states (Oliveira *et al.* 2012). QUAL2Kw was used to simulate the load-carrying ability of the Kali Surabaya River, and the pollutant load for BOD and chemical oxygen demand (COD) was larger than the purification capacity of the river (Aliffia & Karnaningroem 2019). The seasonal variation of assimilation capacity for the Karun River, Tehran, has been determined using this model and stated that the different scenarios adopted for modeling, which were reducing wastewater flow, wastewater concentrations, and increasing the flow, enhanced the river characteristics (Moghimi Nezhad *et al.* 2018). Although there are several models to predict the pollutants' fate, due to easy accessibility, the ability to simulate maximum contaminants, and the availability of uncertainty analysis, QUAL2Kw is the most suitable tool for the calculation of the load-carrying ability of a water body (Darji *et al.* 2022). Hence, this study has used this framework to predict the assimilative capacity of the Yamuna River, Delhi, and water quality management of the severely polluted stretch of Yamuna, Delhi, by flow augmentation and reducing the BOD and COD pollutant load. The urban reach of Delhi is one of the most contaminated river stretches in India, and Delhi contributes 79% of the pollutant load (Joshi *et al.* 2022). This segment carries wastewater from different industries and municipal sewerage in Delhi (Parmar & Singh 2015). This segment of the Yamuna River is getting polluted by disposing of 22 outfalling drains within the 22 km from Wazirabad to Okhla (CPCB 2006). Before entering Delhi, the river shows medium water quality. After entering Delhi, due to discharging a massive BOD load and lack of fresh water, the river segment becomes a sewerage line (Upadhyay *et al.* 2011). Hence, it is the most crucial task to maintain the ecological health of this reach. The wastewater with little or no treatment has deteriorated the river reach (CPCB 2006). Uninterrupted wastewater input with excessive organic pollutants from different sources decreases river quality. The DO concentration becomes low when the river maintains low flow and receives huge wastewater flow (Gain & Giupponi 2015). Hence, flow augmentation is required to manage the water quality and increase the DO of such a polluted reach. The DO concentration of this river reach shows a sharp decline to zero or undetectable after discharging wastewater from the Nazafgarh drain, which is the prime contributor to waste load (Parmar & Singh 2015). Several studies appraised this segment's river quality (Kumar *et al.* 2019). However, little work has been done on the pollution-carrying capacity of this reach. The assimilation capacity of the Delhi reach of the Yamuna River was done by the Central Pollution Control Board of India (CPCB, 82) for COD and chloride, and four major contributing drains have been considered. Although the study was related to the discharging pollutant load from the point sources, Paliwal & Sharma (2007) used QUAL2E to assess the pollutant load-carrying capacity and suggested maintaining 10 cumecs of water flow to maintain the river water quality and only considered the BOD load. Parmar & Keshari (2014) used QUAL2E to study the waste load allocation for this stretch and recommended that flow augmentation was unsuitable for this reach. However, these studies did not include assessing the assimilation capacity by varying BOD and COD with the model QUAL2Kw and improving the load-carrying capacity with flow augmentation. Verma *et al.* (2022) suggested that this river reach required a combination of management options including load reduction, flow augmentation, and external aeration. The present study aims to understand the requirement of flow augmentation to enhance the assimilation capacity of the Yamuna River, Delhi, with suggested effluents standard and hence to manage the desired water quality of the river.

## 2. METHODOLOGY AND STUDY AREA

### 2.1. Study area

The Yamuna River flows from Yamunotri, India, and is the lengthiest tributary to the Ganga River adjoining Prayagraj after traversing 1,376 km from the origination. Before getting into Delhi at Palla, the river crosses 348 km, and the length of Delhi's reach is 48 km. At Palla, the river water shows desirable water quality conditions with low BOD and sufficient DO concentration (Joshi *et al.* 2022). After reaching Wazirabad, around 23 km from upstream (Joshi *et al.* 2022), most indigenous water withdraws and supplies to Delhi, and the perennial river contains little or no fresh water. From the Wazirabad barrage to Okhla upstream, the river feeds 16 main drains containing around 3,000 MLD of wastewater with 265 TPD of BOD load (Delhi Pollution Control Committee (DPCC) 2020). National Green Tribunal (2014) also reported that the national capital

of India leads to pollution of the Yamuna River Delhi stretch by drains containing domestic and industrial sewage. Due to the disposal of untreated and partially treated wastewater from different sewage treatment plants through these drains, the river stretch becomes mostly anoxic after the Wazirabad barrage. It contains high oxygen-demanding substances, microorganisms, and nutrients. The study includes Delhi's 22-km urban river reach between the Wazirabad barrage and the upstream of Okhla barrage and 16 main outfalling drains between these distances. The climatic condition of this area varies between hot in summer and cold in winter. The average summer temperature is 32 °C, with a maximum temperature of 45 °C. At the same time, the average temperature in winter is 12–13 °C, and the lowest temperature is around 2 °C (Arora & Keshari 2021). The monsoon period starts from late June to September, and the highest average rainfall was approximately 515 mm in August (Joshi *et al.* 2022). During this time, wastewater dilutes with rainwater and improves river quality. Hence, variation in water quality has been observed during the monsoon period.

