SCHOLARLY PUBLICATIONS

A CURRENT AWARENESS BULLETIN
OF RESEARCH OUTPUT

@DTU

(106th Edition)

OCTOBER 2021

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 Sixth** 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 **October, 2021**. 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 22nm 10T FinFET SRAM Cell with Improved Write and Read performance in Sub-Threshold Region, ***6.Monica Gupta***, Kirti Gupta and ***3.Neeta Pandey***, Electronics, DTU

2. A Comparison of Similarity Measures for Neighbourhood Based Collaborative Filtering Recommender Systems, ***3.Rohit Beniwal***, ***8.Rahul Khairwal***, ***8.Ritvik Mahajan*** and ***8.Sarthak Pratap Narayan Singh***, CSE, DTU

3. A Green Data Collection & Transmission Method for IoT-Based WSN in Disaster Management, Aridaman Singh Nandan, Samayveer Singh, Aruna Malik and ***3.Rajeev Kumar***, CSE, DTU

4. A Modified-Zeta Converter based Onboard Battery Charging with Improved THD, ***7.Chanchal Chaudhary***, ***3.Saurabh Mishra*** and ***3.Uma Nangia***, Electrical, DTU

5. A Neural Network Based Image Steganography Method using Cyclic Chaos and Integer Wavelet Transform, ***3.Manoj Kumar*** and ***8.Tahera Hussaini***, CSE, DTU

6. A Review of Blockchain-Based Applications and Challenges, ***8.Pratima Sharma***, ***3.Rajni Jindal*** and Malaya Dutta Borah, CSE, DTU

7. A NOVEL LOW POWER DESIGN OF BRENT KUNG ADDER HAVING FAULT TOLERANT, ***7.Sumit Pahuja*** and ***3.Gurjit Kaur***, Electronics, DTU

8. An Automated Multi-Web Platform Voting Framework to Predict Misleading Information Proliferated during COVID-19 Outbreak using Ensemble Method, **6.Dee*pika Varshney*** and ***3.Dinesh Kumar Vishwakarma***, IT, DTU

9. An Evolution of Style Transfer from Artistic to Photorealistic : A Review, ***7.Gaurav Sohaliya*** and ***3.Kapil Sharma***, IT, DTU

10. An Explicit Output Current-mode Quadrature Sinusoidal Oscillator and a Universal Filter Employing Only Grounded Passive Components - a Minimal Realisation, Tajinder Singh ARORA, ***8.1.Manish GUPTA*** and Shiv Narain GUPTA, Electronics, DTU

11. An Improved Fusion Technique for IR and VIS Image using LWT and PCA, ***7.Urvashi Rawat*** and ***3.Sudipta Majumdar***, Electronics, DTU

12. Analysis of barriers intensity for investment in big data analytics for sustainable manufacturing operations in post-COVID-19 pandemic era, ***3.Narender Kumar***, ***3.Girish Kumar*** and Rajesh Kr Singh, Mechanical, DTU

13. ANN and ANFIS Based Approach for Very Short Term Load Forecasting : A Step Towards Smart Energy Management System, ***6.Arshi Khan*** and ***3.M. Rizwan***, Electrical, DTU

14. Applications of Adaptive Long Short-Term Memory to Active Filtering, ***3.Alka Singh***, Electrical, DTU

15. Applied convolutional neural network framework for tagging healthcare systems in crowd protest environment, ***8.Gaurav Tripathi***, Kuldeep Singh and ***3.Dinesh Kumar Vishwakarma***, ECE and IT, DTU

16. APPROXIMATION BY A COMPOSITION OF APOSTOL-GENOCCHI AND PˇALTˇANEA-DURRMEYER OPERATORS, ***6.NAV SHAKTI MISHRA*** AND ***3.NAOKANT DEO***, Applied Mathematics, DTU

17. Assessing the sustainability of municipal solid waste management using life-cycle analysis, ***8.K. Patel*** and ***3.S. K. Singh***, Environmental, DTU

18. Automatic Image Colorization with Convolutional Neural Networks, ***8.Mohammad Amir Qureshi***, ***8.Mayank Gautam***, ***8.Kaushal Pal*** and ***3.N. Jayanthi***, Electronics, DTU

19. Banking Reforms and Competition: Performance of Indian Banking Industry, Renu Gupta and ***3.K. V. Bhanu Murthy***, USME, DTU

20. BGR to HSV based Text Extraction from Manuscripts Using Slidebars, ***7.Mayank Singh*** and ***3.Prof. S. Indu***, Electronics, DTU

21. BOHR PHENOMENON FOR K-QUASICONFORMAL HARMONIC MAPPINGS AND LOGARITHMIC POWER SERIES, ***6.KAMALJEET GANGANIA***, Applied Mathematics, DTU

22. Booth-Encoded Karatsuba: A Novel Hardware-Efficient Multiplier, ***8.Riya JAIN***, ***8.1.Khushbu PAHWA*** and ***3.Neeta PANDEY***, Electronics and Electrical, DTU

23. Centellosides: pharmaceutical applications and production enhancement strategies, ***8.Ritika Luthra***, Garvit Satija and Arpita Roy, Biotechnology, DTU

24. Circularly Polarized Antenna for ISM (5.8 GHz), Satellite Communications and UWB Applications, Navneet Sharma, Anubhav Kumar and ***3.Asok De***, Electronics, DTU

25. Comparative Analysis of Non-Isolated DC-DC Converters for Solar-Photovoltaic System, ***7.Shivangi Yadav***, ***3.Saurabh Mishra*** and ***7.Garima***, Electrical, DTU

26. Comparative Analysis of Image Fusion Techniques for Infrared and Visible Images, ***7.Urvashi Rawat*** and ***3.Sudipta Majumdar***, Electronics, DTU

27. Comparison of IRPT, ANN and ANFIS based Control Algorithm for Shunt Compensation in Grid Connected Systems, ***6.Shreya Sinha*** and ***3.Ankita Arora***, Electrical, DTU

28. Defects assisted luminescence in m-HfO2 nanocrystals: An experimental and theoretical study, ***6.Rajesh Kumar***, Ankush Vij and ***3.Mukhtiyar Singh***, Applied Physics, DTU

29. Design and Modeling of a Filter Assisted 6-Pulse VSC-IGBT Based D-Statcom for Reactive Power Compensation, ***8.Biswabijet Panigrahy***, ***8.Bushra Arif***, ***8.Ritul Panchal***, ***3.Radhe shyam Saha*** and ***3.Krishna Dutt***, Electrical, DTU

30. Detection, Localization and Retrieval of tampering/Alteration in Microsoft Word Documents for Authentication and Copyright Protection using Multiple Techniques, ***8.Manmohan***, ***8.Lavanisha Prasad***, ***8.Nitin*** and ***3.Jeebananda Panda***, Electronics, DTU

31. Drowsiness Detection using Light Framework, ***8.Amit Pratap Singh***, ***3.Ritu Agarwal***, ***8.Chirag Gupta*** and ***8.Himanshu***, IT, DTU

32. EDGly: detection of influential nodes using game theory, ***3.Minni Jain***, ***8.Aman Jaswani***, ***8.Ankita Mehra*** and ***8.Laqshay Mudgal,*** CSE, DTU

33. Effect of swirl flow on the performance of parallel hub axial annular diffuser, ***6.1.Hardial Singh*** and ***3.B.B.Arora***, Mechanical, DTU

34. Estimation and Prediction of Deterministic Human Intent Signal to augment Haptic Glove aided Control of Robotic Hand, ***6.Rajesh Kumar***, Pimmy Gandotray, Brejesh Lally, Arzad A. Kheraniz and ***3.Sudipto Mukherjee***, Mechanical, DTU

35. Exploiting feature space using overlapping windows for improving biometric recognition, Surinder Kaur, ***7.1.Gopal Chaudhary***, Smriti Srivastava, Manju Khari, Ruben Gonzalez Crespo and Javalkar Dinesh Kumar, Computer, DTU

36. Fake News Detection using Semi-Supervised Graph Convolutional Network, ***3.Priyanka Meel*** and ***3.Dinesh Kumar Vishwakarma***, IT, DTU

37. Friedmann cosmology with decaying vacuum density in Brans-Dicke theory, ***3.C. P. Singh*** and Joan Sol `a Peracaula, Mathematics, DTU

38. Grounded-Capacitor Analog Inverse Active Filters using CMOS OTAs, ***6.Ajishek Raj***, ***3.Ram Bhagat***, ***3.Pragati Kumar*** and ***3.D. R. Bhaskar***, Electrical, DTU

39. Heart Failure Prediction using Voting Ensemble Classifier, ***7.Shubham Kaushik*** and ***3.Rajesh Birok***, Electronics, DTU

40. High-Capacity Reversible Data Hiding using Deep Learning, ***8.Mohd Asif***, ***8.Lokesh Kumar***, ***8.Gaurav Swami*** and ***3.Ankita Arora***, Electrical, DTU

41. Identifying and Mitigating Gender Bias in Hyperbolic Word Embeddings, ***8.Vaibhav Kumar***, ***8.1.Tenzin Singhay Bhotia***, **8.*Vaibhav Kumar*** and Tanmoy Chakraborty, IT, DTU

42. Implementation and Comparison of Direct mapped and 4-way Set Associative mapped Cache, ***3.Dr. Gurjit Kaur***, ***8.Riyanshi Arora*** and ***8.Sakshi Shailesh Panchal***, Electronics, DTU

43. Intelligent Testing in Software Industry, ***8.Anushka Lal*** and ***3.Girish Kumar***, CSE and Mechanical, DTU

44. Interfacing of Digital and Resolver Type Systems for Transmission of Heading Angle, ***6.Vijendra Kumar Singh*** and ***3.Dr. Yashna Sharma***, Electronics, DTU

45. Internet of Things Based Energy Harvesting using Piezoelectric Discs, ***3.Deva Nand***, ***8.Aakash Yadav***, ***8.1.Pradeep Kumar*** and ***8.Srinivas Kumar***, Electronics, DTU

46. LCNT incorporated domino with footed diode inverter, ***6.Neetika Yadav***, ***3.Neeta Pandey*** and ***3.Deva Nand***, Electronics, DTU

47. Matrix games with proportional linguistic payoffs, ***6.Parul Chauhan*** and ***3.Anjana Gupta***, Mathematics, DTU

48. Mechanical and Microstructural Properties of Rubberized Concrete After Surface Modification of Waste Tire Rubber Crumb, ***6.Rahul Kumar*** and ***3.Nirendra Dev***, Civil, DTU

49. Multi-modal Fusion using Fine-tuned Self-attention and Transfer Learning for Veracity Analysis of Web Information, ***3.Priyanka Meel*** and ***3.Dinesh Kumar Vishwakarma***, IT, DTU

50. NOMA-based UAV system under finite blocklength regime with analysis in Rician fading channel, Shardul Thapliyal, Rajoo Pandey and ***7.1.Chhagan Charan***, Electronics, DTU

51. Numerical simulation of cascaded cyclone separator for nanosize aerosol, Sonawane Chandrakant R, Atul Kulkarni, Rohan V. Sawant, Anand Kumar Pandey, Hitesh Panchal, Kishor Kumar Sadasivuni and ***3.Anil Kumar***, Mechanical, DTU

52. On Certain Exact Differential Subordinations Involving Convex Dominants, ***3.S. Sivaprasad Kumar*** and ***7.Shagun Banga***, Mathematics, DTU

53. Online Robust MPC based Emergency Maneuvering System for Autonomous Vehicles, ***8.1.Vivek Bithar***, Punit Tulpule and Shawn Midlam-Mohler, Production, DTU

54. Ovarian Cancer: Potential biomarkers and nanotechnology based diagnostic tools, ***7.Shruti Sounkaria*** and ***3.Prakash Chandra***, Biotechnology, DTU

55. Performance comparison of the solar-driven supercritical organic Rankine cycle coupled with the vapour-compression refrigeration cycle, ***6.Yunis Khan*** and ***3.Radhey Shyam Mishra***, Mechanical, DTU

56. PM-BLDC Motor Drive for Solar Water Pumping with Modified Boost Converter, ***7.Shivangi Yadav***, ***3.Saurabh Mishra*** and ***6.Garima***, Electrical, DTU

57. Realization of IIMC based Higher-Order Floating Fractional-Order Element, ***6.Garima Varshney***, ***3.Neeta Pandey*** and ***3.Rajeshwari Pandey***, Electronics, DTU

58. Semantic Segmentation using Generative Adversarial Networks with a Feature Reconstruction Loss, ***7.Gaurav Sohaliya*** and ***3.Kapil Sharma***, IT, DTU

59. Sentiment Analysis of News Headlines Using Simple Transformers, ***7.Anurag Singh*** and ***3.Goonjan Jain***, Applied Mathematics, DTU

60. Stacked Sparse Autoencoder and Machine Learning Based Anxiety Classification Using EEG Signals, ***8.Shikha, 8.Manan Agrawal***, ***8.Mohd Ayaan Anwar*** and ***3.Divyashikha Sethia***, Software Engineering, DTU

61. State Estimation Along WLS-Phasor Measurements in Power System, ***6.Ankit Kumar Dhaikar*** and ***3.S.T. Nagarajan***, Electrical, DTU

62. Studying Electromagnetic field pattern for Breast Cancer Detection by Hexagonal Patch Antenna, ***6.Shruti Awasthi*** and ***3.Dr. Priyanka Jain***, Electronics, DTU

63. The family of Bernstein–Kantorovich operators with shifted knots, ***6.Ram Pratap*** and ***3.Naokant Deo***, Mathematics, DTU

64. The study of effect of vitamin C and Achyranthes aspera seeds enriched diets on the growth, biochemical composition, digestive enzyme activities and expressions of genes involved in the biosynthesis of fatty acids in Snow trout Schizothorax richardsonii (Gray, 1832), Guddu Kumar, ***3.Jai Gopal Sharma***, Ravi Kumar

Goswami, Avanish Kumar Shrivastav, ***6.Neelesh Kumar***, Suresh Chandra, and Rina Chakrabarti, Biotechnology, DTU

65. Thermal-Electrochemical Modeling and Analysis of Different Cathode-Anode Pairs for Lithium-ion Battery, ***6.Snigdha SHARMA***, ***3.Amrish K. PANWAR*** and ***3.Madan Mohan TRIPATHI***, Electrical and Applied Physics, DTU

66. Threshold Type Binary Memristor Emulator based on DVCC, ***7.Juli Kumari Roy*** and ***3.Neeta Pandey***, Electronics, DTU

67. TWO-WAY METRIC LEARNING WITH MAJORITY AND MINORITY SUBSETS FOR CLASSIFICATION OF LARGE EXTREMELY IMBALANCED FACE DATASET, ***7.Ashu Kaushik*** and ***3.Seba Susan***, IT, DTU

68. Unidirectional Onboard Battery Charging for Electric Vehicle using Interleaved Luo Converter, ***7.Chanchal Chaudhary***, ***3.Saurabh Mishra*** and ***3.Uma Nangia***, Electrical, DTU

69. WCA Optimized Fuzzy PD controller for Stabilizing the Two Wheel Self-Balancing Robot, ***6.Likhitha Kakinada*** and ***3.Kuldeep Singh***, Electrical, DTU

1. *Vice Chancellor*
2. *Pro Vice Chancellor*
3. *Faculty*
4. *Teaching-cum-Research Fellow*
5. *Asst. Librarian*
6. *Research Scholar*
7. *PG Scholar*
8. *Undergraduate Student*

1.1. *Ex Vice chancellor*
2.1. *Ex Pro Vice Chancellor*
3.1. *Ex Faculty*
4.1. *Alumni*
5.1 *Others*
6.1. *Ex Research Scholar*
7.1. *Ex PG Scholar*
8.1. *Ex Undergraduate Student*

2021 8th International Conference on Signal Processing and Integrated Networks (SPIN) | 978-1-6654-3564-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SPIN52536.2021.9566009

A 22nm 10T FinFET SRAM Cell with Improved Write and Read performance in Sub-Threshold Region

Monica Gupta
Dept. of Electronics and Communication, Delhi Technological University, New Delhi, India
monica.gupta.me@gmail.com

Kirti Gupta
Dept. of Electronics and Communication, Bharati Vidyapeeth's College of Engineering, New Delhi, India
kirtigupta22@gmail.com

Neeta Pandey
Dept. of Electronics and Communication, Delhi Technological University, New Delhi, India
neetapandey@dce.ac.in

***Abstract*— Although voltage scaling presents an attractive alternative to overcome the issue of power consumption in the nanometer regime, it degrades the performance of static random access memories. This paper presents a 10T SRAM cell at 22 nm FinFET technology to achieve improved write and read performance in the sub-threshold region. A write assist transistor incorporated in the latch structure helps the node voltages to flip easily during a write operation and improve the write performance. In addition, the issue of high and data-dependent read bitline leakage is resolved by maintaining similar operating conditions in the read port of un-selected cells independent of stored data. This enhances the read bitline voltage swing compared to the other CMOS cells. The FinFET based design of the cell maintains a high read current in the read port and reduces read access time. Compared to the conventional 8T CMOS cell, the proposed design provides 19 %, 99.9 % and 70 % improvement in RSNM, read access time and read bitline voltage swing respectively at 0.3 V. Apart from significant improvement observed in write margin and write access time, an improvement of 17 % and 77 % are also noted in hold static noise margin and retention voltage in the sub-threshold region.**

Keywords—10T, SRAM, FinFET, read bitline voltage swing, access time, noise margin, sub-threshold

I. INTRODUCTION

WITH the growing need for system-on-chips and faster processors, the demand for high performance, low-power and reliable SRAMs is increasing drastically [1]. In the current technology regime, an attractive approach to reduce the power requirement is supply voltage scaling [2]. Recently, the designers are focusing on sub-threshold region so that the the point with minimum energy consumption can be reached [3]. However, in the nanometer regime, maintaining the performance under 20 nm is becoming difficult with bulk CMOS transistors [2]. FinFET based designs are emerging as the most promising substitute for planar CMOS based circuits due to good scaling ability, high ON current, reduced Vth variations, better Subthreshold-Slope and Short-Channel-Effect[1][4].