## 2.2. Data and monitoring sites

The study obtained the data from the DPCC responsible for collecting data for the Delhi reach and all the drains outfalling between the distance. The monitoring stations included in this study covered five stations of DPCC S1 (Wazirabad downstream), S2 (ISBT), S3 (ITO), S4 (Nizamuddin bridge), and S5 (Okhla Upstream). The coordinates of these stations are shown in Table 1. The DO, BOD, COD, and pH water quality data were collected for March 2021 and April 2022. For the Delhi region, March–May is the low flow period due to negligible rainfall, known as the pre-monsoon period. The 16 outfalling drains were taken as point sources, and data were collected from DPCC. Due to data constraints, only four parameters were collected and simulated. The model QUAL2Kw was selected for this study, and the average data of March 2021 were used for calibration and those of April 2022 were used for confirmation.

## 2.3. Model setup

QUAL2Kw model has been used in this study to simulate BOD, COD, DO, and pH. This framework divided the reach into unequal, properly mixed segments of the same hydrological and water quality conditions (Kang *et al.* 2020). The 22-km river reach has been divided into 14 segments depending on the confluence of drains as point sources. Due to the instability of the model, drains outfalling in a very small distance have been taken in one segment. The model capabilities and descriptions are found in the user manual of QUAL2Kw. The framework is suitable for the river to reach with more or less constant pollutant loads and flow (Oliveira *et al.* 2012). The input data comprised headwater flow and water quality data, 16 outfalling drains wastewater flow, and quality data as point sources. These point sources carry domestic and industrial wastewater and discharge from sewage treatment plants. Delhi receives deficient rainfall; hence, surface runoff is very low. Besides this, some diffused sources of pollutant loads are used for cattle bathing, washing clothes, and bathing people. The model was calibrated using low flow and dry period data for March 2021. Geometrics and hydraulics data are shown in supplementary Table S1. Calibration was done repeatedly until the predicted values came closer to actual conditions. The manning constant and slope of the river have been taken as 0.05 and 0.0002, respectively (Group & Group 2001). The Manning equation was used because of limited data availability. The BOD values were taken as CBOD<sub>f</sub> and COD as generic constituents. The constituents included in the model are flow, temperature, BOD, DO, COD, pH, conductivity, and alkalinity. Due to data constraints, nutrient data were not included in this study. For the slow-moving river with shallow depth, the O'Connor–Dobbins equation has been used to calculate reaeration constants (Paliwal *et al.* 2007). The BOD and DO deal with the mechanism of sedimentation and settling, but due to low oxygen availability, 25% settleable BOD (CPCB, 82) decomposes in anoxic conditions and hence no trade of DO (Paliwal *et al.* 2007). Again, product methane rises upward, and due to

**Table 1** | Locations of monitoring stations

Monitoring sites	Coordinates
S1	28°42'47.27"N, 77°13'54.95"E
S2	28°40'16.87"N, 77°14'1.72"E
S3	28°37'42.34"N, 77°15'12.59"E
S4	28°35'29.62"N, 77°16'17.52"E
S5	28°32'40"N, 77°18'49"E

Source: DPCC.

buoyant forces, settled substances resuspend (Kazmi & Hansen 1997). Due to high turbidity, sunlight is obstructed (Kazmi 2000), and hence, phytoplankton activities are negligible. The DO variation due to photosynthesis and respiration is insignificant for this reach (Parmar & Keshari 2014). An exponential model was chosen for oxygen inhabitation for CBOD, and the calculation step was set to 5.625 for model stabilization. Except for the monsoon period, flow conditions are almost the same throughout the year (CPCB 2000). The non-monsoon flow prevails most of the year, and critical flow is essential to determine the assimilation capacity. For calibration and validation,  $1 \text{ m}^3/\text{s}$  flow is assumed at the upstream point. The headwater flow and quality are shown in supplementary Table S2. The point source input values are shown in supplementary Table S3. The Yamuna River, Delhi, is polluted with loads from different nonpoint sources. Although groundwater recharge is negligible for this area, pollution from nearby slum areas, cattle bathing, and agricultural runoff should be considered (Kazmi & Hansen 1997). In this study,  $1 \text{ mg/l}$  of distributed BOD load has been adjusted after a  $5 \text{ km}$  distance. The model was run until simulated values agreed with observed values. The framework was auto-calibrated for a population size of 100 and 50 generations, and simulation was done with new datasets for April for confirmation. The root-mean-square error was calculated to verify the calibration result with validation results.

#### 2.4. Scenario generation for assessment of assimilation capacity

The QUAL2Kw has been applied to assess the assimilation capacity of this polluted stretch. Hence, four cases were studied, generating 41 scenarios varying the BOD and COD load with flow augmentation. Supplementary Table S4 shows input BOD and COD loads with point source flow. The head water flow has been increased at 10 cumecs intervals for different scenarios. The scenarios were generated to achieve the water quality suggested for this river stretch, i.e., Class 'C' by the CPCB. For this criterion, river water should maintain DO greater than  $4 \text{ mg/l}$  and BOD less than  $3 \text{ mg/l}$ . The upstream flow has been increased to maintain this requirement by adjusting different BOD and COD loads. Flow augmentation of 10 cumecs increment was done upstream for developing scenarios of four cases shown in supplementary Figure S1.