In a conventional 6T CMOS cell, read stability and write abilty can't be optimized simultaneously. So, an isolated read port is introduced in a conventional 8T CMOS cell design to ensure read stabilty almost equal to hold stability. However, the high and data-dependent read bitline (RBL) leakage of the read port results in reduced RBL voltage swing. Moreover, at low supply voltages, the reduced drive of the MOS transistors fails to flip the node voltages making it difficult to perform the write operation.

Many improvised cell structures are suggested in the literature to address these issues [5][6]. One such structure is presented in [5]. The modified read port design of the cell reduces the RBL leakage slightly but fails to address the data-dependency issue [7]. Another 10T cell [6] removes the data-dependency problem but at the cost of increased RBL leakages [7]. Moreover, none of the cell addresses the degraded write performance in sub-threshold region. Therefore, in the paper a new FinFET SRAM cell design with improved performance in sub-threshold region is proposed.

The paper is prepared as follows. The existing CMOS cell designs are briefly reviewed in section II. The proposed 10T FinFET cell structure with the operation is described in section III. The performances of all the cells are compared in section IV. The conclusion is drawn in section V.

II. EXISTING CMOS CELL DESIGNS

The structure of a SRAM cell comprises of a latch to store either '1' or '0' at the internal storage nodes and access transistors to read the stored data or write new one. In an array of SRAM cells, the cells are further organized in many rows and columns. There are several ways in which the circuit for the cell can be realized. The classification of cells are usually done on the basis

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:37:37 UTC from IEEE Xplore. Restrictions apply.

of total transistors required in the circuit execution. A few such circuits are reviewed below.

A. CONVENTIONAL 8T CMOS SRAM CELL (8T_CONV)

The 8T_CONV cell structure (Fig. 1(a)) consists of a latch (MUL-MDL, MUR-MDR) that stores the data at internal storage nodes D and DB. The differential write operation is performed by the access transistors (MAL and MAR) along with the bitlines (WBL and WBLB) and wordline signal (WWL). An isolated read port (M1 and M2) reads the data stored at node DB using bitline (RBL) and wordline signal (RWL). The read cycle begins with the pre-charging of RBL to V_{DD} and pulling of RWL signal high in the selected cell at the positive and negative edge of the clock cycle respectively. When the cell stores '0' at node D, the RBL discharges and a high read current flows through the transistors M1 and M2. Alternatively, when node D stores '1', the RBL voltage remains high. However, due to the flow of data-dependent and high RBL leakage current in un-selected cells (having RWL low) connected in the column, RBL discharges, resulting in reduced RBL voltage swing and wrong sensing of stored data [7]. To store the new data, the values are applied on the complementary bitlines, WBL and WBLB, and WWL is pulled high at the positive and negative edge of the clock cycle respectively. However, in sub-threshold region due to increased strength of pMOS transistor than nMOS transistor at slow-fast process corner, the write operation suffers resulting in degraded write performance.

B. 10T CMOS SRAM CELL (10T_CAL) [5]

A 10T_CAL [5] uses a modified four transistors based read port (Fig. 1(b)). The design partially addresses the RBL leakage current issue of un-selected cells by reducing one component of the leakage current. The RBL leakages remain high and data-dependent resulting in reduced RBL voltage swing [7]. Moreover, the stacked transistors deteriorate the read current values and increase read access time.

C. 10T CMOS SRAM CELL (10T_SH) [6]

Another cell, introduced in [6], addresses the high and data-dependent RBL leakage issue of 8T_CONV cell. The cell successfully overcomes the data-dependency problem but at the cost of increased RBL leakages resulting in insignificant improvement in RBL voltage swing. In addition, read access time remains high due to degraded read current values [7].

The existing cell designs are therefore unsuccessful in addressing the increased read access time and deteriorated RBL voltage swing problems of an isolated read port. Moreover, the use of same structure for performing write operation as in 8T_CONV cell, results in poor write performance in sub-threshold region. This necessitates a new cell design that can address these issues and provide improved read and write performance in sub-threshold region.

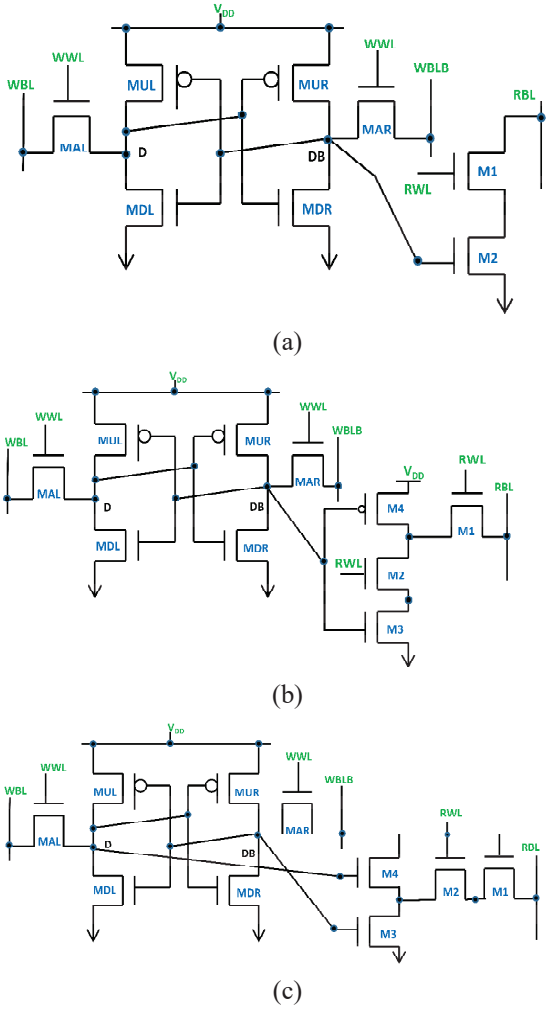


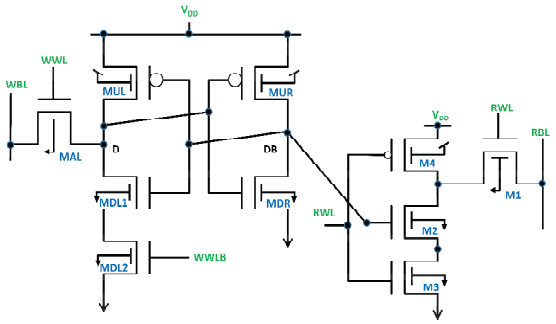
Fig. 1. Schematic (a) 8T_CONV cell (b) 10T_CAL cell [5] (c) 10T_SH cell [6]

III. PROPOSED 10T FINFET SRAM CELL

The architecture of the proposed cell that performs single-ended operations for reading and writing process is presented in Fig. 2. It comprises of the basic latch structure (MUL-MDL and MUR-MDR) to store one bit of information in a complementary way at its internal storage nodes D and DB. It performs write operation using access transistor MAL, wordline signal WWL and corresponding bitline WBL. A write assist transistor MDL2, driven by complementary signal WWLB, is incorporated in the latch structure to assist the write operation.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:37:37 UTC from IEEE Xplore. Restrictions apply.

An isolated read port (M1, M2, and M3) is used to accomplish the read operation using wordline signal RWL along with the bitline RBL. A transistor M4 is introduced to enhance the read performance of the cell.



(a)

Mode / Control signal	Read	Write	Hold
WWL	0	1	0
WWLB	1	0	1
RWL	1	0	0

(b)

Fig. 2. Proposed 10T FinFET cell (a) Schematic diagram (b) Operating conditions.

A. Write-mode

During write operation, the new value is applied on WBL through the write driver. The complementary signals WWL and WWLB are then pulled high and low respectively in the selected cell. To improve the write performance, a write assist transistor (MDL2) driven by control signal WWLB turns OFF and disables the path between storage node D and circuit ground. This helps the node voltages to flip quickly during write operation without disrupting the strong feedback mechanism of the latch structure. The advantage is that the write performance of the selected cell is improved without causing the stability issues in the other cells connected to the same column.

B. Read-mode

At the beginning, RBL is pre-charged using pre-charge circuit to V_{DD} and then RWL signal is turned high in the selected cell along the row. Now, if DB holds '1' then RBL voltage reduces or it remains high otherwise. The change in RBL voltage is further used for sensing the stored data and complete the operation. Usually, the RBL leakage current that flows through the isolated read port of an un-selected cell (having RWL low) is data-dependent. Due to this, the RBL voltage discharges even when the selected cell holds '0' at DB resulting in reduced RBL voltage swing and wrong sensing of stored data. To get over this problem, a transistor M4 is incorporated in the cell. The transistor M4 turns ON in the un-selected cells and maintains a voltage close to V_{DD} at drain of M4 irrespective of the value stored at DB. This reduces the drain to source voltage across M1. The result is significant reduction in RBL leakage current independent of stored data. Thus, achieving low and data-independent RBL leakages and improved RBL voltage swing. The high read current of FinFET based cell provides significant improvement in read access time in sub-threshold region

C. Hold-mode

Both the wordline signals (RWL and WWL) are kept low while WWLB is held high during this mode. This isolates the internal storage nodes from the write and read bitlines. The write assist transistor MDL2 conducts and connects the pull-down transistor MDL1 to the circuit ground. In addition, the stacking effect of two pull-down transistors (MDL1 and MDL2) decreases the leakages resulting in high stability of stored data [8].

IV. SIMULATION AND DISCUSSIONS

The performance of the proposed 10T FinFET SRAM cell is evaluated and compared with 8T_CONV, 10T_CAL and 10T_SH CMOS cells based on different parameters in three operating modes. All the simulations are performed in Tanner TSPICE tool using 22 nm FinFET and CMOS PTM model parameters as mentioned in Table I.

TABLE I. FINFET AND CMOS PARAMETERS

FinFET parameters		**CMOS parameters**	
name	**value**	**name**	**value**
Gate length	22 nm	Length of channel	22 nm
Oxide thickness	1.1 nm	Channel width	44 nm
Fin thickness	15 nm	Threshold voltage nMOS, Vthn	0.68858 V
Fin height	28 nm	Threshold voltage pMOS, Vthp	-0.63745 V
Source/drain doping	3e26 cm^{-3}		
SS	75 mV/decade		
DIBL	59 mV/V		

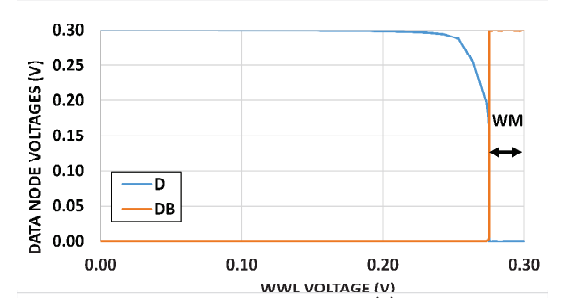
A. Write-mode

The key parameters such as write ability and access time are evaluated for comparison. The ease with which the writing process can be accomplished is measured in terms of write margin [10]. The time required to complete the writing process is evaluated in terms of access time.

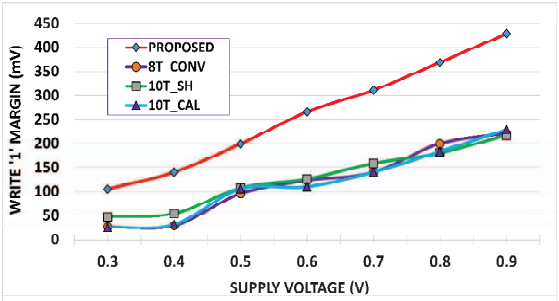
1) Write Margin (WM) [10]: The WM is measured using the method mentioned in [10] (Fig. 3(a)) and the corresponding

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:37:37 UTC from IEEE Xplore. Restrictions apply.

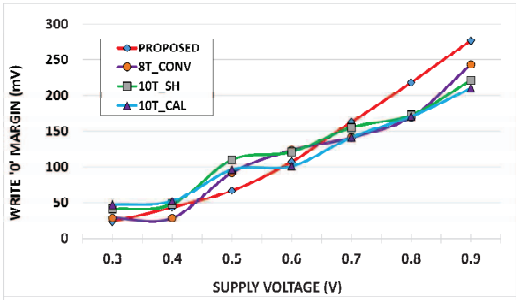
values are displayed in Fig. 3(b) and 3(c) for WM '1' and '0' respectively at different voltages. As can be seen, the proposed cell provides reasonable values of WM '1' at all voltages. The improvement achieved in WM '1' is 299 %, 123 % and 322 % compared to 8T_CONV, 10T_SH and 10T_CAL cells respectively at 0.3 V. The proposed cell also provides improvement in WM '0' (Fig. 3(c)) at higher voltages whereas the performance remains comparable to other cells at lower voltages.



(a)



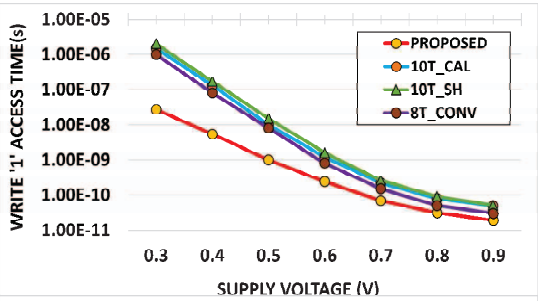
(b)



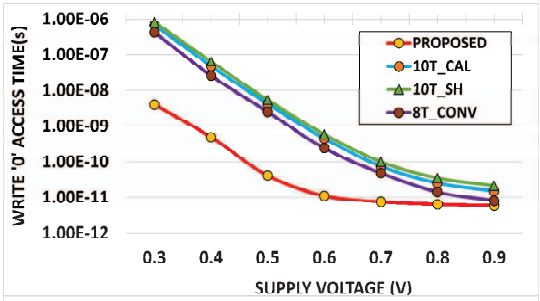
(c)

Fig. 3. Write Margin (WM) at 27 °C (a) Simulated plot for proposed cell (b) WM '1' (c) WM '0'.

2) Access time (T_{write}) [11]: This parameter is the measure of how much faster the cell can overwrite the stored data. T_{write} '1' is defined as the time required for the internal storage node holding '0' value to increase to 0.9 times the supply voltage. Alternatively, T_{write} '0' is defined as the time required for same node to decrease to 0.1 times the supply voltage when holding '1'. The plots in Fig. 4 show the estimated values at different voltages. The proposed cell shows upto 98.6 % and 99.5 % improvement in T_{write} '1' and '0' respectively over other cells at 0.3 V.



(a)



(b)

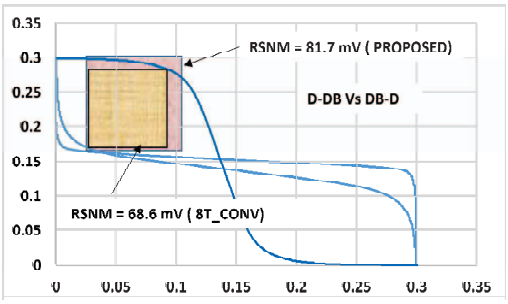
Fig. 4. Access time (T_{write}) at 27 °C (a) T_{write} '1' (b) T_{write} '0'

B. Read-mode

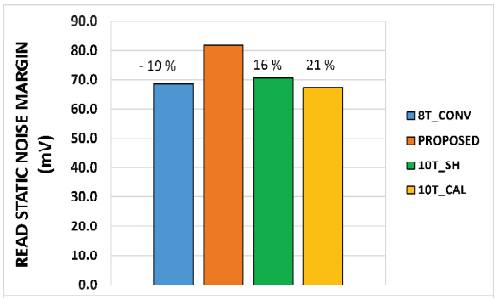
In this mode, the static noise margin is measured to gauge the impact of external interference on stored data. The time to read the stored data is evaluated as read access time. The read bitline voltage swing is captured to measure the ability of the design to correctly sense the stored value as '0' and '1' during read cycle.

1) Static Noise margin (RSNM) [9]: The parameter is measured using butterfly curve method [9] and is plotted in Fig. 5(a) at 0.3 V. It can be observed that the proposed cell shows significant improvement in read stability in comparison to 8T_CONV cell due to more steeper characteristics. The results in Fig. 5(b) shows that the proposed design achieves 19 %, 16 % and 21 % improvement in RSNM value compared to 8T_CONV, 10T_SH and 10T_CAL cells respectively.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:37:37 UTC from IEEE Xplore. Restrictions apply.



(a)



(b)

Fig. 5. Static noise margin (RSNM) (a) Simulated butterfly curves (b) RSNM values at 0.3 V and 27 °C

2) Access Time (T_{read}) [7]: The T_{read} is measured at different supply voltages for all the cells and the values are plotted in Fig. 6. The proposed cell provides significant reduction in T_{read} values irrespective of supply voltage. An improvement of up to 99.9 % is observed compared to other cells in the sub-threshold region.

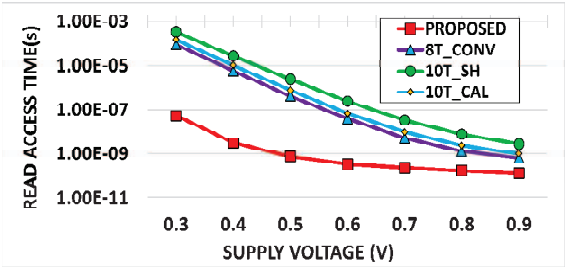


Fig. 6. Access Time (T_{read}) at 27 °C

3) Voltage Swing (ΔV_{RBL}) [7] : The ΔV_{RBL} is measured as the change in RBL voltage when the cell stores '0' and '1' respectively. The normalized values of this parameter for all the cells are captured at different voltages and are plotted in Fig. 7. The results depict that the proposed cell achieves an improvement of 70 %, 49 % and 37 % in comparison to 8T_CONV, 10T_SH and 10T_CAL cells respectively in sub-threshold region.