### 3. RESULTS AND DISCUSSION

#### 3.1. Calibration and confirmation

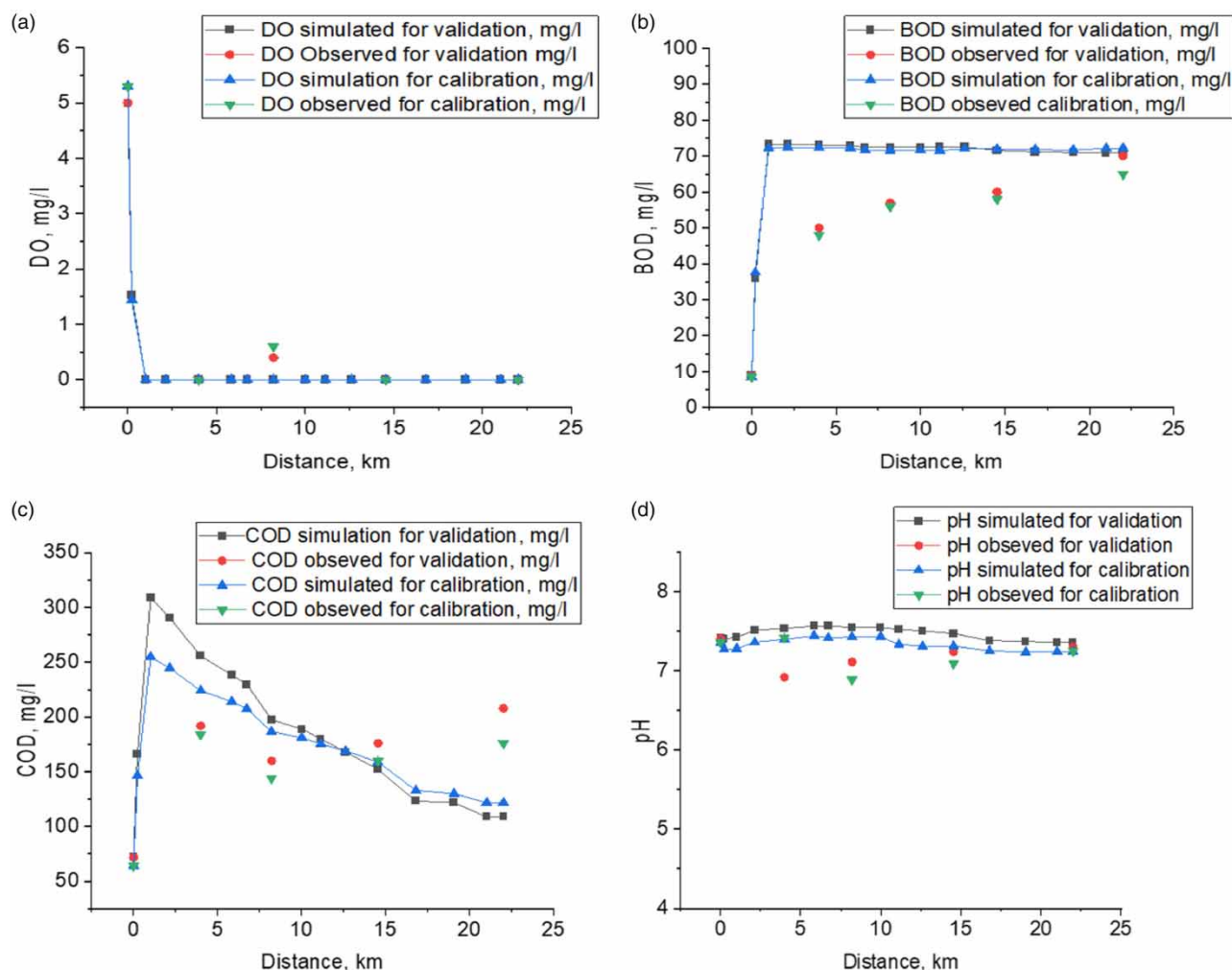
Figure 2(a)–2(d) shows DO, BOD, COD, and pH calibration and validation. Figure 2(a) shows that the DO reduced to zero after joining D1 (Najafgarh drain), the highest pollutant load contributor into this stretch, contributing around 58% of the total pollutant stress (Paliwal *et al.* 2007). Due to the high oxygen-demanding substances and low fresh flow, this river reach has become a sewerage line without DO. Figure 2(b) and 2(c) shows that after the outfalling of D1, BOD and COD values increased sharply. The RMSEVs for DO, BOD, COD, and pH were 16.28, 24.55, 24.09, and 4.5% for calibration and 17.03, 24.6, 35.19, and 4% for confirmation. Some errors are unavoidable as the single average values have been taken as monthly averages, and sampling times might be varied for different monitoring stations of 22 km long reach. Furthermore, point sources' wastewater qualities might vary depending on collection time and sampling procedure. More accurate predictions may be possible by collecting samples hourly for each monitoring station. Despite some inaccuracy, the QUAL2Kw framework has shown to be quite applicable for this river reach and can be adopted for water quality management purposes for data-limited conditions (Sharma *et al.* 2017; Verma *et al.* 2022).

#### 3.2. Strategies for assessment of the assimilation capacity of the river reach

Three strategies have been studied for the assessment of assimilation capacity. Table 2 shows the strategy adopted for assessment. Figure 4 shows the BOD, COD, and DO profiles without BOD and COD with headwater input shown in Table 3. Figure 3 shows the predicted DO, BOD, and COD profiles, and it has been observed that the river has a very low assimilative capacity. It can be concluded that with the flow of 1 cumec with  $2.8 \text{ mg/l}$  BOD and  $12 \text{ mg/l}$  COD, river reach is not able to maintain the required DO ( $\geq 4 \text{ mg/l}$ ) and BOD ( $\leq 3 \text{ mg/l}$ ). Hence, this reach is needed augmentation of flow upstream. As flow is deficient, the stream's reaeration capacity becomes poor; therefore, after some distance, DO reduction happens, and BOD increases. From Figure 3, it was observed that COD decreased, and thus, the oxygen requirement for COD was high. So, it needs to consider COD load, which was not considered in previous studies (Paliwal *et al.* 2007).

Figure 4 shows that with existing flow and pollutant load, DO concentration decreased to 0 throughout the river reach, and BOD concentration was also above  $60 \text{ mg/l}$  after outfalling of D1. Hence, pollutant load also requires to reduce with flow augmentation.





**Figure 2** | Simulated and observed values of DO, BOD, COD, and pH for calibration and validation.

**Table 2** | Strategies for assessment of assimilation capacity of the river reach

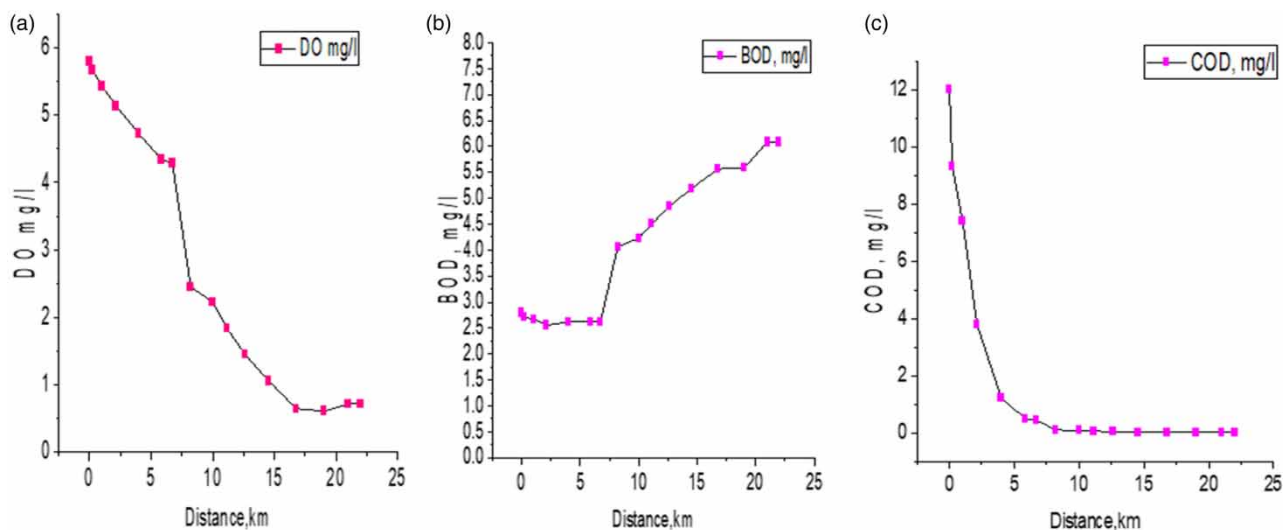
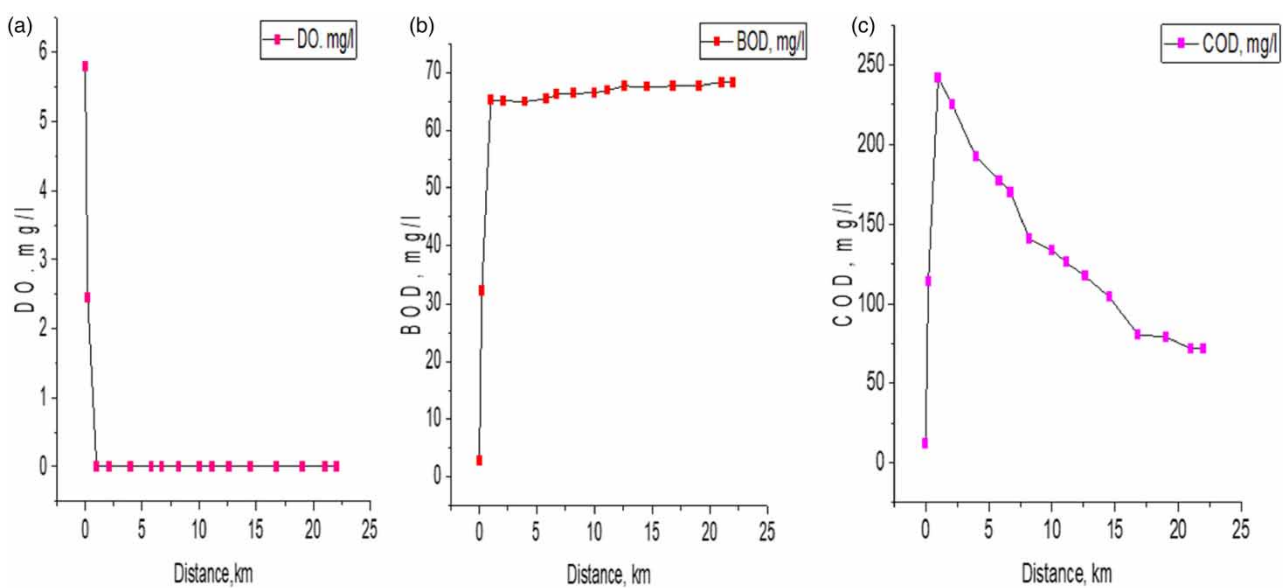
No.	Strategies
1	Without any pollutant load
2	With existing BOD and COD load
3	Four cases were generated with load modification and flow augmentation

In strategy 3, the flow has been increased from 10 cumecs up to 120 cumecs for scenarios s1–s12 of four cases, and BOD and COD loads have kept changing, as shown in Table S4. In case 1, 12 scenarios have been generated, increasing flow from 10 to 120 cumecs. BOD and COD have been kept at 10 and 50 mg/l, respectively, for all point sources, as effluent standards were set for this river stretch by the National Green Tribunal (NGT) of India.