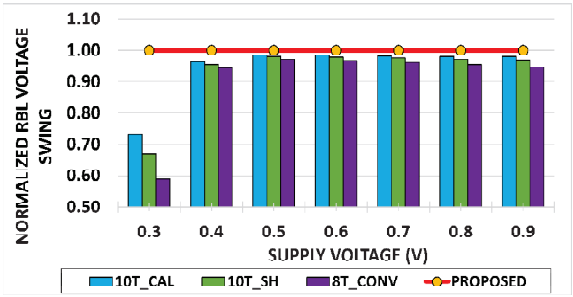


Fig. 7. Normalized Read Bitline Voltage Swing (ΔV_{RBL}) at 27 °C

C. Hold-mode

The key parameters namely static noise margin and leakage power are estimated and compared in this mode.

1) Static Noise margin (HSNM) [12]: The stability of the data at complementary nodes D and DB in the presence of leakages is measured using this parameter by plotting butterfly curves between D-DB and inverted DB-D (Fig. 8). An improvement of 17 %, 14 % and 21 % is seen for proposed cell with respect to 8T_CONV, 10T_SH, 10_CAL cells respectively at 0.3 V indicating improved stability of stored data.

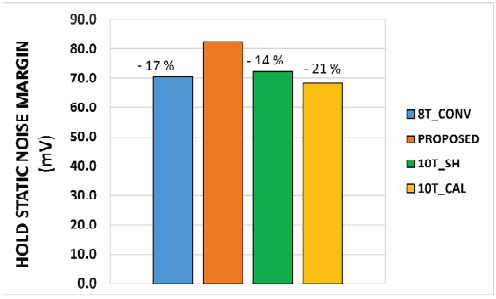


Fig. 8. Static noise margin (HSNM) at 0.3 V and 27 °C.

2) Retention Voltage (DRV) [12]: The data stored in the cell becomes unstable as the cell supply reduces below particular value. So, the supply voltage at which the latch attains metastablility gives the measure of this parameter. The lower value of this parameter helps in reducing the power requirements in hold-mode. The plot in Fig. 9 shows that the proposed cell provides impeccable performace at all voltages. A percentage improvement of 77 %, 82 % and 83 % is observed compared to 8T_CONV, 10T_SH and 10_CAL cells respectively at 0.3 V.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:37:37 UTC from IEEE Xplore. Restrictions apply.

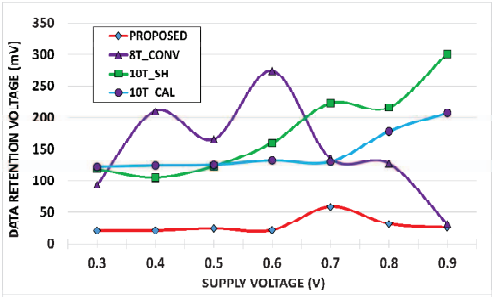


Fig. 9. Retention voltage (DRV) at 27 °C.

3) Leakage Power (P_{leak}) [12]: A plot of P_{leak} at different supply voltages (Fig. 10) depicts that values are comparable for all the cells in sub-threshold region. At higher voltages, the leakage power of proposed cell is more as the write assist transistor (MDL2) remains ON during hold mode and increases the overall leakage current of the latch structure compared to other cells.

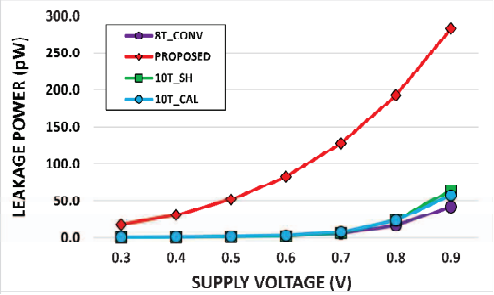


Fig. 10. Leakage power (P_{leak}) at 27 °C.

V. Conclusion

For sub-threshold operation, a 10T FinFET cell with improved write and read performance at 22 nm technology node is presented in this research paper. The proposed cell shows an improvement of up to 322 % and 99.5 % in write margin and access time respectively in write-mode over other cells due to the use of assist transistor which eases the writing process. A 21 %, 99.9 % and 70 % improvement is also observed in static noise margin, access time and voltage swing respectively in read-mode due to high read current values and, low and data-independent read bitline leakages. It also achieves up to 21 % and 83 % improvement in static noise margin and retention voltage respectively in hold-mode.

References

[1] Almeida RB, Marques CM, Butzen PF, Silva FRG, Ricardo R, Meinhardt C, "Analysis of 6 T SRAM cell in sub-45 nm CMOS and FinFET technologies", Microelectronics Reliability, 88-90(6), 196-202, 2018.

[2] Farkhani H, Peiravi A, Kargaard JM, Moradi F, "Comparative study of FinFETs versus 22nm bulk CMOS technologies: SRAM design perspective", 27th IEEE International System-on-Chip Conference (SOCC), 449-454, 2014.

[3] Guduri M, Dokania V, Verma R, Islam A, " Minimum energy solution for ultra-low power applications", Microsyst. Technol., 25(5), 1823–1831, 2019.

[4] Karapetyan S, Schlichtmann U, "20nm FinFET-based SRAM cell: Impact of variability and design choices on performance characteristics", 14th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD), 1-4, 2017.

[5] Calhoun BH, Chandrakasan AP, "A 256-kb 65-nm sub-threshold SRAM design for ultra-low-voltage operation", IEEE J Solid-State Circuits, 42(3), 680-688, 2007.

[6] Gupta S, Gupta K, Calhoun BH, Pandey N, "Low-power near-threshold 10T SRAM bit cells with enhanced data-independent read port leakage for array augmentation in 32-nm CMOS", IEEE Trans Circuits Syst I Regul Pap, 66(3), 978-988, 2018.

[7] Gupta M, Gupta K, Pandey N, " A data-independent 9T SRAM cell with enhanced ION/IOFF ratio and RBL voltage swing in near threshold and sub-threshold region", Int J Circ Theor Appl., 49, 953-969, 2021.

[8] Narendra S, Borkar S, De V, Antoniadis D, Chandrakasan A, "Scaling of stack effect and its application for leakage reduction", Int. Symp. Low Power Electronic and Design, 195-200, Aug. 2001.

[9] Gupta M, Gupta K, Pandey N, "A design of low leakage cache memory cell for high performance processors", Journal of Information and Optimization Sciences, 40(2), 279-290, 2019.

[10] Makino H et al, " Reexamination of SRAM cell write margin definitions in view of predicting the distribution", IEEE Transaction of Circuits and Systems II: Exp. Briefs, 58(4), 230- 234, 2011.

[11] Ahmad S, Gupta M K, Alam N, Hasan M, "Single-Ended Schmitt-Trigger-Based Robust Low-Power SRAM Cell", IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 24(8), 2634-2642, 2016.

[12] Gupta M, Gupta K, Pandey N, "A 32-nm Sub-Threshold 9T SRAM Bitcell with Improved Read and Write performance", International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 781-787, 2018.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:37:37 UTC from IEEE Xplore. Restrictions apply.

2021 Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 28-29, 2021

A Comparison of Similarity Measures for Neighbourhood Based Collaborative Filtering Recommender Systems

Rohit Beniwal
Department of Computer Science and Engineering
Delhi Technological University
New Delhi, India
dr.rohitbeniwal@gmail.com

Rahul Khairwal
Department of Computer Science and Engineering
Delhi Technological University
New Delhi, India
rahulkhairwal@gmail.com

Ritvik Mahajan
Department of Computer Science and Engineering
Delhi Technological University
New Delhi, India
ritvik16@gmail.com

Sarthak Pratap Narayan Singh
Department of Computer Science and Engineering
Delhi Technological University
New Delhi, India
sarthakpratap@gmail.com

2021 Asian Conference on Innovation in Technology (ASIANCON) | 978-1-7281-8402-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ASIANCON51346.2021.9544901

***Abstract*—As data continues to grow rapidly at an exponential rate, the importance of data processing systems to extract information is also increasing continuously. Recommender Systems are information processing software that seek to aid decision making by providing relevant suggestions from a given pool of items. In this domain, earlier studies have presented a comparison of similarity measures mostly using Root Mean Squared Error and Mean Absolute Error and to the best of our knowledge, they lacked analysis based on the computation time. Therefore, in this paper, we present a comparison of three similarity measures, namely Pearson Correlation Similarity, Cosine Vector Similarity, and Mean Square Difference on two different datasets known as MovieLens 100k and MovieLens 1M and compared the results using three accuracy metrics namely, Root Mean Square Error, Mean Absolute Error, and Fraction of Concordant Pairs. Our results show that in terms of accuracy, Pearson Correlation Similarity is the best performing similarity measure. Moreover, we also consider the time of simulations for each measure to add another dimension to performance evaluation. As a result, Mean Squared Difference was found to be the fastest among all the three similarity measures**

Keywords—Recommender System, Collaborative Filtering, Similarity Measures, Neighbourhood-based models

I. INTRODUCTION

With rapid and exponential growth in data in the past few years and expected to increase as time progresses, recommender systems are extensively being employed in a wide variety of applications ranging across a large number of domains. Recommender systems are information processing software tools that seek to aid user decision making by providing relevant suggestions from a given pool of items [1]. Different applications use different types of recommender systems. Bobadilla et al. [2], Yang et al. [3], Katarya and Verma [4], and Zhang et al. [5] explain various methodologies for classifying recommender systems. Yadav and Katarya [6] also provides a diagrammatical visualisation of these categories.

Collaborative Filtering (CF) is one such approach in which the past behaviour of a user or community is used to predict the item in which the target user may be interested. Further, model-based CF recommender systems can be subdivided into two types, namely User-Based Collaborative Filtering (UBCF) and Item-Based Collaborative Filtering (IBCF) systems. Neighbourhood-based models are one of the extensively used algorithms for rating prediction in these types of recommender systems [7].

UBCF algorithms predict ratings for a target user considering the ratings given by other similar users whereas the IBCF algorithms exploit the similarity between items to achieve the same task. Various similarity metrics are used to determine the relationships that exist between users or items. In this research paper, we used both UBCF and IBCF to compare the three-similarity metrics.

The succeeding sections of this paper are organized as follows: Section II discusses the related work; Section III describes the methodology and evaluation metrics used for comparing the similarity measures; Section IV provides the implementation details followed by Section V, which examines the obtained results. Finally, section VI concludes the research paper and provides direction for future work.

II. RELATED WORK

In a study done by Hassanieh et al. [8], seven similarity measures were compared on the MovieLens 100k dataset using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) evaluation metrics. Pearson Correlation Similarity (PCS) was found as the worst performing among all the seven measures. In this study, authors proposed performing these comparisons on larger datasets using more no. of evaluation metrics and analysing time-based performance of these similarity measures as their future scope.

Suganeshwari and Ibrahim [9] examined the performance of five similarity measures using both UBCF and IBCF on

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:38:08 UTC from IEEE Xplore. Restrictions apply.

the MovieLens 100k dataset. Results showed that IBCF was more effective than UBCF.

Cacheda et al. [10] compared PCS, Cosine Vector Similarity (CVS), Mean Square Difference (MSD), and Weighted Pearson Correlation (WPC) along with two novel metrics.

The outcomes showed the shortcomings of numerous measures in extracting information from user profiles particularly when the dataset is sparse.

III. METHODOLOGY

In this section, we provide the basic workflow that we plan to follow in this research paper. A brief outline of the steps to be followed is as shown below by Figure 1.

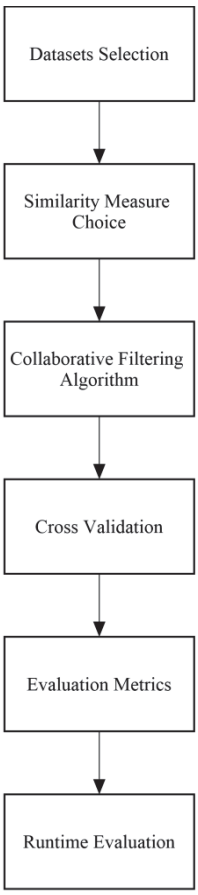


Fig. 1. Workflow Description

A. Dataset Selection

In this subsection, we will choose a suitable dataset in order to perform the similarity measure for neighbourhood based collaborative filtering recommender systems.

B. Similarity Measure Choice

In this subsection, we will examine the three similarity measures that we will use in this research paper. These three similarity measures are PCS, CVS, and MSD.

1) Pearson Correlation Similarity

PCS is a widely used measure for computing similarity. In the case of UBCF, given two users 'x' and 'y' and a set of items I_{xy} which contains the mutually rated items. Pearson correlation is computed as follows [11]

$$\mathrm{PCS}(x,y) = \frac{\sum_{i \in Ixy}(r_{xi}-\overline{r_x})(r_{yi}-\overline{r_y})}{\sqrt{\sum_{i \in Ixy}(r_{xi}-\overline{r_x})^2}\cdot\sqrt{\sum_{i \in Ixy}(r_{yi}-\overline{r_y})^2}} \quad (1)$$

Where r_{xi} signifies the rating given by user 'x' to item 'i'. $\overline{r_x}$ denotes the mean rating given by user 'x'. On similar lines, for IBCF, given two items 'p' and 'q', a user 'x' and a set I_{pq} containing the users who have scored both the items, the Pearson correlation for items 'p' and 'q' is given by

$$\mathrm{PCS}(p,q) = \frac{\sum_{x \in Ipq}(r_{px}-\overline{r_p})(r_{qx}-\overline{r_q})}{\sqrt{\sum_{x \in Ipq}(r_{px}-\overline{r_i})^2}\cdot\sqrt{\sum_{x \in Ipq}(r_{qx}-\overline{r_q})^2}} \quad (2)$$

PCS values range between -1 to +1, where +1 denotes perfect similarity while -1 denotes perfectly opposite user/items or inverse relation between users. A value 0 denotes that there is no relation. A major drawback of PCS is that it returns a high (low) similarity value when there are few mutually rated items, despite the differences in the rating values. [12]

2) Cosine Vector Similarity

Cosine Vector Similarity considers data objects as vectors in n-dimensional space. The similarity index on a scale of 0 to 1 is given by cosine of the angle between these vectors. For a user-based approach, CVS score for two users 'x' and 'y' is given by [11]

$$\mathrm{CVS}(x,y) = \frac{\sum_{i \in Ixy} r_{xi}.r_{yi}}{\sqrt{\sum_{i \in Ixy}(r_{xi})^2}\cdot\sqrt{\sum_{i \in Ixy}(r_{yi})^2}} \quad (3)$$

Similarly, for an item-based approach, CVS between two items 'p' and 'q' is given by

$$CVS(p,q) = \frac{\sum_{x \in Ipq} r_{pu}.r_{qu}}{\sqrt{\sum_{x \in Ipq}(r_{pu})^2}\cdot\sqrt{\sum_{x \in Ipq}(r_{qu})^2}} \quad (4)$$

Terms have the same meaning as given above. Unknown ratings are treated as 0's leading to a CVS value of 0. A value of 0 signifies no relation and a value 1 signifies perfect similarity. CVS suffers from the fact that it fails to account for the differences in variance and mean of the ratings. Also, if the mean user/item rating is subtracted from the user/item ratings, CVS becomes equivalent to PCS. [13]

3) Mean Squared Difference

MSD evaluates the similarity between two users (for UBCF) or items (for IBCF) as the reciprocal of the average (mean) of the squared differences given a set of common rated items. The formula for computing MSD is as follows for UBCF (Equation 5) and IBCF (Equation 6):

$$\mathrm{MSD}(x,y) = \frac{|I_{xy}|}{\sum_{i \in Ixy}(r_{xi}-r_{yi})^2} \quad (5)$$

$$\mathrm{MSD}(p,q) = \frac{|I_{pq}|}{\sum_{u \in Iij}(r_{pu}-r_{qu})^2} \quad (6)$$

MSD is a value greater than 0. A smaller value of MSD indicates a better similarity One limitation of MSD is that it cannot account for negative correlations like PCS, which might improve the overall prediction accuracy.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:38:08 UTC from IEEE Xplore. Restrictions apply.

C. Collaborative Filtering Algorithm

The collaborative filtering algorithm that we plan to use is a k-NN inspired algorithm that accounts for the mean rating given by the user. It takes 'k' as an input value, which specifies the maximum number of neighbours used to compute the predicted rating and the choice of similarity measure used to determine similar users or items. The predicted ratings are given by Equation 7 and 8 for UBCF and IBCF respectively.

$$\widehat{r_{xi}} = \mu_x + \frac{\sum_{y \in N_i^k(x)} sim(x,y).(r_{yi} - \mu_y)}{\sum_{y \in N_i^k(x)} sim(x,y)} \quad (7)$$

$\widehat{r_{xi}}$ refers to the rating predicted by the system that user 'x' is expected to give to item 'i', computed from all the similar users 'y' in the neighbourhood of 'x' as specified by the value of k ($N_i^k(x)$). sim(x,y) refers the similarity score between users 'x' and 'y' as determined by the selected choice of similarity measure. μ_x and μ_y denote the average (or mean) rating values provided by users 'x' and 'y' respectively.

$$\widehat{r_{xp}} = \mu_p + \frac{\sum_{q \in N_x^k(p)} sim(p,q).(r_{xq} - \mu_q)}{\sum_{x \in N_x^k(p)} sim(p,q)} \quad (8)$$

$\widehat{r_{pf}}$ refers to the rating predicted by the system that user 'x' is expected to give to item 'p', computed from all the similar items 'q' in the neighbourhood of 'p' as specified by the value of k ($N_x^k(p)$). sim(p,q) refers to the similarity score between 'p' and 'q' as determined by the selected choice of similarity measure. r_{xq} is the actual rating given by user 'x' to item 'p'.μ_p and μ_q denotes the average score provided for items 'p' and 'q' respectively.

D. Cross Validation

"Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model. In typical cross-validation, the training and validation sets must cross-over in successive rounds such that each data point has a chance of being validated against" [14] . Cross validation has the advantage of being less biased over other simple methods such as a train test split. For our experiments, we will use a k-fold cross validation procedure. As examined in earlier studies, a value of ten for k has shown a low bias and a moderate variance, therefore, we will use a 10-fold cross validation procedure in order to implement our methodology.

E. Evalution Metrics

In this subsection, in order to compare the similarity metrics as discussed above, we will use three evaluation metrics as described below.