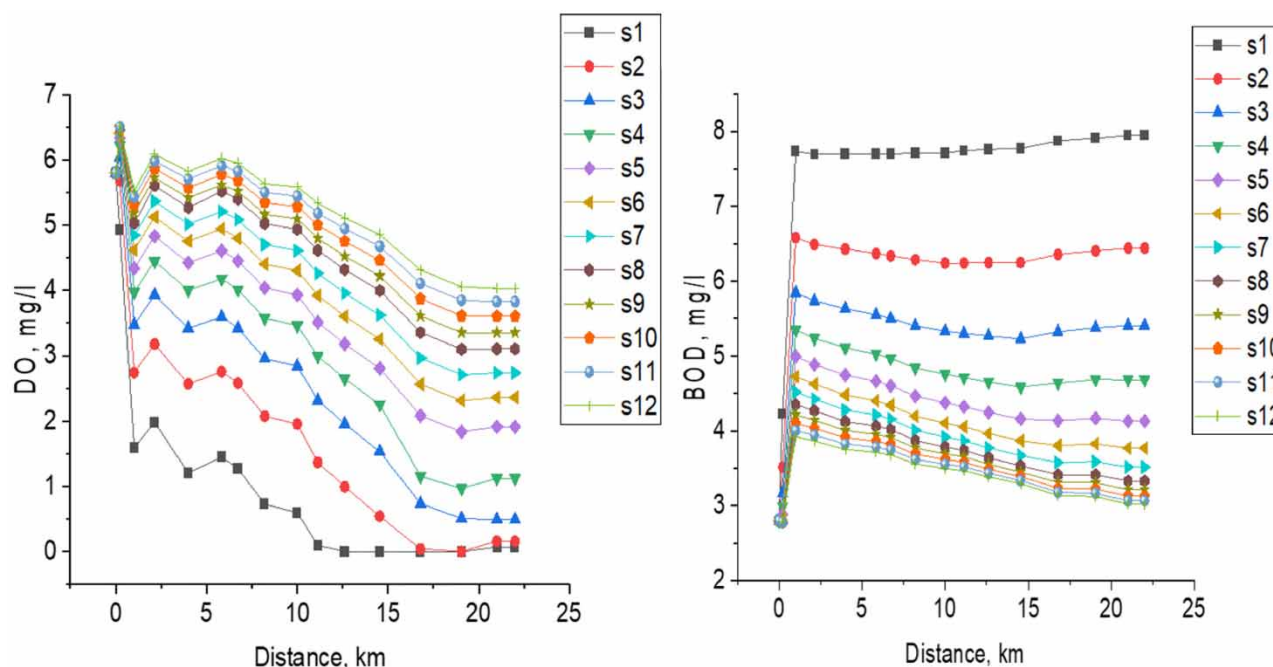
Figure 5 shows that with flow 120 cumecs, scenario 12 (s12), the reach can assimilate 10 mg/l of BOD and 50 mg/l of COD from each point source. Although DO maintained the required value ( $\geq 4$  mg/l) throughout the reach, the BOD level was higher than 3 mg/l. In case 2, BOD has been kept at 10 mg/l, and COD reduced to 25 mg/l in each point source and flow has been increased from 10 to 90 cumecs (s1–s9). Figure 6 shows that around 90 cumecs of flow augmentation upstream can maintain DO concentration. Although, at some distance, BOD is higher than 3 mg/l. In case 3, BOD has been reduced

**Table 3** | Headwater input for different strategies

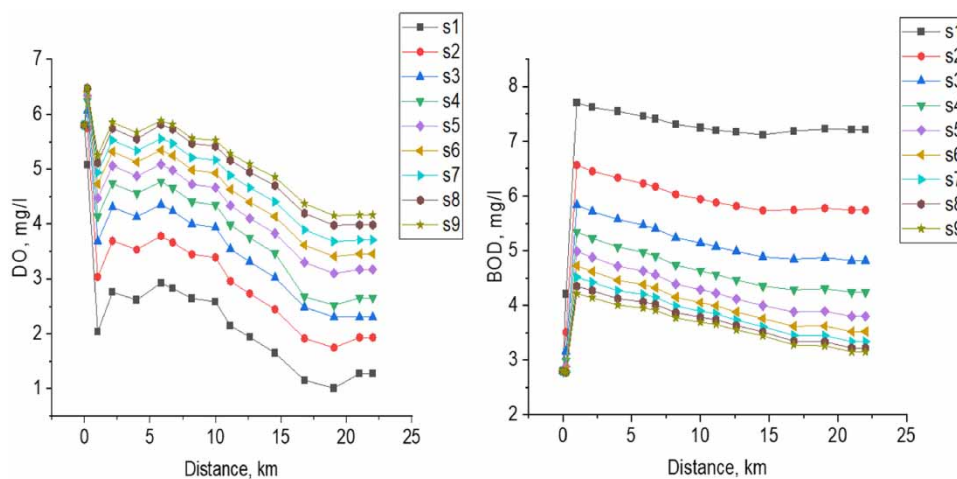
Parameters	Quantity
Flow	1 m <sup>3</sup> /s
DO	5.8 mg/l
BOD	2.8 mg/l
COD	12 mg/l
pH	7.4

**Figure 3** | Predicted DO, BOD, and COD without load.**Figure 4** | DO, BOD, and COD profiles with existing load.



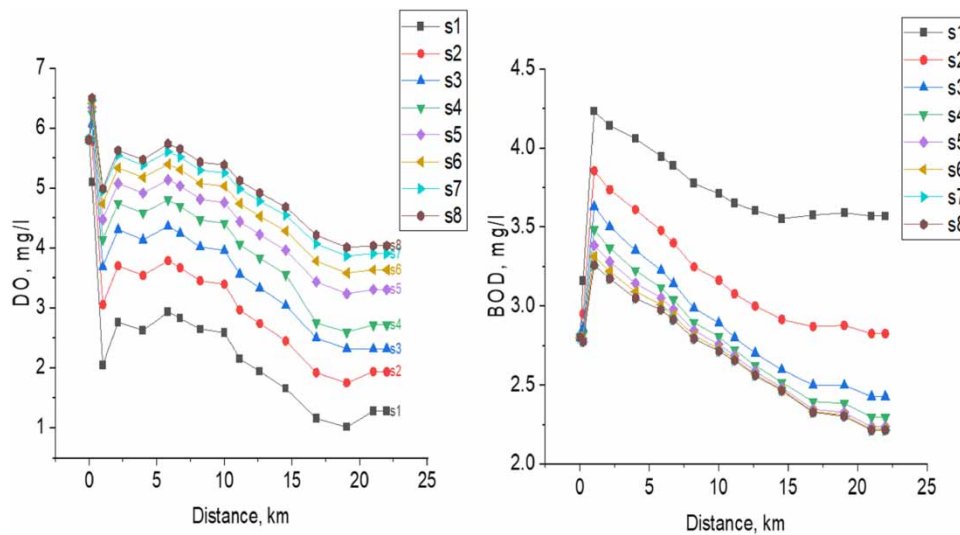


**Figure 5** | Scenarios for case 1 with varying flow.

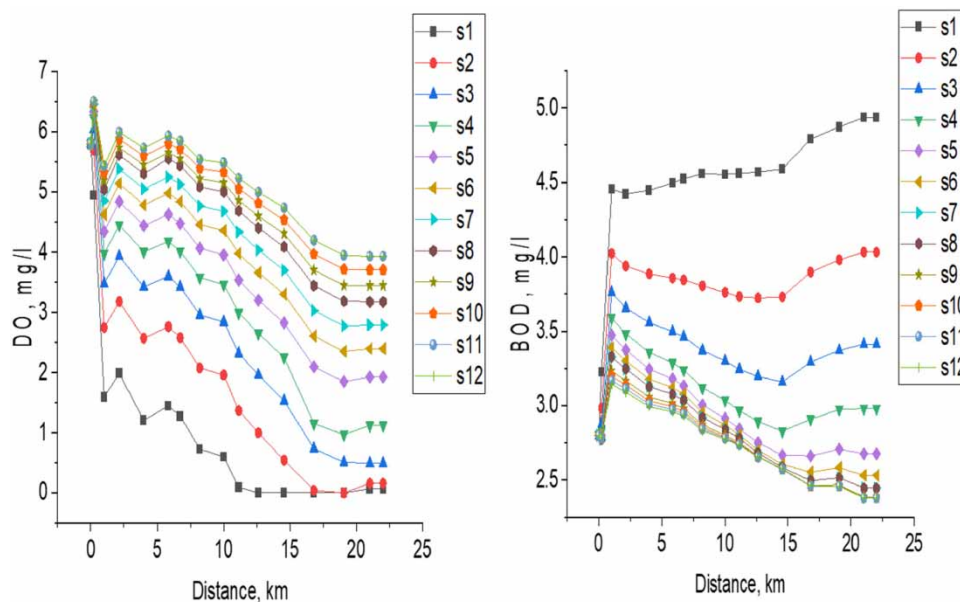


**Figure 6** | Scenarios for case 2 with varying flow.

to 5 mg/l in each point source, and COD has been kept at 25 mg/l. BOD has been observed to be maintained below 3 mg/l after 5 km upstream. Maintaining DO above 4 mg/l requires around 90 cumecs of flow upstream. Therefore, in case 4, BOD has been reduced in D1, D11, D12, and D15 to 5 mg/l; the rest have been kept at 10 mg/l. COD is also marked as the effluent standard prescribed by NGT. Figure 8 shows that 120 cumec flow upstream can maintain BOD and DO within the specified values. In case 3, with 80 cumecs of upstream flow, the reach can assimilate 31.33 TPD of BOD and 142.85 TPD of COD load. Kazmi & Hansen (1997), concluded that BOD and DO concentrations for effluent drains should be 10 and 4 mg/l to maintain the river water quality. Hence, the increase of DO in the point sources might be increased to improve assimilation capacity. They also suggested a 40 cumec upstream flow increment. Paliwal & Kansal (2007) indicated that some drains need to be diversified, and flow augmentation is required to maintain the required standard. The river reach requires a



**Figure 7** | Scenarios for case 3 with varying flow.



**Figure 8** | Scenarios for case 4 with varying flow.

combination of management options, including diversification of major drains with flow augmentation and advanced treatment (Verma *et al.* 2022). Some segments also require external aeration.