1) Root Mean Squared Error

In context to recommender systems, RMSE measures the square root of the average (mean) of squared differences between the rating values predicted by the system and the actual rating value given by the user. RMSE is always a positive value. Given two RMSE values, the lower value indicates better system performance. The formula for RMSE is

$$\text{RMSE} = \frac{1}{|\hat{R}|}\sqrt{\sum_{\widehat{r_{xi}} \in \hat{R}} (r_{xi} - \widehat{r_{xi}})} \quad (9)$$

2) Mean Absolute Error

MAE gages the absolute difference of the system's predicted rating value from the actual user rating value. Like RMSE, a lower value of MAE signifies enhanced performance. MAE is computed by the following formula

$$MAE = \frac{1}{|\hat{R}|}\sum_{\widehat{r_{xi}} \in \hat{R}} |r_{xi} - \widehat{r_{xi}}| \quad (10)$$

3) Fraction of Concordant Pairs

Given a test set say X, the number of concordant pairs for a given system user 'x' is calculated by counting those which are ranked correctly by the prediction algorithm. The mathematical definition is given as [15]

$$n_c^x = |\{p, q)| \left(\widehat{r_{xp}} > \widehat{r_{xq}} \text{ and } r_{xp} > r_{xq}\right) \text{ or } \left(\widehat{r_{xp}} < \widehat{r_{xq}} \text{ and } r_{xp} < r_{xq}\right)| \quad (11)$$

where p and q are two different items, 'r' represents predicted ratings and 'r' represents actual ratings. In a similar way the number of dis-concordant pairs, n_d^x can be calculated for user x. The formula for computing FCP is [15]

$$FCP = \frac{\sum_{x \in U} n_c^x}{\sum_{x \in U} n_c^x + \sum_{x \in U} n_d^x} = \frac{n_c}{n_c + n_d} \quad (12)$$

where U denotes the set all the users

F. Runtime Evaluation

The program runtimes will be computed for all the different collaborative filtering algorithms discussed above in order to add another dimension to performance evaluation.

IV. Implementation

In this section, we present the implementation details of the proposed methodology including the various tools and technologies used. In order to implement the proposed methodology, we used the Python programming language.

A. Dataset Selection

In order to perform the similarity measure for neighbourhood based collaborative filtering recommender systems, we chose the MovieLens 100k (ML-100K) and MovieLens 1M (ML-1M) data sets. "MovieLens data sets are compiled by the Group Lens Research Project at the University of Minnesota." [16]. Each data set consists of the rating values given by users to a movie on a specified rating scale along with simple demographic information of the user. The following table 1 provides a comparison of properties between the two datasets.

TABLE I. MovieLens 100K vs MovieLens 1M

Dataset/Property	ML - 100K	ML - 1M
Total Ratings	100,000	1,000,209
User Count	943	6040
Movie (Item) Count	1682	3706
Genres	19	18
Rating Scales	1-5	1-5
Sparsity Value	93.70	95.53

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:38:08 UTC from IEEE Xplore. Restrictions apply.

B. *Choice of Similarity*

The choice of similarity measure as provided by the similarities module in the 'surprise' package was set using the 'sim_options' parameter.

C. *Collaborative Filtering Algorithm*

KnnWithMeans class from the 'surprise' package was used to implement the prediction algorithm.

D. *Cross Validation*

To evaluate the similarity measures, we used the 10-fold cross validation procedure. Further, it was used for reporting accuracy and computation times.

E. *Evaluation Metrics*

The three accuracy evaluation metrics namely, MAE, RMSE, and FCP as discussed above were passed as inputs to the 'cross_validate' method.

F. *Runtime Evaluation*

Here, Table 2 summarizes the system specifications, which we used to perform for all the simulations.

TABLE II. IMPLEMENTATION ENVIRONMENT DETAILS

S.No.	Property	Specification
1	Processor and Processing Power	Intel Core i5 4440, 2.7GHz quad-core
2	Operating System	macOS High Sierra, version 10.13.6
3	RAM	8GB (4GB * 2) at 1600MHz DDR3 memory
4	Programming Language	Python 3.8
5	Programming Platform	Jupyter Notebook

V. RESULTS AND ANALYSIS

In this section, we contrast the three similarity measures implemented on ML-100K and ML-100M datasets for both UBCF and IBCF using the three accuracy evaluation metrics discussed previously. We also provide data regarding the time taken for each simulation along with the accuracy metric results

The value of all accuracy metrics has been obtained by averaging values for 50 different values of 'k' parameter ranging between 2 and 100 (both inclusive) in steps of 2. A general shape is shown by **Error! Reference source not found.**

Figure 3 as shown below describes the average RMSE for all the similarity measures implemented on both the datasets, using both UBCF and IBCF approaches.

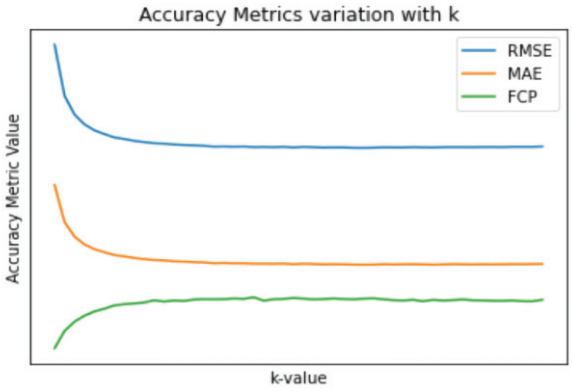


Fig. 2. Accuracy Metrics Variation with k

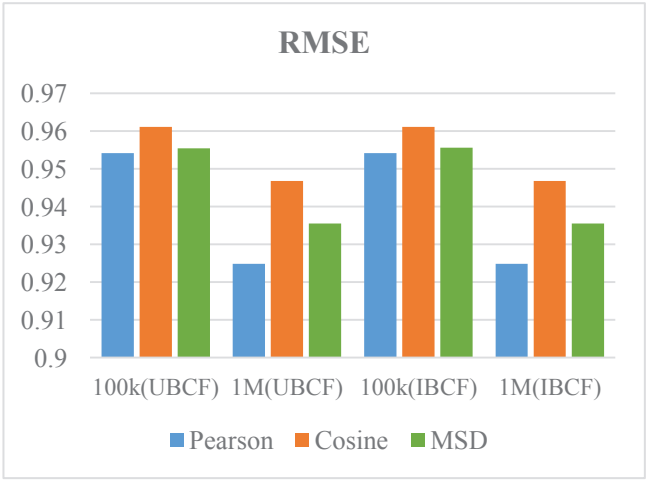


Fig. 3. Average RMSE Values

From the above figure, the following three conclusions can be drawn. First, all similarity measures give almost the same prediction accuracy for UBCF and IBCF approaches (as values are the same up to the 4th decimal point). Second, the accuracy of all similarities is more when implemented on the ML-1M dataset as compared to the ML-100K dataset, which is shown by a lesser RMSE. Third, in all cases, the order of performance, indicated by low RMSE, is PCS > MSD > CVS.

TABLE III: PYTHON MODULE DESCRIPTIONS

S.No.	Library/Module	Description	Purpose
A	pandas	Python Library for data analysis and manipulation	To read the dataset and create a data frame for further analysis
B	surprise	Python Scikit Library for designing, creating, evaluating and analysing recommender systems	To implement a) Similarity measures b) Rating prediction algorithms c) Cross validation Procedures d) Accuracy Metrics
C	matplotlib	Python Plotting Library	Result visualization and analysis
D	NumPy	Python Library for Numerical Computations	Mathematical computations and array handling
E	time	Python Module to manage time related tasks	To compute run time of the simulations

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:38:08 UTC from IEEE Xplore. Restrictions apply.

The same pattern is shown when the similarity measures are compared by using MAE as shown below by Figure 4.

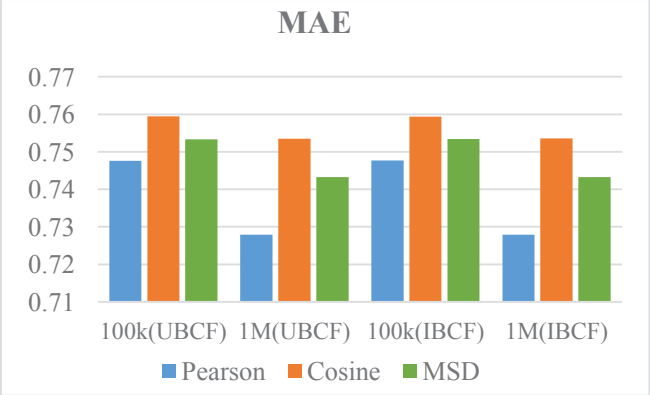


Fig. 4. Average MAE Values

In the above figure, MAE values are lower than RMSE only because RMSE tends to strongly penalize large errors and do not indicate better performance.

FCP values are shown below by following Figure 5.

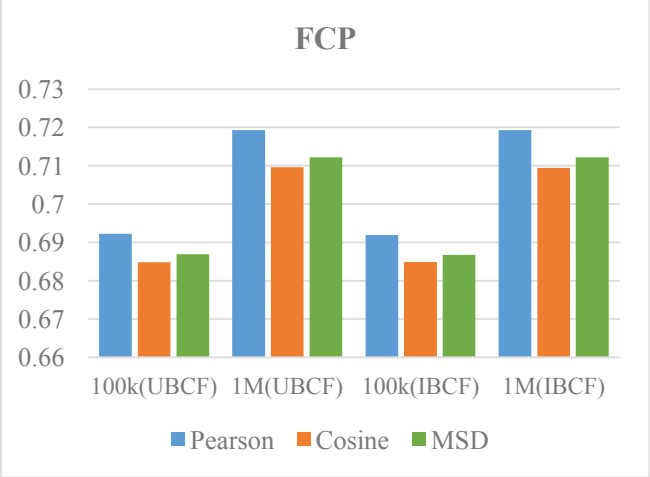


Fig. 5. Average FCP Values

FCP shows a similar performance trend like RMSE and MAE as higher FCP values indicate a better performance.

Table 4 shows the run time of all different simulations rounded off to the nearest second.

TABLE IV: RUN TIMES FOR SIMULATIONS

Similarity Measure	MovieLens – 100 k		MovieLens - 1M	
	UBCF	IBCF	UBCF	IBCF
PCS	6208	5415	93376	128698
CVS	5747	3386	84484	117306
MSD	4168	2351	67261	98051

From Table 4, it can be seen in terms of execution time that MSD is the fastest one while PCS is the slowest one, and CVS comes in between the two for all cases. Also, it can be concluded that for ML-100k dataset, IBCF is a way faster than UBCF while the trend is reversed on the ML-1M dataset.

VI. CONCLUSION AND FUTURE SCOPE

In this research paper, we presented a comparison of three similarity measures, namely PCS, CVS, and MSD on two different datasets known as MovieLens 100k and MovieLens 1M and compared the results via UBCF and IBCF algorithms using three accuracy metrics namely, RMSE, MAE, and FCP. Our results have shown that PCS has the best prediction and ranking accuracy, but took the maximum time out of all-similarity measures. MSD has the least time for computation, but provided the second-best accuracy. CVS has the least accuracy of predictions among all similarity measures. The accuracy of predictions for both UBCF and IBCF algorithms for the respective similarity measures was very close.

In the future, a similar study can be conducted on a larger dataset (ML-10M, ML-20M or ML-25M) so as to get more precise results. More similarity measures and evaluation metrics can also be considered.

REFERENCES

[1] F. Ricci, L. Rokach and B. Shapira, Introduction to Recommender Systems Handbook, Springer, 2011, pp. 1-35.

[2] J. Bobadilla, F. Ortega, A. Hernando and A. Gutiérrez, "Recommendor Systems Survey," in Knowledge-Based Systems, vol. 46, 2013, pp. 109-132.

[3] X. Yang, Y. Guo, Y. Liu and H. Steck, "A survey of collaborative filtering based social recommender systems," in Comput. Comun., vol. 41, 2014, pp. 1-10.

[4] R. Katarrya and O. Verma, "Recommender system with grey wolf optimizer and FCM," in Nueral Compt. Appl., 2016, pp. 1-9.

[5] F. Zhang, T. Gong, V. E. Lee, G. Zhao, C. Rong and G. Qu, "Fast algorithms to evaluate collaborative filtering recommender systems," in Knowledge-Based Systems, vol. 96, 2015, pp. 96-103.

[6] D. K. Yadav and R. Katarya, "Study on Recommender System using Fuzzy Logic," in Second International Conference on Computing Methodologies and Communication, 2018.

[7] M. Elahi, F. Ricci and N. Rubens, "A Survey Of Active Learning In Collaborative Filtering Recommender Systems," in Computer Science Review, vol. 20, Elsevier, 2016, pp. 29-50.

[8] L. A. Hassanieh, C. A. Jaoudeh, J. B. Abdo and J. Demerjian, "Similarity Measures for Collaborative Filtering Recommender Systems," in Middle East and North Africa Communications Conference(MENACOMM), Jounieh, 2018.

[9] G. Suganeshwari and S. S. Ibrahim, "A Comparison Study on Similarity Measures in Collaborative Filtering Algorithms for Movie Recommendation," International Journal of Pure and Applied Mathematics, vol. 119, pp. 1495-`1505, 2018.

[10] F. Cacheda, V. Carneiro, D. Fernández and V. Formoso, "Comparision of collaborative filtering algorithms: limitations of current techniques and proposals for scalable, high-performance recommender system," ACM Trans. Web, pp. 1-33, 2011.

[11] C. Desrosiers and G. Karypis, "A Comprehensive Survey Of Neighborhood-Based Recommendation Methods," in Recommender Systems Handbook, Springer, 2011, pp. 107-144.

[12] Idrissi and Z. N., "A systematic literature review of sparsity issues in recommender systems," Social Network Analysis and Mining, vol. 10, no. 15, 11 February 2020.

[13] M. D. Ekstrand, J. T. Riedl and J. A. Konstan, "Collaborative Filtering Recommender Systems," Journal Foundations and Trends in Human-Computer Interaction, vol. 4, no. 2, pp. 81-173, February 2011.

[14] P. Refaeilzadeh, L. Tang and H. Liu, "Cross-Validation," in Encyclopedia of Database Systems, L. LIU and M. T. ÖZSU, Eds., Springer, Boston, MA, 2009.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:38:08 UTC from IEEE Xplore. Restrictions apply.

[15] Y. Koren and J. Sill, "Collaborative Filtering on Ordinal User Feedback," in International Joint Conference on Artificial Intelligence, Beijing , 2013.

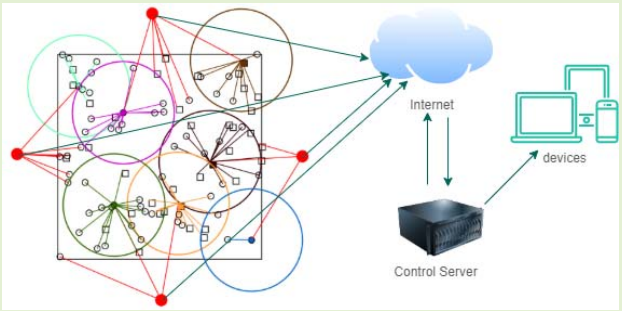
[16] F. M. Harper and J. A. Konstan, "The MovieLens Datasets: History and Context," Transactions on Interactive Intelligent (Tiis), p. 19, December 2015.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:38:08 UTC from IEEE Xplore. Restrictions apply.

A Green Data Collection & Transmission Method for IoT-Based WSN in Disaster Management

Aridaman Singh Nandan, Samayveer Singh, *Member, IEEE*, Aruna Malik, *Member, IEEE*, and Rajeev Kumar

***Abstract*—Being as an integral part of data exchange in disaster management, Internet of Thing (IoT) is an essential component to update the disaster management information to all the connected nodes. In such scenarios, the nodes are deployed with a limited battery, batteries of these nodes are non-rechargeable and such networks consume huge energy in data exchange. Therefore, the practical implementation of such type of networks is a difficult task as routing avoids the paths and consume a tremendous amount of energy during collection/transmission of data. This paper proposes an optimized Genetic Algorithm (GA) based green data collection/transmission method for IoT based WSN in disaster management by satisfying multiple constraints i.e., optimizing intra-cluster distance, systematic utilization of node's energy in the cluster and reducing hop count. The proposed direct data collection/transmission and movable sink strategies shorten the communication distance between the sink and cluster head (CH) which diminishes the hotspot problem. The direct data collection helps in transmitting data directly to the sink when the sinks are nearer to the sensor nodes with respect to CH. Further, the incorporated dynamic sensing range minimizes overlapping of sensing range of CH along with a significant decrement in the transmission energy. The simulation results show that the proposed protocol outperforms the existing protocols on the performance metrics namely network's remaining energy, lifetime, stability period, and throughput per rounds.**



***Index Terms*—Internet of Things, sensor nodes, clustering, direct data collection, adjustable sensing range, movable sinks.**

I. Introduction

NOW a days worldwide, various terrible natural and man-made disasters such as hurricanes, earthquakes, floods, nuclear explosions and transportation accidents, are hazardous to the human being life and others. [1] These natural and man-made disasters are not only damage the real-estate infrastructure like roads, buildings, etc. but also heavily damage the backbone of ICT infrastructure which leads to breakdown in regular communication modes. Moreover, diagnosing and recovery of the ICT infrastructure requires a lot of time and efforts which may halt/delay rescue operations [1]. Hence, the need of on spot random deployable and automatic configurable disaster response rescue system for making connectivity and communication possible has been on the forefront. The automatic rescue system can be deployed using IoTs based wireless sensor network (WSN) in an effective and scalable manner for handle such disasters in real time.