#### 4. CONCLUSION

The QUAL2Kw model assessed the assimilation capacity of Yamuna's most polluted stretch. The model is appropriate for this reach as it can be simulated with low data availability. Thus, it is ideal for decision-making tools like India, where limited data are available. This study revealed that the river's assimilation capacity was low due to high BOD and low DO levels. The wastewater enters the river from 16 drains and also diffused sources. Najafgarh drains added the highest wastewater quantity with elevated BOD and COD levels; thus, after adjoining this drain, the river's water quality fell to inferior. These conditions prevailed over the 22 km of this reach. In this study, the upstream flow increment with reduction of BOD and COD has been

studied. In strategy 1, wastewater from all drains was curtailed, and the desired standard of reach was not found. Improvement of the assimilation capacity of this river is a very challenging job, as the less upstream water with low DO and high BOD. In strategy 3, four cases were established with 41 scenarios with an increment of flow and reductions of BOD and COD. These cases suggested that load reduction and flow increment can improve the assimilation capacity of the river reach. This reach required substantial load cutting with flow dilution to enhance the water quality. Both the remedy options are very complicated and economically unfeasible. The study also revealed that COD and BOD are responsible for DO deterioration. It is also noted that the nitrogenous substances would improve the estimation of DO. As these drains carry domestic water containing nitrogenous waste, it has been suggested that regular monitoring of ammonium, organic nitrogen, and nitrate nitrite is also required.

## DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories.

## CONFLICT OF INTEREST

The authors declare there is no conflict.

## REFERENCES

- Aliffia, A. & Karnaningroem, N. 2019 Simulation of pollution load capacity using QUAL2Kw model in Kali Surabaya River (Cangkir-Sepanjang segment). *IOP Conference Series: Earth and Environmental Science* **259** (1). <https://doi.org/10.1088/1755-1315/259/1/012019>.
- Arora, S. & Keshari, A. K. 2021 Pattern recognition of water quality variance in Yamuna River (India) using hierarchical agglomerative cluster and principal component analyses. *Environmental Monitoring and Assessment* **193** (8). <https://doi.org/10.1007/s10661-021-09318-1>.
- Assimilation capacity of point pollution load The River Yamuna in the Union Territory of Delhi-Central Board for the Prevention and Control of India, 1982.
- Basant, N., Gupta, S., Malik, A. & Singh, K. P. 2010 Linear and nonlinear modeling for simultaneous prediction of dissolved oxygen and biochemical oxygen demand of the surface water – a case study. *Chemometrics and Intelligent Laboratory Systems* **104** (2), 172–180. <https://doi.org/10.1016/j.chemolab.2010.08.005>.
- Chapra, S. C., Camacho, L. A. & McBride, G. B. 2021 Impact of global warming on dissolved oxygen and bod assimilative capacity of the world's rivers: modeling analysis. *Water (Switzerland)* **13** (17). <https://doi.org/10.3390/w13172408>.
- Cox, B. A. 2003 A review of currently available in-stream water-quality models and their applicability for simulating dissolved oxygen in lowland rivers. *Science of the Total Environment* **314–316** (03), 335–377. [https://doi.org/10.1016/S0048-9697\(03\)00063-9](https://doi.org/10.1016/S0048-9697(03)00063-9).
- CPCB 2006 Water Quality Status of Yamuna River, Assessment and Development of River Basin, pp. 1–115. Available from: [www.cpcb.nic.in](http://www.cpcb.nic.in)
- Darji, J., Lodha, P. & Tyagi, S. 2022 Assimilative capacity and water quality modeling of rivers: a review. *Aqua Water Infrastructure, Ecosystems and Society* **71** (10), 1127–1147. <https://doi.org/10.2166/aqua.2022.063>.
- Gain, A. K. & Giupponi, C. 2015 A dynamic assessment of water scarcity risk in the Lower Brahmaputra River Basin: an integrated approach. *Ecological Indicators* **48**, 120–131. <https://doi.org/10.1016/j.ecolind.2014.07.034>.
- González, S. O., Almeida, C. A., Calderón, M., Mallea, M. A. & González, P. 2014 Assessment of the water self-purification capacity on a river affected by organic pollution: application of chemometrics in spatial and temporal variations. *Environmental Science and Pollution Research* **21** (18), 10583–10593. <https://doi.org/10.1007/s11356-014-3098-y>.
- Group, C. E. & Group, M. 2001 Water quality modeling of a stretch of the river Yamuna by Ajit Pratap Singh 1\* and S. K. Ghosh 2 1. December, 16–18.
- Joshi, P., Chauhan, A., Dua, P., Malik, S. & Liou, Y. A. 2022 Physicochemical and biological analysis of river Yamuna at Palla station from 2009 to 2019. *Scientific Reports* **12** (1). <https://doi.org/10.1038/s41598-022-06900-6>.
- Kang, G., Qiu, Y., Wang, Q., Qi, Z., Sun, Y. & Wang, Y. 2020 Exploration of the critical factors influencing the water quality in two contrasting climatic regions. *Environmental Science and Pollution Research* **27** (11), 12601–12612. <https://doi.org/10.1007/s11356-020-07786-5>.
- Kumar, B., Singh, U. K. & Ojha, S. N. 2019 Evaluation of geochemical data of Yamuna River using WQI and multivariate statistical analyses: a case study. *International Journal of River Basin Management* **17** (2), 143–155. <https://doi.org/10.1080/15715124.2018.1437743>.
- Mandal, P., Upadhyay, R. & Hasan, A. 2010 Seasonal and spatial variation of Yamuna River water quality in Delhi, India. *Environmental Monitoring and Assessment* **170** (1–4), 661–670. <https://doi.org/10.1007/s10661-009-1265-2>.
- McIntyre, N. R. & Wheeler, H. S. 2004 A tool for risk-based management of surface water quality. *Environmental Modelling and Software* **19** (12), 1131–1140. <https://doi.org/10.1016/j.envsoft.2003.12.003>.
- Moghim Nezad, S., Ebrahimi, K. & Kerachian, R. 2018 Investigation of seasonal self-purification variations of Karun River, Iran. *Amirkabir Journal of Civil Engineering* **49** (4), 193–196. <https://doi.org/10.22060/ceej.2016.866>.