The IoTs based WSNs are a group of sensor nodes (SNs) that possess data collection/transmission, processing and communication capabilities and send it to the disaster management server with the help of the sinks. The sinks are the radio receiver/transmitter that serves as the data collector from all the connected SNs and provide an interface between the end user and network. The sinks can be fixed or movable in nature. The SNs are the core components of the IoT based WSNs and performance depends greatly on the ability of SNs. However, there are some limitations associated with the SNs such as a limited battery, memory, short transmission range, etc. Almost every task including data collection, processing and transmission result in energy drainage and it makes the whole network vulnerable to shut down due to energy breakdown [3]. The network becomes non-operational and communication links are broken which results in network fail. The replacement of batteries in hostile environments is not a viable option [4]. The optimization of the energy is the NP-hard problem because of the complex structure of the IoTs based WSNs. In disaster management, these systems,

Manuscript received August 29, 2021; revised October 3, 2021; accepted October 3, 2021. Date of publication October 4, 2021; date of current version November 12, 2021. The associate editor coordinating the review of this article and approving it for publication was Prof. Dongsoo Har. *(Corresponding author: Rajeev Kumar.)*

Aridaman Singh Nandan, Samayveer Singh, and Aruna Malik are with the Department of Computer Science and Engineering, Dr. B. R. Ambedkar National Institute of Technology Jalandhar, Jalandhar, Punjab 144011, India.

Rajeev Kumar is with the Department of Computer Science and Engineering, Delhi Technological University, Delhi 110042, India (e-mail: rajeevkumar@dtu.ac.in).

Digital Object Identifier 10.1109/JSEN.2021.3117995

1558-1748 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:42:59 UTC from IEEE Xplore. Restrictions apply.

sources SNs communicate the data with multiple destinations SNs/sinks or vice versa. Therefore, the fundamental goal of researchers in this field is directed to increase the longevity of such a network by satisfying the quality of service (QoS) constraints. Many resources used by the traditional methods to designed the solution of such type of networks.

When the IoTs based sensor networks are deployed for the data collection/transmission, the survival period of such networks is the main concern due to inefficient clustering and data collection process, much overhead, poor routing mechanism, in-scalability, long haul transmission and complex structure. Therefore, the optimized cluster head (CH) election decides the fate of the networks that deliver the disaster information to the end-users on an emergency basis. To overcome the above-discussed issues, an optimized genetic algorithm (OptimizedGA) based on direct data collection (DDC) and adjustable sensing range (ASR) using static and dynamic sinks in IoT based WSNs is proposed to maximizing the lifetime. The significant contribution of the proposed work can be summarized as follows:

- In the proposed work, a sensing system is introduced that deliberates the clustering of SNs in the disaster management scenarios. The CH election approach using GA is proposed by combining parameters of energy, density, distance and comparative bias for its fitness function which favours high energy CHs and helps in avoiding in-stability problem of the network.
- The proposed work further introduces the concept of DDC with clustering which allows the nodes (that are nearer to sink) to send their data directly to sinks instead of creating CH and send their data to CHs.
- In this work, both static and moveable sinks are deployed inside and outside of the network to diminish the hot-spot problem.
- Furthermore, the proposed work incorporates the idea of adjustable sensing range (ASR) to avoid overlapping coverage of CH. It is beneficial to minimize the overlapping sensing area of SNs to reduce energy consumption.
- The proposed work manages node and link failure by automatically performing re-clustering as soon as any such failure detected.

Rest of the paper is organized as follows: Section 2 discusses the literature review. The system model is discussed in Section 3. Section 4 discusses the proposed strategies. The simulation results and discusses are deliberated in Section 5. Finally, the paper is concluded in Section 6 with its future work.

II. Literature Review

In this section, various GA based clustering protocols are discussed by considering the effectiveness of energy-efficient protocols for disaster management scenarios. Hamidouche *et al.* discuss a GA based method for improving the lifetime and QoS of WSNs [4]. This method prolongs the lifetime and increases the QoS by using multiple fitness parameters like node density, remaining energy, and distance. This work writhed from the hotshot problem due to inefficient data collection process. Jha *et al.* discuss an energy optimization method by using GA [5]. This method uses the energy for calculating the fitness function. The random selection of CHs makes this method inefficient. Al-Shalabi *et al.* discuss an energy-efficient multi-hop path using an enhanced GA [6]. This paper chooses an optimal multi-hop path CHs and sinks to reduce power consumption. A fitness function is developed using GA which selects the optimum CHs and also optimizes the path for data collection. The proposed protocol contains too much overhead. Bhola *et al.* propose an extension of LEACH protocol using GA for energy-efficient WSNs [7]. This work considers the remaining energy of nodes to develop fitness function which helps in find optimal route. It suffers the long haul transmission in the data collection process. Rani *et al.* discuss the dynamic clustering approach based on GA for IoT applications [8]. This method used a dynamic methodology for clustering and relay nodes for a cluster. Ragavan and Ramasamy [9] utilize GA based scheme for dynamic deployment and applied Tabu search to come up with the optimal route. It endures hot spot problems and instability of the network.

Pal at al. propose an energy-efficient clustering technique using GA [10]. This method modifies the steady-state phase of the LEACH protocol by using three parameters such as compactness, separation, and number of CHs. The CHs election is inefficient due to the network remaining energy is not conveyed. Nandan *et al.* discuss an efficient cluster head election based on optimized genetic algorithm for movable sinks [11]. These methods select CH by combining distance, energy with the node density parameter. This method gives the poor routing mechanism which hamper the network lifetime. Sahoo *et al.* discuss a particle swarm optimization (PSO) based clustering with sink mobility in heterogeneous WSN [12]. This method deliberates the lower energy nodes and shortest energy nodes in clustering. Sahoo *et al.* discuss a hybrid approach for optimizing clusters based in WSNs called GAPSO-H [13], an extension of [12]. This method uses energy consumption rate in addition to distance, node degree and energy.

Dehghani *et al.* discuss an enhanced energy-aware cluster-based routing algorithm by using GA [14]. It suffers complex topology, more overhead, and hot spot problem of the networks. Wang *et al.* discuss GA for energy-efficient clustering and routing in WSNs [15]. This method considered the fitness function using total energy consumption which improves energy efficiency. The fitness function also considered the load balancing among the nodes. It is not performed well due to the limitations of the fitness parameter. Chauhan *et al.* discuss an evolutionary algorithm for CH selection in heterogeneous WSNs [16]. However, remaining energy and distance are only considered in clustering process.

Ramesh at al. discuss an application based systems for remote health monitoring and management using IoTs sensors [17]. However, these methods are not able to maintain the tradeoff between load balancing and network reliability during mobility. Furthermore, there are many routing and clustering approaches [18]–[21] discuss for clustering and data collection but most of the approaches suffer from the load balancing for energy efficient data delivery. Verma *et al.* discuss a GA based optimized CH selection for single and

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:42:59 UTC from IEEE Xplore. Restrictions apply.

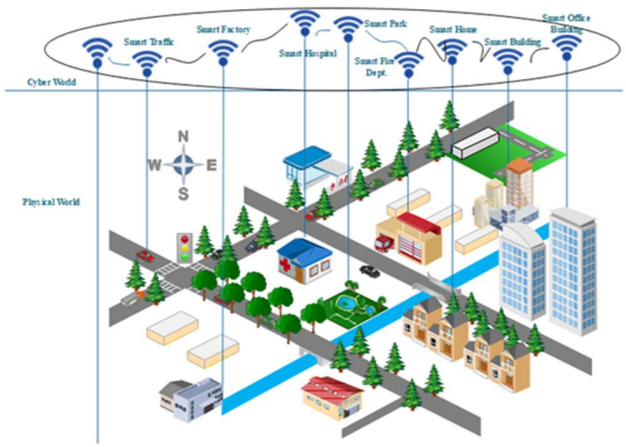


Fig. 1. IoTs-based WSNs scenario for disaster management.

multiple data sink in heterogeneous WSNs [22]. This method uses various parameters such as distance, remaining energy and node density for fitness function along with single and multiple data sinks. It is suffering the complex topology and upholds the too much overhead. Based on the above study, it is observed that the metaheuristic algorithms provide sustainable solutions and also gives an opportunity of minimizing energy consumption and computation time.

III. System Model

The routing problem is formulated for the disaster management scenario by considering smart sensors for mapping the objects of the cyber and physical world as shown in Fig. 1. The sinks are directly connected to the control server which collects/transfer the data among the connected sensors then the information forwarded to the end user to provide the required service. An example is considered for the fire scenario in which sensors are placed at various places (such as smart traffic, smart factory, smart hospital, smart homes and building) of the city. The decision of collecting and transmitting the rescue information is initiated by the control server and forwarded through the intermediate nodes to the end users. The proposed IoT based WSNs scenario is considered for two type of network environments with a predefined amount of energy and nodes called 1 and 2-stair heterogeneity. The networks consisting of n number of SNs.

1-stair heterogeneity (Str-1-Het): It contains only a single type of nodes in the network i.e., normal nodes. Therefore, all the nodes in Str-1-Het having an equal amount of initial energy. The Str-1-Het contains n nodes and E_{init} will be the energy of individual nodes. Therefore, the total network energy is $E_{Total} = E_{init} * n$.

2-stair heterogeneity (Str-2-Het): It consists of two types of energy nodes in the network i.e., intermediate and normal nodes. The normal and intermediate nodes are equipped with E_{init} and $E_{init} * (1 + \alpha)$ initial energy in the network filed. The intermediate nodes have α more energy than the normal nodes which are equipped with $\mathfrak{p}_1$ fraction. Let $\mathfrak{p}_1$ be the proportion of normal nodes ($n_{\text{nrm}} = n*(1-\mathfrak{p}_1)$ and intermediate ($n_{\text{int}} = n * \mathfrak{p}_1$) nodes. Thus, the total network energy of Str-2-Het network is $E_{init} * n * (1 + \mathfrak{p}_1 * \alpha)$.

A radio energy consumption model consists of a sensor system, signal processing systems, analog to digital converters and a radio transceiver. The radio transceiver consumes more energy in communication than that of the other processes. The energy consumed by a SN to transfer a packet of x bits at a distance d is given by $E_{tran}(x, d)$ as follows:

$$E_{tran}(x,d) = \begin{cases} x * (E_{ele} + E_{efs} * d^2) & if\ d \le d_o \\ x * (E_{ele} + E_{amp} * d^4) & if\ d > d_o \end{cases} \quad (1)$$

where d is the distance between the sending and receiving nodes. The transmitter amplifier characteristics are given by E_{ele}, E_{efs} and E_{amp} which are representing the energy dissipated by the transceiver circuitry during communication, energy consumption by the amplifier in free space and multipath propagation model, respectively. Energy consumed in data receiving is given by $E_{rec}(x)$:

$$E_{rec}(x) = x * E_{ele} \quad (2)$$

Data aggregation at CH consumes energy given by $E_{ag}(x)$:

$$E_{ag}(x) = w * x * E_{da} \quad (3)$$

where E_{da} is the energy required to aggregate 1-bit data and w is the number of data packets. Energy expenditure during aggregation is denoted by $E_{ag}(x)$ for l bit packets.

IV. Proposed Work

This section explains the optimized GA for CH election, direct data collection (DDC) method, ASR strategy and stable and movable multiple sink strategies.

A. Direct Data Collection (DDC) Method

In clustering, it is absolutely unnecessary to shape CH near the sink. It prompts unfortunate giving off of data first to CH and thereafter to the BS, while, the SN itself is adequately near clearly speak with the sink and send appropriate data packets to it. DDC by sinks reduce bothersome handing-off of data and movement of data bundles over a progressively stretched out distance to show up at the sink and hence, by and large, improve the lifetime of the sensor arrangement. The multi-hop communication consuming framework's energy near the sink is cleared by using the thought of DDC near sink. Thus, the clustering methodology settles on making CH's, yet moreover take care that no CH is characterized in areas closer to sink for propelling DDC, as such isolating nodes near sink as autonomous nodes for transmission. SNs near sink up-to a predefined range/distance are equipped for DDC process. Other than independent nodes, nodes which are deficiently located in some sparse territory can directly send data to the nearest sink. The SNs arranged closer to sink clearly transmit to the sink. However, this doesn't infer that they can never take part in the clustering process, or can turn into common or CHs. Over the network's lifetime, the amount of active nodes exhaust and a period may come when a sensible productive plan is to make groups near the sink. By then, the computation will take that course and make clusters near the sink as well.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:42:59 UTC from IEEE Xplore. Restrictions apply.

The nodes are adequately near sink to be seen as independent nodes anyway are still picked as CH. Hence, this thought of DDC diminishes unwanted data bundles movement.

$$Dn_i = \begin{cases} Independent\ SN, & \text{if} \sqrt{\left|x_{Dn_i} - x_s\right|^2 + \left|y_{Dn_i} - y_s\right|^2} < D^s \\ Common\ Node, & otherwise \end{cases} \tag{4}$$

where x_{Dn_i} and y_{Dn_i} designate the coordinates of SNs, and x_s & y_s are the coordinate position of the closest available sink for transmission. For i^{th} node to fit the bill for DDC, it should lie in a specific separation range regarding the nearest sink. D^s portrays the base separation needed from the nearest sink to straightforwardly speak with it without consideration of any CHs.

B. Adjustable Sensing Range (ASR) Method

The concept of adjustable sensing range describes a novel structure for growing lifetime of WSNs. The positions of SNs with fixed sensing and communication ranges achieves extremely high dense cluster groups containing SNs which are similarly in locale of some other dense cluster. To slash down gross energy use, it is useful to confine the overlapping sensing zone of the SNs. The sufficiency of the strategies lies in modifying the sensing extent of the CHs reliant on their position compare to sinks. This facilitates the nodes arranged near BS to communicate directly to sink rather than being a node of some cluster. The adaptable detecting range handles the hot-spot issue and diminishes energy usage of SNs. In this paper, we use the sensing detecting extent of the CHs rather than the fixed sensing range. The ASR_{ch} i.e. adjustable sensing range of an SN is calculated as follow-

$$ASR_{ch} = \left(1 - \left(\frac{\mathfrak{D}_{max} - \mathfrak{D}(BS, SN_i)}{\mathfrak{D}_{max} - \mathfrak{D}_{min}}\right)\right) * R_{max} \tag{5}$$

where R_{max} is a node's maximum sensing reach, $\mathfrak{D}_{min}$ and $\mathfrak{D}_{max}$ are the min and max distance from BS, $\mathfrak{D}(BS, SN_i)$ mean the separation between base station and i'th sensor hub.

C. Stable and Movable Multiple Sink Strategies

The mobile sinks in IoT based WSNs can improve sensor network execution, for example, expanding energy effectiveness and throughput. The total roundabout sink movement methodology is proposed for movable sinks for collecting information from network nodes to use the versatility. The arbitrary sink movement isn't appropriate for data assimilation consistently because in most pessimistic scenario, it might put sinks intently moving or perhaps in a region where nodes are meager. In the proposed situation, sinks move in a predefined coordinate, without knowing topological structure of deployed SNs. Generally, we accept that SNs know the topographical area of moving sinks by using some service. The sinks can move outside of the system territory in a circular manner. The centre point of this virtual circle is assumed to be the middle point $(\mathfrak{C}_X, \mathfrak{C}_Y)$ of the deployed network. The radius $\mathfrak{R}$ of the circle is kept adequate to keep the sinks outside the network region. In the work, the sinks setup coordinate and produce their movement positions alongside their underlying calculated position are given as follow:

$$x = \mathfrak{R} * \cos(\Theta) + \mathfrak{C}_X \tag{6}$$

$$y = \mathfrak{R} * \sin(\Theta) + \mathfrak{C}_Y \tag{7}$$

where scopes of Θ are $0 \le \Theta \le \pi$, $0 \le \Theta \le \pi$, and $0 \le \Theta \le \pi/2$ for complete roundabout 180° movement and single quadrant rotation, separately. As the operation continues, sinks make a move to their next position, henceforth gathering data from different locales of the deployed network. All sinks can either move in clock-wise or anti-clockwise direction.

D. Fitness Parameters for Cluster Formation

There are four fitness assessment boundaries are investigated for CH election as given below:

1) Node's Energy Factor: A CH dissipates considerably more energy conversely with normal nodes and it loses energy at much snappier rate as different SNs are transmitting data to it for extra dealing. The node's remaining energy stock transforms into a noteworthy unit for consolidation in fitness work specification and energy boundary is given in (8) as follow:

$$\mathfrak{f}_1 = \sum_{s=1}^{n} \varepsilon_{residual}(s) \big/ \varepsilon_{max}(s) \tag{8}$$

where $\varepsilon_{residual}$ indicates the remaining energy of s^{th} hub and ε_{max} presents introductory most extreme energy of node when network started its operation. It is obvious that higher estimation of $\mathfrak{f}_1$ add to an unrivaled fitness regard.

2) Node-Sink Distance: More energy is consumed if the transmission distance between the sending and receiving node is higher and the opposite way around. This ensures CH has the least correspondence transmission distance with its closest available sink, thusly, encouraging energy conservation. The communication distance optimization fitness boundary is given as follow:

$$\mathfrak{f}_2 = \sum_{q=1}^{N} \left(\frac{\mathfrak{D}_{(n-S)}(q)}{\mathfrak{D}_{F(n-S)}} + \frac{n}{\sum_{s=1}^{N} \mathfrak{D}_{(n-S)}(s)}\right) \tag{9}$$

In eq. (9), Euclidean distance between t^{th} hub and closest sink is introduced by $\mathfrak{D}_{(n-S)}$ however, a separation between farthest hub and sink is spoken by $\mathfrak{D}_{F(n-S)}$. It is clear that more the estimation of $\mathfrak{f}_2$ more it favors decision of a node as CH which develop least correspondence separation among CH and its relating sink.

3) Node Density: The higher number of nodes in the oversight of CH helps in ensuring and flexibly keeping up the rest of the energy of CNs which other-wise would be swallowed comprehensively. Picked CH is surrounded by various nodes for instance has high cluster node density around its incorporated territory. This fitness boundary related to density of clusters is portrayed as follow:

$$\mathfrak{f}_3 = \left(\frac{\sum_{q=1}^{n_C} \mathfrak{D}_{(n-NN)}(q)}{n_C} * \frac{1}{\mathfrak{D}_{(n-FN)}(q)}\right) \tag{10}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:42:59 UTC from IEEE Xplore. Restrictions apply.

In eq. (10) q runs from 1 to n_C for example number of nodes in corresponding cluster. $\mathfrak{D}_{(n-NN)}(q)$ forms the distance between q^{th} SN and its neighboring nodes in cluster. $\mathfrak{D}_{(n-FN)}(q)$ presents the Euclidean distance between q^{th} hub and farthest SN inside same cluster.

4) Comparative Bias: This fitness factor doesn't dismiss this individual capacity of SN and is assessed only a solitary time during the introduction stage. $\mathfrak{f}_4$ is deciphered as follows:

$$\mathfrak{f}_4 = \begin{cases} 1 - \mathfrak{n}_{int}/n & for\ intermediate\ node \\ 1 - \mathfrak{n}_{nrm}/n & for\ normal\ node \end{cases} \quad (11)$$

$\mathfrak{f}_4$ favors the assurance of high energy nodes for CH as this factor is generally related to the proportion of energy provided in SNs toward the beginning stage.

The fitness work holds order over the structure as a single figure of authenticity as how close the given cluster association structure is to achieve the predefined aim of minimization. As far as GA fitness function, the framework execution metric of proposed work is given by eq. (12).

$$\mathrm{F} = 1/\mathfrak{f}_1 + \mathfrak{f}_2 + \mathfrak{f}_3 + \mathfrak{f}_4 \quad (12)$$

The proposed GA for glancing through possibly conceivably cluster structure is composed to minimize fitness esteem F over the evolutionary technique of system chromosomes centering for perfect framework execution.

E. Proposed Technique for Cluster Formation

GA is a heuristic methodology that follows a redundant procedure for looking optimized solutions for procuring potential system framework execution by recreating strategy of natural evolution. The growing new candidates are ideal in quality over its past one. As chromosomes hold CH information of system, direct copying of these chromosomes to next generation won't produce any improvement in structure. Therefore, for significant transformative advancement of ideal clusters certain hereditary activities are required. This procedure ensures that best cluster advancement is passed from one generation to straightaway, subsequently, tweaking it in accordance by network state.

1) Encoding Information: Genotype & Phenotype: The network structure data held by chromosome is called genotype. Genotype encoding is basically a correspondence array of proportional size to that of number of nodes in network with each value either being 0 or 1. The value estimation of 0 infers that its relating network index SN is a cluster node, estimation of 1 corresponds SN as CH. Consequently, chromosome separates into a twofold exhibit of size equivalent to add up to sensors in arrange encoding common nodes as 0 and CH as 1. The veritable hereditary data is genotype while, the portrayal of that data is phenotype. In this work, digital information is placed in whole number variable and communicated as a color. A node stacked up with shading is made when digital data is 1 and 0 worth makes unfilled node.

2) Crossover: Crossover is a hereditary activity combining two chromosomes to make another posterity chromosome. In our circumstance, we utilize 1 point crossover, in which offspring chromosome shares half genetic data on network structure from first parent chromosome and other half from second parent chromosome. This ensures genotype information of cluster structures is being passed on to coming candidate solutions. Child chromosome bitstream contains half genetic data from first parent and half from second parent. The crossover probability picks whether crossover should happen or the parent chromosome should be clearly passed to generation to come.

3) Mutation: Newly created cluster structures are just recombination of existing ones, hence doesn't contain any variation in terms of converting a common node to cluster node and vice-versa. In out implementation scenario, mutation operation simply toggles value either from 0 to 1 or 1 to 0 i.e. converting a node to CH and vice-versa. Mutation rate should be wisely chosen, as the greater value may either overpopulate network with CHs or either completely remove them. After mutating a node to CH it is possible that the new CH may come in the vicinity of already existing CH, thus, producing an invalid chromosome which must be rejected.

F. Working Process of the Proposed Technique

In this section, step by step operations of the proposed technique are discussed. The mating pool algorithm is given as follows:

Algorithm 1 Mating Pool Algorithm

Input: Chromosome population $\mathrm{PC^P}$
Output: Returns mating pool array *mating_pool*

1. $\mathrm{PC^P}$ denotes chromosome population for create mating pool. Initialize *sum = 0* which holds sum of fitness of all chromosomes in $\mathrm{PC^P}$.
2. $sum = \sum_{k=1}^{len(PC^P)} PC_k^P.fitness$ // Add up all fitness values.
3. Initialize probability distribution array $Prob_F = []$ *(empty array)* identifying probability with fitness level of corresponding chromosome.
4. **for** x = 1 to *len*($\mathrm{PC^P}$) **do**
   $Prob_F.push((PC_x^P.fitness\ /\ sum) * 100)$
5. *mating_pool = []* storing chromosome indices from $\mathrm{PC^P}$.
6. **for** x = 1 to *len*($\mathrm{PC^P}$) **do**
7. i = $ceil(Prob_F[x] * PC_x^P.fitness)$
8. **for** j = 1 to i **do** // more value of i, higher is chance for selection
9. *mating_pool.push(x)* //add x′th chromosome j times to mating pool
10. **Return** *mating_pool*

The algorithm for CH election is given as follows:

The proposed works into two stages i.e. setup and steady stage. In the setup phase, the procedure starts with the organization of homogeneous and heterogeneous nodes. After the SNs and sink deployment, the clustering process is followed. The CH determination is done utilizing a GA. The network factors and GA parameters are introduced. After CH selection is resolved, information communication is initiated in the steady stage. As CH gets information from its CNs, it collects

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:42:59 UTC from IEEE Xplore. Restrictions apply.

Algorithm 2 Proposed Technique for CH Election

1. Initialize total number of SNs *N, p_size* as population size of genetic chromosomes crossover rate as C_{rate}, mutation rate as M_{rate}, α, β, γ and $gen_count = 0$
2. Create SNs array $SN = \{k = 1,2,3.....N\}$ filled with SNs of either homogeneous or heterogeneous types.
3. Let $PC^P = []$ hold current genetic population array.
4. **For** t = 1 to *p_size* **do //** Random genotype information generated
   $PC_t^P = generate_chromosome($ N)
5. Calculate each chromosome's fitness from t = 1 to *p_size* i.e. $C_t^P.fitness = Fitness(PC_t^P)$
6. **While** *termination critera not met* **do**
   i. fittest = for each Fitness(PC_i^P) > Fitness(PC_j^P) where i **is not equal** to j and $1 \leq i, j \leq pop_size$
   ii. **For** j = 1 to $len(PC^P)$ // Normalize population fitness
   PC_j^P.fitness = PC_j^P.fitness / fittest.fitness
   iii. Based on *rank selection method* create a mating pool for parent selection process. $mating_pool = mating_pool_procedure($ $PC^P)$
   iv. Initialize a new array $offspring_pop = []$ for holding offspring population and set $offset = 1$ *and* $elitisim = True$
   v. **IF** *elitism* is *true* **then** *offspring_pop.push(fittest)*
   vi. **while** $len(offspring_pop) \leq p_size$ **repeat**
   a) IF random(1) < Crate then
   $parent_{first} = PC^P[mating_pool[random(0, len(mating_pool))]]$
   $parent_{second} = PC^P[mating_pool[random(0, len(mating_pool))]]$
   $child =$ Crossover($parent_{first}$, $parent_{second}$)
   offspring_pop.push(child)
   b) Else directly add parent to next generation
   $offspring_pop.push(PC^P[offset])$
   $offset{+}{+}$
   vii. $PC^P = offspring_pop$ // New evolved population.
   viii. **For** k = 1 to $len(PC^P)$ **do //** Evaluate fitness of population
   $PC_k^P.fitness = Fitness(PC_k^P)$ // calculate fitness of k'th chromosome
7. **endWhile**

the data and forwards it to the sink if there should be an occurrence of single sink though, in the event of multiple sinks, it finds the closest found sink. After each dispersal, the node's energy is checked. In the event that it happens that it is equivalent to or under zero, the node should be dead and is spoken to by setting the $dead_node_i$ banner as true.

G. Complexity of Proposed and Existing Protocols

The proposed and existing protocols are based on the GA and their complexity depends on the way to implementation of GA specially on the digital representation of an entity being processed. The genetic operators and their implementation (as a chromosome), population size is chosen and calculation of fitness function; all indeed decide the overall complexity of the algorithm. The estimation depends on each individual operation being performed, e.g. mutation, crossover, fitness parameter calculation etc. The chromosome length which is also equal to the number of nodes in the network linearly affects the fitness calculation which can be put as $4 \times O(N)$ where N is the size of individuals (chromosomes) or the number of nodes in the proposed network. Mutation operation contributes $O(N)$ as it loops through every chromosome bit. The crossover takes two chromosomes and creates a child by linearly transferring chromosome data; therefore, adding a linear time of $O(N)$. After fitness calculation, the fittest chromosome from the population is linearly looked for which it adds the linear time of $O(p_size)$, where p_size is the population. So, the overall complexity for running the algorithm for g number of generations can be rationalized as $O(g \times (p_size \times (4 \times N) + (N + N) + p_size)) = O(g \times p_size \times n)$

V. Simulation Study and Discussions

In this section, the simulation is prepared by considering the disaster management scenarios for the smart city where different smart transportation system, smart hospitals, smart gas stations, smart apartments, smart central parks, smart schools, smart supermarkets, smart stadiums, etc. are considered in the virtual physical world. All the objects of the virtual physical world are mapped into the cyber world. The architecture of a smart city for all the essential applications is shown in Fig. 1. The disaster management server directly connected to the sink where multiple SNs help in forwarding the messages to the end-users to provide the required services. An example of a fire scenario has been considered by deployed 100 sensors with 1 or 4 sinks in the physical world. Thus, the decision of packet forwarding is taken at the disaster management control server, and packets are dispersed to the respective nodes through intermediate nodes. The SNs are deployed in the various objects of the smart city region $100 \times 100m^2$ by considering the total 70 Jules energy of the networks. These networks are simulated in MATLAB version R2014a with an Intel(R) Core(TM) i7 CPU and processor at 2.4 GHz with RAM 8GB. The radius is considered $70m$ for the sinks mobility outside the networks from the centre whereas inside the network it is $40m$.

The initial energy of the nodes in case of Str-1-Het is 0.7 J whereas in case of Str-2-Het networks the energies of the stair-1 and stair-2 nodes are 0.5 J and 1 J, respectively. The number of stair-1 and stair-2 nodes for Str-2-Het networks are 60 and 40, respectively. The threshold distance parameter ($\boldsymbol{d_0}$) is 87m and the number of nodes and energy fraction of intermediate nodes are $\alpha = 1$ and $\mathrm{þ}_1 = 0.4$, respectively. The energy consumption in running transmitter and receiver ($\varepsilon_{\mathbf{ele}}$), in amplification for $\boldsymbol{d} \leq \boldsymbol{d_0}$ ($\varepsilon_{\mathbf{efs}}$), in amplification for $\mathbf{d} > \boldsymbol{d_0}$($\varepsilon_{\mathbf{amp}}$), in data aggregation ($\varepsilon_{\mathbf{da}}$) are 50nJ/bit, $10pJ/bit/m^2$, $0.0013pJ/bit/m^4$ and $5nJ/bit/signal$, respectively. The p_size, crossover rate (C_{rate}), mutation rate (M_{rate}), crossover-type, population selection method, chromosome count, number of generations are 2000 bits, 30, 0.5, 0.006, 1-point crossover, rank election way, 30, and 30-50, respectively.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:42:59 UTC from IEEE Xplore. Restrictions apply.

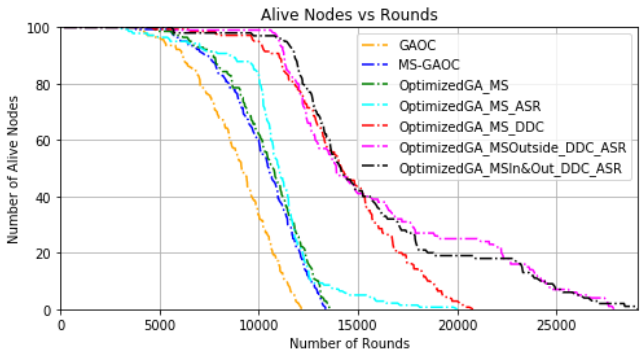


Fig. 2. Alive nodes vs. rounds for Str-1-Het.

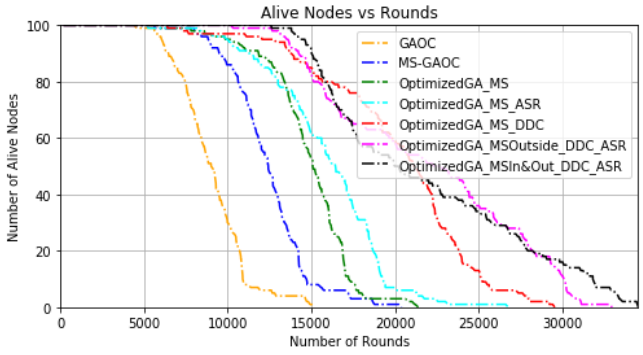


Fig. 3. Alive nodes vs. rounds for Str-2-Het.

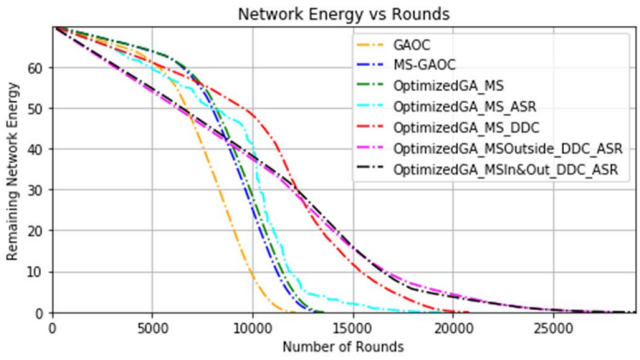


Fig. 4. Remaining network energy vs. rounds for Str-1-Het.

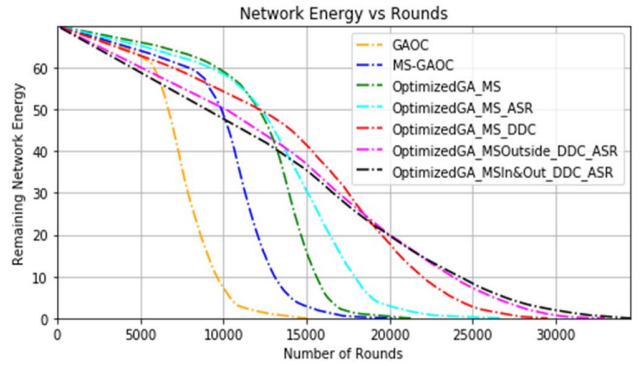


Fig. 5. Remaining network energy vs. rounds for Str-2-Het.

The performance of the proposed and existing protocols is analyzed on several performance metrics such as stability period, network lifetime, network remaining energy and throughput. In this work, five different protocols are proposed namely optimized GA with movable multiple sink (OptimizedGA-MS), optimized GA with multiple movable sinks and ASR (OptimizedGA-MS-ASR), optimized GA with movable multiple sinks and DDC (OptimizedGA-MS-DDC), optimized GA with outside movable multiple sinks, DDC and ASR (OptimizedGA-MSOutside-DDC-ASR), and optimized GA with inside & outside movable multiple sinks, DDC and ASR (OptimizedGA-MSIn&Out-DDC-ASR). Furthermore, all the proposed protocols are compared with the existing protocols i.e. GAOC and MS-GAOC [22] as these have outperformed from various routing strategies as discussed in the literature.

A. Network Lifetime and Stability Period

The comparative analysis of the active nodes per round for the existing and proposed protocols by considering 70 J network energy for Str-1-Het and Str-2-Het is shown in Figs. 2 & 3. It is evident from results that the percentage increment in the network lifetime for OptimizedGA-MS, OptimizedGA-MS-ASR, OptimizedGA-MS-DDC, OptimizedGA-MSOutside-DDC-ASR and OptimizedGA-MSIn&Out-DDC-ASR is 14.69%, 55.81%, 45.57%, 75.18% and 85.16% in case of Str-1-Het and 1.5%, 55.66%, 46.28%, 99.75% and 106.6% in case of Str-2-Het, respectively, as compared to the MS-GAOC. Such an improvement can be checked by the way that fitness function is fathomed with various fitness traits for CH determination. The high lingering energy factor and comparative bias further positively dominate for a choice of intermediate nodes as CH. In addition, the arrangement of nodes according to defined strategy contribute in lessening hot-spot issue. Similarly, un-important relaying of packets first to CH and afterwards to BS in any event, when BS is close enough for direct transmission is dealt with by DDC idea. This conserves high measure of energy by decreasing traffic of information bundles and becomes the predominant purpose behind energy preservation. Likewise, making detecting range adjustable prunes the system structure by making little clusters close to sink and bigger ones away from sink.

B. Network Remaining Energy

Figs. 4 & 5 show the system's remaining energy metric estimating energy utilization regarding the number of rounds is exhibited. More energy preservation is procured because of reduction in correspondence distance among sender and beneficiary nodes. Additionally, chosen CH has a higher number of neighboring nodes bringing about firmly bound cluster arrangement. In addition, energy conservation gets the advantage from the choice of intermediate nodes as CH. It very well may be understood that OptimizedGA_MS protocol streamlines energy utilization with the consolidation of fitness boundaries for cluster election. The sink movement

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:42:59 UTC from IEEE Xplore. Restrictions apply.

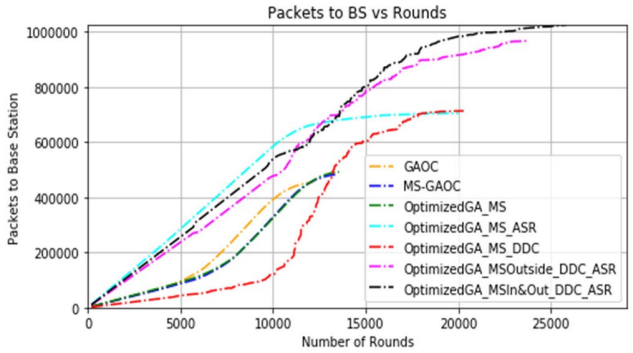


Fig. 6. Packets sent to BS vs. rounds for Str-1-Het.