- Neilson, B. T., Hobson, A. J., VonStackelberg, N., Shupryt, M. & Ostermiller, J. (2013). *Using Qual2K Modeling to Support Nutrient Criteria Development and Wasteload Analyses in Utah*, pp. 1–49.
- Oliveira, B., Bola, J., Quinteiro, P., Nadais, H. & Arroja, L. 2012 *Application of Qual2Kw model as a tool for water quality management: Cértima River as a case study*. *Environmental Monitoring and Assessment* **184** (10), 6197–6210. <https://doi.org/10.1007/s10661-011-2413-z>.
- Ostroumov, S. A. 2005 *On some issues of maintaining water quality and self-purification*. *Water Resources* **32** (3), 305–313. Translated from *Vodnye Resursy*, Vol. 32, No. 3, 2005, pp. 337–346.
- Paliwal, R. & Sharma, P. 2007 Application of QUAL2E for the river Yamuna: to assess the impact of pointloads and to recommend measures to improve water quality of the river. *Environment* **2702**.
- Paliwal, R., Sharma, P. & Kansal, A. 2007 *Water quality modelling of the river Yamuna (India) using QUAL2E-UNCAS*. *Journal of Environmental Management* **83** (2), 131–144. <https://doi.org/10.1016/j.jenvman.2006.02.003>.
- Parmar, D. L. & Keshari, A. K. 2014 *Wasteload allocation using wastewater treatment and flow augmentation*. *Environmental Modeling and Assessment* **19** (1), 35–44. <https://doi.org/10.1007/s10666-013-9378-y>.
- Parmar, S. & Singh, V. 2015 *Water quality parameters of river Yamuna in Delhi after 20 years of the Yamuna action plan* (Vol. 6, Issue 4).
- Pinto, U. & Maheshwari, B. L. 2011 *River health assessment in peri-urban landscapes: an application of multivariate analysis to identify the key variables*. *Water Research* **45** (13), 3915–3924. <https://doi.org/10.1016/j.watres.2011.04.044>.
- Sharma, D., Kansal, A. & Pelletier, G. 2017 *Water quality modeling for urban reach of Yamuna river, India (1999–2009), using QUAL2Kw*. *Applied Water Science* **7** (3), 1535–1559. <https://doi.org/10.1007/s13201-015-0311-1>.
- Shrestha, S. & Kazama, F. 2007 *Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan*. *Environmental Modelling and Software* **22** (4), 464–475. <https://doi.org/10.1016/j.envsoft.2006.02.001>.
- Upadhyay, R., Dasgupta, N., Hasan, A. & Upadhyay, S. K. 2011 *Managing water quality of River Yamuna in NCR Delhi*. *Physics and Chemistry of the Earth* **36** (9–11), 372–378. <https://doi.org/10.1016/j.pce.2010.03.018>.
- Vagnetti, R., Miana, P., Fabris, M. & Pavoni, B. 2003 *Self-purification ability of a resurgence stream*. *Chemosphere* **52** (10), 1781–1795. [https://doi.org/10.1016/S0045-6535\(03\)00445-4](https://doi.org/10.1016/S0045-6535(03)00445-4).
- Verma, N., Singh, G. & Ahsan, N. 2022 *Development of water quality management strategies for an urban river reach: a case study of the river Yamuna, Delhi, India*. *Arabian Journal of Geosciences* **15** (24). <https://doi.org/10.1007/s12517-022-11030-4>.
- Wei, G. L., Yang, Z. F., Cui, B. S., Li, B., Chen, H., Bai, J. H. & Dong, S. K. 2009 *Impact of dam construction on water quality and water self-purification capacity of the Lancang River, China*. *Water Resources Management* **23** (9), 1763–1780. <https://doi.org/10.1007/s11269-008-9351-8>.
- Zare Farjoudi, S., Moridi, A. & Sarang, A. 2021 *Multi-objective waste load allocation in river system under inflow uncertainty*. *International Journal of Environmental Science and Technology* **18** (6), 1549–1560. <https://doi.org/10.1007/s13762-020-02897-5>.

First received 14 June 2023; accepted in revised form 12 September 2023. Available online 25 September 2023