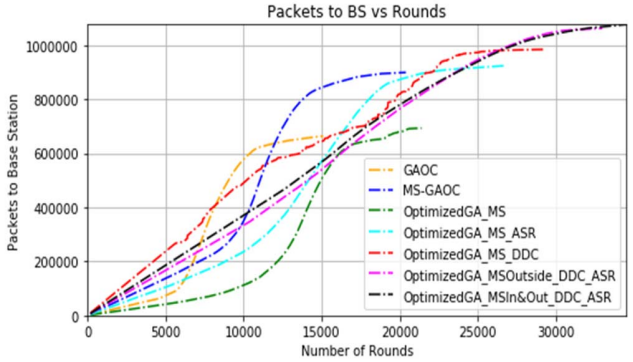


Fig. 7. Packets sent to BS vs. rounds for Str-2-Het.

decrease energy utilization rate when contrasted with static sink situation; creating better outcomes. In context of energy, ASR & DDC factor limits cluster development close to the BS subsequently, evacuating un-required packet transferring which in any case could have expended measure of energy.

C. Throughput

The packet transmission per rounds for Str-1-Het and Str-2-Het environment as shown in Figs. 6 & 7. A gigantic improvement in throughput is 2.2%, 45.96%, 43.77%, 94.85% and 96.13% in case of Str-1-Het & 3.24%, 46.35%, 38.33%, 58.22%, and 60.32% in case of Str-2-Het for the protocols OptimizedGA-MS, OptimizedGA-MS-ASR, OptimizedGA-MS-DDC, OptimizedGA-MSOutside-DDC-ASR and OptimizedGA-MSIn&Out-DDC-ASR, respectively. The fitness parameters for CH choice joined with sink movement and DDC assortment advances life expectancy anticipation and throughput. Communicating more data is just conceivable if nodes live for a longer time. Moreover, a great deal of CH's energy is safeguarded when sink moves toward near it; as it's cluster size decrease to zero making nodes to legitimately communicated with sinks directly.

The un-vital transferring of data when the sink is close enough for direct correspondence spares a lot of network energy which shows up back in the type of expanded throughput. CH election race supported by progressively clusters make nodes to transmit data for long and bring network longevity.

TABLE I
COMPARATIVE ANALYSIS OF PROPOSED METHODS WITH OTHER ALGORITHMS FOR DIFFERENT METRICS

Protocol	Total Energy	Stability Period	HND	Network Lifetime	Throughput
GAOC [22]	70	4573	9011	15097	663054
MS-GAOC [22]	70	4978	12400	18638	671434
PSO-ECSM [12]	140	6301	12558	19071	637880
GAPSO-H [13]	140	6387	12946	24988	792952
DDMPEA [16]	70	5870	10627	21000	168000
EEWC [10]	100	1324	1643	4770	43954
OptimizedGA_MS	70	5937	15075	21376	693220
OptimizedGA_MS_DDC	70	6112	17978	29040	982622
OptimizedGA_MS_ASR	70	6211	20200	27131	928798
OptimizedGA_MSOutside_DDC_ASR	70	6345	21869	32650	1062369
OptimizedGA_MSIn&Out_DDC_ASR	70	6524	22512	34511	1076434

Table I shows the comparative analysis of proposed methods with other algorithms by considering different metrics such as stability period (first node dead), half node dead (HND), network lifetime (last node dead) information along with the throughput. It is evident from Table I that the proposed variants perform better than the existing one due to its efficient CHs election and data collection process.

D. Load Balancing Analysis of Deployed Sinks

This section analyzes the load balancing to conveyed sinks per rounds. The data packets being transmitted from the CHs should be impartially distributed among the different deployed closest sinks by dynamically finding out their optimum distance from multiple moving sinks. In static sink condition, CHs to sink distance stand separated to be consistent anyway under mobile sink circumstance, this separation is evaluated constantly before CHs decide to send data to the closest open sink. As Fig. 8 shows that under movable sink scenario, sinks are getting around proportionate number of packets and shows the absolute sum of data packets being accumulated by the deployed sinks. The sink movement of OptimizedGA protocol ensures that sinks are getting equal weights from dense and inadequate SN territories. This circumstance is moved by sink movement to and fro and gathering data from both dense and meager network locales nearby ensuring transmission distance optimization. In OptimizedGA_MSIn&Out_DDC_ASR protocol, 2 sinks are inside and the other 2 are outside the network domain. As such, inside sinks collect a more prominent number of packets than outer sinks as they are a great deal closer to SNs. The outside sinks assemble packets from nodes at the boundaries of the system. Additionally, the thought of consideration of DDC doesn't affect the load request demand on relating sinks. As clustering process proceeds, ASR contribute by altering the cluster size and SNs keep changing their tendency between CH, cluster and autonomous node. Moving sinks keep gathering data from eagerly transmitting CHs, therefore, equally circling sink load. In this manner, the proposed system significantly conveys the solicitation load

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:42:59 UTC from IEEE Xplore. Restrictions apply.

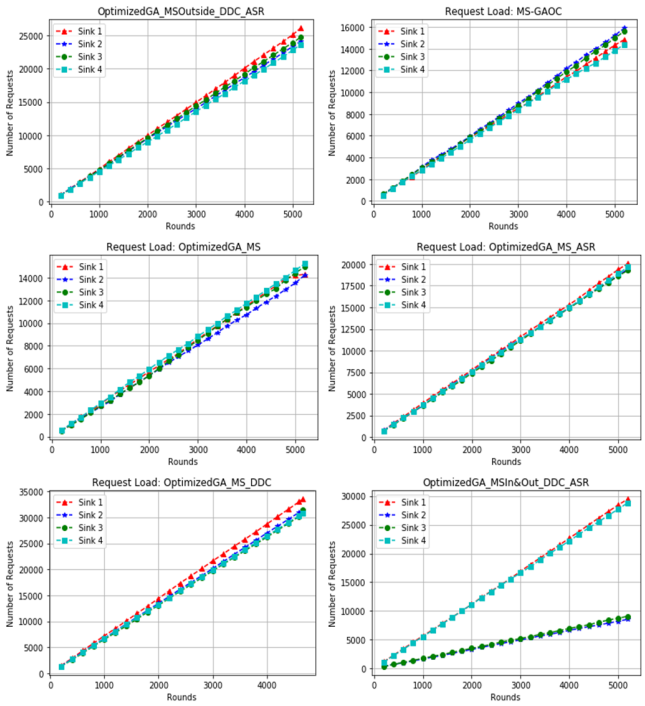


Fig. 8. Number of requests received by the sinks for the proposed and existing multiple sink scenario.

among the multiple sinks helping in improving sensor network nature of CH structuring.

VI. Conclusion and Future Works

In this paper, a set of OptimizedGA protocols i.e., based on DDC and ASR using stable and movable sinks have been proposed for providing extended stable period in the execution of the network. The minimum and maximum percentage increment in the network lifetime is 14.69% & 85.16% and 1.5% & 106.6%, for Str-1-Het and Str-2-Het, respectively, as compared to the MS-GAOC. There are multiple reasons for such a gigantic enhancement in the network lifespan. Firstly, the selection of the highest capable node as CH using the fitness function where comparative bias parameter encourages the selection of higher energy nodes as CH. Secondly, the DDC method using movable sink inside and outside the network region diminishes remaining energy at a slower rate in comparison to static sink environment. Finally, the ASR optimizes cluster size with respect to the closest sink and reduces the number of hops to deliver packets to respective sinks. This confirms the effectiveness of the proposed protocols by maintaining the proper load balancing and alleviating the hot spot problem. The future work can be extended by including the developing fault tolerance system and secure data transmission using authentication mechanisms [23].

References

[1] X. Fu, H. Yao, O. Postolache, and Y. Yanga, "Message forwarding for WSN-assisted opportunistic network in disaster scenarios," *J. Netw. Comput. Appl.*, vol. 137, pp. 11–24, Jul. 2019.

[2] S. Chand, S. Singh, and B. Kumar, "Heterogeneous HEED protocol for wireless sensor networks," *Wireless Pers. Commun.*, vol. 77, no. 3, pp. 2117–2139, Aug. 2014.

[3] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *Proc. 33rd Annu. Int. Conf. Syst. Sci. Hawaii*, 2001, pp. 3005–3014.

[4] R. Hamidouche, Z. Aliouat, and A. M. Gueroui, "Genetic algorithm for improving the lifetime and QoS of wireless sensor networks," *Wireless Pers. Commun.*, vol. 101, no. 4, pp. 2313–2348, Aug. 2018.

[5] S. K. Jha and E. M. Eyong, "An energy optimization in wireless sensor networks by using genetic algorithm," *Telecommun. Syst.*, vol. 67, no. 1, pp. 113–121, Jan. 2018.

[6] M. Al-Shalabi, M. Anbar, T.-C. Wan, and Z. Alqattan, "Energy efficient multi-hop path in wireless sensor networks using an enhanced genetic algorithm," *Inf. Sci.*, vol. 500, pp. 259–273, Jun. 2019.

[7] J. Bhola, S. Soni, and G. K. Cheema, "Genetic algorithm based optimized leach protocol for energy efficient wireless sensor networks," *J. Ambient Intell. Hum. Comput.*, vol. 11, no. 3, pp. 1281–1288, 2020.

[8] S. Rani, S. H. Ahmed, and R. Rastogi, "Dynamic clustering approach based on wireless sensor networks genetic algorithm for IoT applications," *Wireless Netw.*, vol. 26, no. 4, pp. 2307–2316, May 2020.

[9] P. S. Ragavan and K. Ramasamy, "Optimized routing in wireless sensor networks by establishing dynamic topologies based on genetic algorithm," *Cluster Comput.*, vol. 22, no. 5, pp. 12119–12125, Sep. 2019.

[10] R. Pal, S. Yadav, and R. Karnwal, "EEWC: Energy-efficient weighted clustering method based on genetic algorithm for HWSNs," *Complex Intell. Syst.*, vol. 6, no. 2, pp. 391–400, Jul. 2020.

[11] A. S. Nandan, S. Singh, and L. K. Awasthi, "An efficient cluster head election based on optimized genetic algorithm for movable sinks in IoT enabled HWSNs," *Appl. Soft Comput.*, vol. 107, Aug. 2021, Art. no. 107318.

[12] B. M. Sahoo, T. Amgoth, and H. M. Pandey, "Particle swarm optimization based energy efficient clustering and sink mobility in heterogeneous wireless sensor network," *Ad Hoc Netw.*, vol. 106, Sep. 2020, Art. no. 102237.

[13] B. M. Sahoo, H. M. Pandey, and T. Amgoth, "GAPSO-H: A hybrid approach towards optimizing the cluster based routing in wireless sensor network," *Swarm Evol. Comput.*, vol. 60, Feb. 2021, Art. no. 100772.

[14] S. Dehghani, B. Barekatain, and M. Pourzaferani, "An enhanced energy-aware cluster-based routing algorithm in wireless sensor networks," *Wireless Pers. Commun.*, vol. 98, no. 1, pp. 1605–1635, Jan. 2018.

[15] T. Wang, G. Zhang, X. Yang, and A. Vajdi, "Genetic algorithm for energy-efficient clustering and routing in wireless sensor networks," *J. Syst. Softw.*, vol. 146, pp. 196–214, Dec. 2018.

[16] S. Chauhan, M. Singh, and A. K. Aggarwal, "Cluster head selection in heterogeneous wireless sensor network using a new evolutionary algorithm," *Wireless Pers. Commun.*, vol. 119, no. 1, pp. 585–616, Jul. 2021, doi: 10.1007/s11277-021-08225-5.

[17] M. V. Ramesh *et al.*, "Achieving sustainability through smart city applications: Protocols, systems and solutions using IoT and wireless sensor network," *CSI Trans. ICT*, vol. 8, no. 2, pp. 213–230, Jun. 2020.

[18] L. Kong, D. Zhang, Z. He, Q. Xiang, J. Wan, and M. Tao, "Embracing big data with compressive sensing: A green approach in industrial wireless networks," *IEEE Commun. Mag.*, vol. 54, no. 10, pp. 53–59, Oct. 2016.

[19] X. Fu, G. Fortino, P. Pace, G. Aloi, and W. Li, "Environment-fusion multipath routing protocol for wireless sensor networks," *Inf. Fusion*, vol. 53, pp. 4–19, Jan. 2020.

[20] X. Li, D. Li, J. Wan, C. Liu, and M. Imran, "Adaptive transmission optimization in SDN-based industrial Internet of Things with edge computing," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1351–1360, Jun. 2018.

[21] X. Fu, G. Fortino, W. Li, P. Pace, and Y. Yang, "WSNs-assisted opportunistic network for low-latency message forwarding in sparse settings," *Future Gener. Comput. Syst.*, vol. 91, pp. 223–237, Feb. 2018.

[22] S. Verma, N. Sood, and A. K. Sharma, "Genetic algorithm-based optimized cluster head selection for single and multiple data sinks in heterogeneous wireless sensor network," *Appl. Soft Comput.*, vol. 85, Dec. 2019, Art. no. 105788.

[23] X. Fu and Y. Yang, "Modeling and analyzing cascading failures for Internet of Things," *Inf. Sci.*, vol. 545, pp. 753–770, Feb. 2021.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:42:59 UTC from IEEE Xplore. Restrictions apply.



Aridaman Singh Nandan received the B.Tech. degree in computer science and engineering from Guru Nanak Dev University, Amritsar, Punjab, India, in 2018, and the M.Tech. degree in computer science and engineering from Dr. B. R. Ambedkar National Institute of Technology Jalandhar, Punjab, in 2020. His research areas lie in the area of information security and wireless sensor networks.



Aruna Malik (Member, IEEE) received the B.Tech. degree in computer science and engineering from Uttar Pradesh Technical University, Lucknow, India, in 2007, and the M.Tech. and Ph.D. degrees in computer science and engineering from Dr. B. R. Ambedkar National Institute of Technology Jalandhar, Punjab, India, in 2012 and 2017, respectively. Currently, she is working as an Assistant Professor with the Department of Computing Science and Engineering, Dr. B. R. Ambedkar National Institute of Technology Jalandhar. Her research areas lie in the area of the IoTs, wireless sensor networks, data hiding, and image processing.



Samayveer Singh (Member, IEEE) received the B.Tech. degree in information technology from Uttar Pradesh Technical University, Lucknow, India, in 2007, the M.Tech. degree in computer science and engineering from Dr. B. R. Ambedkar National Institute of Technology Jalandhar, Punjab, India, in 2010, and the Ph.D. degree from the Department of Computer Engineering, Netaji Subhas Institute of Technology, University of Delhi, New Delhi, India, in 2016. Currently, he is working as an Assistant Professor with the Computer Science and Engineering Department, Dr. B. R. Ambedkar National Institute of Technology Jalandhar. His research interests include the Internet of Things, wireless sensor networks, data hiding, and information security.



Rajeev Kumar received the B.Tech. degree in information technology from Uttar Pradesh Technical University, Lucknow, India, in 2007, the M.Tech. degree in information systems from Netaji Subhas Institute of Technology, University of Delhi, India, in 2012, and the Ph.D. degree from the Computer Engineering Department, Netaji Subhas Institute of Technology, University of Delhi, in 2017. He has also worked as a Postdoctoral Researcher at Kyungil University, South Korea. He is currently working as an Assistant Professor with the Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India. His research areas lie in the area of wireless networks, the IoTs, data hiding, and image processing.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:42:59 UTC from IEEE Xplore. Restrictions apply.

2021 8th International Conference on Signal Processing and Integrated Networks (SPIN) | 978-1-6654-3564-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SPIN52536.2021.9566059

A Modified-Zeta Converter based Onboard Battery Charging with Improved THD

Chanchal Chaudhary
Electrical Engineering Department
Delhi Technological University
Shahbad Daulatpur, Main Bawana Road, Delhi NCR-110042
Delhi, India
chanchalchaudhary94@gmail.com

Saurabh Mishra
Electrical Engineering Department
Delhi Technological University
Shahbad Daulatpur, Main Bawana Road, Delhi NCR-110042
Delhi, India
saurabhmishra@dtu.ac.in

Uma Nangia
Electrical Engineering Department
Delhi Technological University
Shahbad Daulatpur, Main Bawana Road, Delhi NCR-110042
Delhi, India
uma_nangia@rediffmail.com

Abstract*— This paper presents onboard battery charging with a modified Zeta converter. Charging system is incorporated with two clamping diodes and two extra switches at Zeta converter input resulting in low THD input current and better charging efficiency by mitigating the problems of conventional Zeta converter. An AC-DC conversion is employed by diode bridge rectifier (DBR) followed by modified Zeta converter as power factor correction (PFC) unit, and the flyback converter is used to synchronize the current of battery with the implementation of closed loop cascaded PI control during constant voltage (CV) mode and constant current (CC) mode. Zeta converter is operated in Discontinuous conduction mode (DCM) mode such that Zeta converter can work with better dynamic response, and low ripple at the output voltage. Linear PI control topology is employed here, which has cascaded control for the better working efficiency of the charger. Proposed converter operation is to achieve low total harmonic distortion (THD) within IEC 61000-3-2 standards. Due to inbuilt isolation, components are less and hence proving the system to be more reliable. Proposed system's efficacy is validated in MATLAB SIMULINK.

Keywords—: Improved Zeta converter, DCM mode, Diode Bridge rectifier (DBR), Power factor correction (PFC), Total Harmonic Distortion (THD).

I. INTRODUCTION

Increasing pollution and changing lifestyles has raised a question on use of conventional resources like petrol, Diesel etc. and hence a new and cleaner fuel source i.e., Electric Vehicle (EV) has become new interest area for researchers. Public transportation electrification has resulted in many environmental as well as commercial benefits by developing a good energy saving implementation and mechanism which consequently became helpful in cutting down the emission of greenhouse gases which are hazardous for our environment. Utilization of maximum energy and demands of high efficiency of fuel has traversed the path of automotive industry towards Electric vehicles very rapidly. Control techniques in motors are also becoming efficient and helpful in promoting EV at the end of vehicular industry. Moreover, there is a high motivating force from the side of government also to develop new and energy efficient methods in the area of EV. The electric vehicle charger's topology involves an AC-DC conversion unit, generally DBR is used, which consequently introduces harmonics in the input AC side. Due to this, THD of input current increases to an unacceptable range thereby compelling the power factor to reach unity. Generally, the input line current of a system with DBR is having exceptionally high THD about 50-60% of the evaluated load due to the line current distortions and hence the conversion efficiency is highly deteriorated. Distortion in power factor is a power quality concern which is important to be addressed as power factor much lower than unity leads to the poor efficiency and introduces the harmonics in the whole connected system as well So, in general EMI filters are used at the input side of DBR which improves the current profile of system but still the THD is not reduced to the required limits so there is a requirement of power factor correction unit. Subsequently, a PFC unit is usually incorporated in the basic topology which not only reduces THD of current at input but also improves power factor close to unity while nowadays in some topologies Isolated converters are also used to improve the performance of charger further. Isolation in the circuit reduces the distortion to further extent. These chargers with Isolated converters are known as two stage chargers.

In power factor correction unit any basic topology like buck, boost, buck-boost or other DC-DC converters like CUK, Zeta, Luo, Landsman etc. can be used. Many of the topologies for chargers are discussed in [1]. Out of conventional topologies buck and boost are not suitable choices, as per [2], for PFC unit, due to limited variation in duty cycle while buck-boost converter is preferred over them as being a basic converter with less components and wide range of variation in duty cycle duly mentioned in [3]. Modified techniques are also helpful in buck-boost topology and offers some advantages over conventional buck boost as given in [4]. Moreover, CUK, SEPIC [5], Zeta are the better and favorable configurations over buck-boost converter due to their current waveforms and better configuration design. CUK is relevant choice in chargers as per advantages mentioned in [6] i.e., Less EMI noise, lower interference etc. while SEPIC as described in [7] is though better than conventional converters but comes with some drawback mentioned in [8] i.e., output current discontinuity and high amount of ripple content in output current. Zeta converter has certain qualities as mentioned in [9] like positive output voltage, finer dynamic response in DCM mode and low ripple present at the output which makes it apt for choosing

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:40:15 UTC from IEEE Xplore. Restrictions apply.

as PFC in EV but in case of DCM mode and high-power rating operations, voltage stresses on PFC switches are quite high in Zeta converter. So, in context of these issues, this paper presents a modified topology of Zeta converter that includes two extra switches and two clamping diodes helping in reducing the stress at input of DBR. Modified topologies as given in [10] and [11] give the required voltage at DC link with lesser duty ratio as compared from conventional topologies. So, they work better in aspect of lowering THD thereby enhancing the power factor at input. ZVS switching can also be useful as discussed in [12]. Flyback converter design is as per [13]. It is mentioned very well in [14] that how high frequency transformers are helpful in mitigating the further ripples in system. There is no inclusion of extra inductors or capacitors to improve efficiency as these limits the power density of the system converter. So, modified topology is worth as PFC unit. It reduces the ripple content and lowers down THD within limits making the system to work more efficiently and maintaining the reliability of it.

II. Configuration and Working Operation

Fig.1 exhibits the configuration of charger, which depicts a DBR (being simplest) used for conversion of AC-DC, an EMI filter with filter inductance L_f and filter capacitance C_f 'followed by modified Zeta converter incorporated with two switches S_1 and S_2 along with two clamping diodes D_1 and D_2. This whole system is cascaded with a flyback converter controlling charging current by CC-CV charging and charging a battery of 48V,80AH. Modified Zeta converter is employed as a power factor correction unit here which is also improving the efficiency by reducing the voltage stress at DBR input.

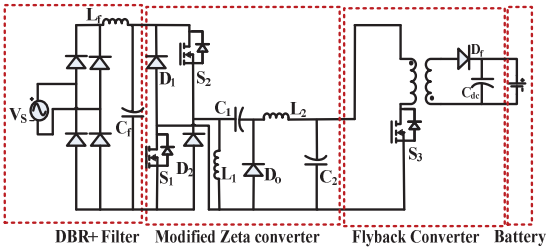


Fig.1. Proposed Configuration

This configuration has three modes of operation. Both the converters are selected to work in DCM mode to make the system efficient. Operating modes for modified Zeta converter as PFC unit is described in Fig.2(a)-Fig.2(c) with switches S_1, S_2 being on and off in DCM mode accordingly. Directions of flow of current in different working loops are shown in figures.

- Mode-1: This mode operation involves switches S_1, S_2 switched simultaneously while inductor L_1 starts charging via supply. Diode D_o is in non-conducting state. As inductor L_2 is storing energy, at the same time voltage across capacitor C_1 starts reducing as can be seen from Fig.2(a)

- Mode-2: From Fig.2(b) it can be seen that switch S_1, S_2 both are in off condition in mode 2 while diodes D_1, D_2 and clamping diodes become forward biased. Capacitance C_1 is getting charged through D_o diode. Inductor L_1 is getting discharged by diodes D_1 and D_2. Inductor L_2 supplies rest of the part of system with the help of capacitor C_2 and diode D_o.

- Mode-3: In this mode, system enters in DCM mode or in other words freewheeling period as clamping diodes and both switches are in off state as shown in Fig.2(c). Inductances L_1 and L_2 has currents in such a way that current in diode is tend to be zero. C_1 i.e., the transfer capacitance supplies energy requirement of load via L_2 and C_2. Switches S_1, S_2 and diodes have peak voltages quite less than the conventional converter and is given by:

$$V_{S1} = V_{S2} = V_S \qquad (1)$$

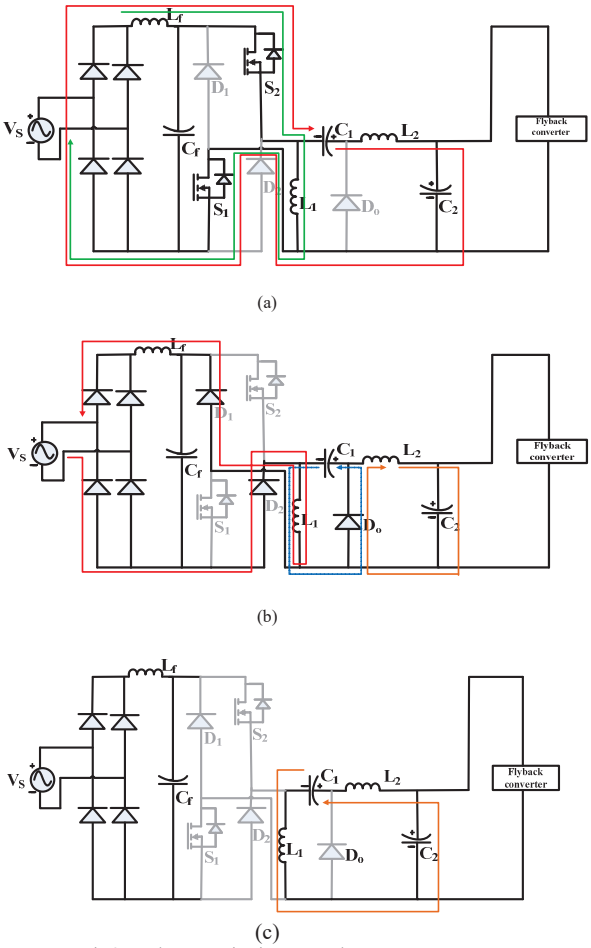


Fig.2. Mode's operation in Improved Zeta converter
(a) Mode1 (b) Mode 2 (c) Mode 3

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:40:15 UTC from IEEE Xplore. Restrictions apply.

III. DESIGN CONSIDERATIONS

This system involves two DC link capacitors i.e., C_2 and C_{dc}. former is the DC link generated at output of Zeta converter whereas latter is associated with flyback converter. . Voltage at flyback is used to charge the Li-ion battery of rated voltage specified as 48V. Switching frequency in flyback converter is chosen to be 50 kHz while for Zeta converter it is 20kHz. To make Zeta converter work in DCM mode, Duty cycle is calculated as:

$$D(t) = \frac{V_s}{V_2 + \sqrt{2}V_s} \tag{2}$$

where V_2 is the voltage at DC link of Zeta converter

Critical value of Input Inductance L_1 is calculated as:

$$L_1 = \left(\frac{V_{s\min}^{\ 2}}{P}\right)\left(\frac{V_s}{V_2 + \sqrt{2}V_s}\right)\frac{T_s}{2} \tag{3}$$

where T_s is the switching time period of Zeta converter.

Critical value of Series capacitance is given by:

$$C_1 = \frac{P}{\mu f\left(V_2 + \sqrt{2}V_{s\max}\right)} \tag{4}$$

where μ is the ripple factor for series capacitance

Critical value of inductance 'L2'is given by:

$$L_2 = \left(\frac{V_{s\min}^{\ 2}}{P}\right)\left(\frac{V_s}{V_2 + \sqrt{2}V_s}\right)\frac{1}{\lambda f} \tag{5}$$

where λ is the ripple factor for current in inductor.

Critical value of DC-link capacitance for Zeta converter is given by:

$$C_2 = \frac{I_2}{4\pi f \Delta V_2} \tag{6}$$

Filter is designed to improve the current profile at the input of Zeta converter. It has an inductance and capacitance designed accordingly to reduce the ripples in input current. Critical value of filter capacitance is designed as:

$$C_{f\max} = \frac{P\tan\theta}{2\pi f_c {V_s}^2} \tag{7}$$

Filter inductance ' L_f ' is calculated as:

$$L_f = \left[\frac{1}{4\pi f_c^{\ 2} C_f}\right] - \left[\frac{\left(0.04{V_s}^2\right)}{2\pi fP}\right] \tag{8}$$

where f_c is cut off frequency of the filter which is chosen to be 1/10 times of the frequency of converter.

In design of flyback converter magnetizing inductance plays an important role. chosen value of magnetizing inductance should be less than critical calculated value. Magnetizing inductance's critical value is given by:

$$L_m = \frac{\left(D_f V_2\right)^2}{2 f_{sw} I_{dc} V_{dc}} \tag{9}$$

where ' f_{sw} 'is flyback converter's switching frequency.

TABLE 1. CHARGER SPECIFICATIONS

S.No.	Parameter	Calculations
1.	Input Inductance ' L_1 '	110μH
2.	Input Series capacitor ' C_1 '	0.88μH
3.	Output capacitor ' C_2 '	2mF
4.	Output Inductance ' L_2 '	2.8mH
5.	Filter Inductance ' L_f '	4mH
6.	Filter capacitance ' C_f '	580nF
7.	DC link capacitor ' C_{dc} '	3mH
8.	Magnetizing inductance ' L_m '	130μH
9.	Zeta converter Switching frequency ' f '	20kHz
10.	Flyback converter Switching frequency ' f_{sw} '	50kHz

1V. CONTROL STRATEGY

The control technique for the proposed model is having two separate PI loops for each i.e., Flyback converter is being operated on cascaded loop of PI control which tracks battery voltage as well as keeps battery current in limits as shown in Fig.3. Battery voltage is being compared with reference voltage to produce an error signal fed to voltage PI controller which in turn is generating error signal to feed into current controller, further pulses are generated with PWM generator unit and switches S_3 while Modified Zeta converter has single PI loop control. It feedbacks DC link voltage after getting compared with reference voltage and gives the pulse for switching S_1 and S_2 used in modified Zeta converter as per Fig.4.

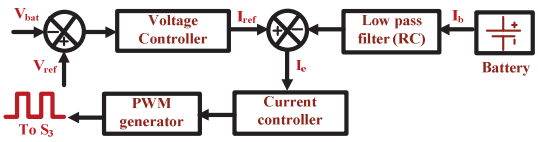


Fig.3. Control loop for Flyback converter

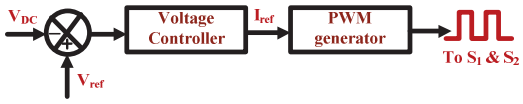


Fig.4. Control loop for Modified Zeta converter

Let A_1, A_2, A_3 be the state space matrix of mode 1, mode 2 and mode 3 respectively. For state space matrix capacitor voltages and inductor currents are chosen as variables. Using KVL and KCL equations for each mode state space matrix can be generated.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:40:15 UTC from IEEE Xplore. Restrictions apply.

Let $x_1 = I_{L1}$, $x_2 = I_{L2}$, $x_3 = V_{C1}$, $x_4 = V_{C2}$ (10)

For mode 1 matrix A_1 is given by: $x_1 = A_1 x + B_1$

$$\dot{x_1} = \frac{V_S}{L_1} \tag{11}$$

$$\dot{x_2} = \frac{V_S}{L_2} + \frac{x_3}{L_2} - \frac{x_4}{L_2} \tag{12}$$

$$\dot{x_3} = \frac{x_1}{C_1} \tag{13}$$

$$\dot{x_4} = \frac{x_1}{C_2} - \frac{x_4}{RC_2} \tag{14}$$

$$A_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{L_2} & -\frac{1}{L_2} \\ \frac{1}{C_1} & 0 & 0 & 0 \\ \frac{1}{C_2} & 0 & 0 & -\frac{1}{RC_2} \end{bmatrix}$$

Similarly, for mode 2 matrix A_2 can be calculated as:

$$\dot{x_1} = \frac{V_S}{L_1} \tag{15}$$

$$\dot{x_2} = \frac{x_4}{L_2} \tag{16}$$

$$\dot{x_3} = \frac{x_1}{C_1} \tag{17}$$

$$\dot{x_4} = \frac{x_2}{C_2} - \frac{x_4}{RC_2} \tag{18}$$

$$A_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{L_2} & -\frac{1}{L_2} \\ \frac{1}{C_1} & 0 & 0 & 0 \\ 0 & \frac{1}{C_2} & 0 & \frac{1}{RC_2} \end{bmatrix}$$

For mode 3 A_3 can be calculated as:

$$\dot{x_1} = \frac{x_4}{L_1} + \frac{x_3}{L_1} \tag{19}$$

$$\dot{x_2} = \frac{x_4}{L_2} + \frac{x_3}{L_2} \tag{20}$$

$$\dot{x_3} = \frac{x_2}{C_1} \tag{21}$$

$$\dot{x_4} = \frac{x_2}{C_2} - \frac{x_4}{RC_2} \tag{22}$$

$$A_3 = \begin{bmatrix} 0 & 0 & \frac{1}{L_1} & \frac{1}{L_1} \\ 0 & 0 & \frac{1}{L_2} & \frac{1}{L_2} \\ 0 & \frac{1}{C_1} & 0 & 0 \\ 0 & \frac{1}{C_2} & 0 & -\frac{1}{RC_2} \end{bmatrix}$$

For DCM mode:

$$\text{Average } A = A_1 d_1 + A_2 d_2 + A_3(1 - d_1 - d_2) \tag{23}$$

where d_1, d_2 are duty ratios for mode 1 and mode 2 respectively. Matrix 'A' is generated after solving equation (23). This is used to find out transfer function of the model. Tuning of PI can be done using various methods. This model includes tuning via Zeigler Nicholas method where transfer function is calculated with the help of steady state analysis and state space modelling. The critical value of 'K_{cr}' can be calculated using Routh Hurwitz criteria which is consisting of auxiliary equation helpful in finding out the value of 'T_{cr}'. By Zeigler Nicholas method, values of 'K_p' and 'K_I' can be calculated as:

$$K_p = 0.45 K_{cr} \tag{24}$$

$$K_I = \frac{1.2 K_p}{T_{cr}} \tag{25}$$

$$T_{cr} = \frac{2\pi}{\omega_{cr}} \tag{26}$$

Critical value of 'K_{cr}' can be calculated by Routh-Hurwitz criteria.

V. RESULTS

The system shows the charging of Li-ion battery of 48V, 80AH Fig.5. (a)-Fig.5. (e) depicts Input Voltage V_s, Input

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:40:15 UTC from IEEE Xplore. Restrictions apply.

Current I_s Battery Voltage V_{bat} , Battery Current I_{bat} and SOC% respectively. Fig.5. (a) and Fig.5. (b) shows that power factor of system is almost unity. THD in input current which is 7.67% as shown in Fig.6. It is getting charged up to 52V-56V. DC voltage generated at the input of flyback converter is found to be in the range of 50V-55V Current remains positive in case of discharging of battery while it becomes negative whenever battery is getting charged. So as current is building up and has been settled up to 14A and negative in nature shows that it is charging the battery with constant 14A. SOC% is steadily increasing showing the charging of battery gradually. It can be clearly seen in Fig.7. that how improved Zeta reduces THD to much larger extent and PFC plays a very important role in reduction of harmonics.

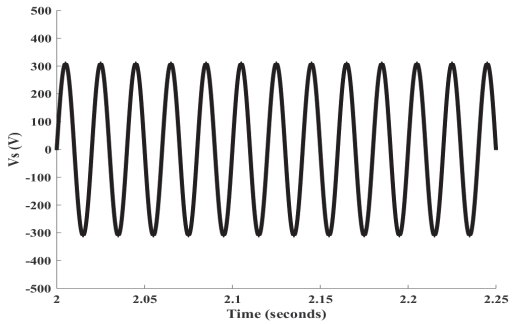


Fig.5. (a) Input Voltage (V_S)

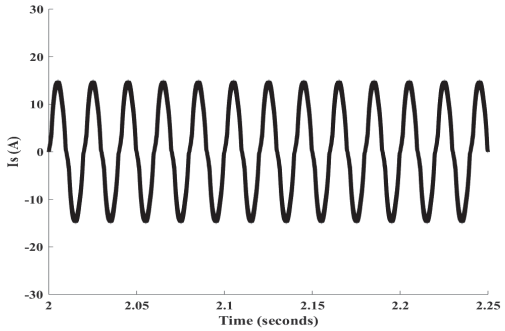


Fig.5. (b) Input Current (I_s)

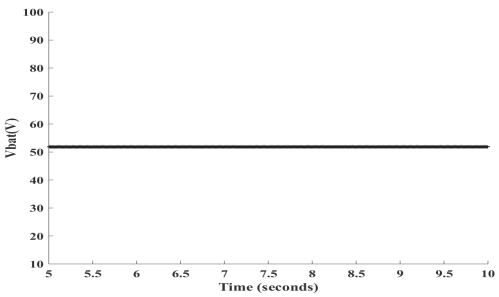


Fig.5. (c) Battery Voltage (V_{bat})

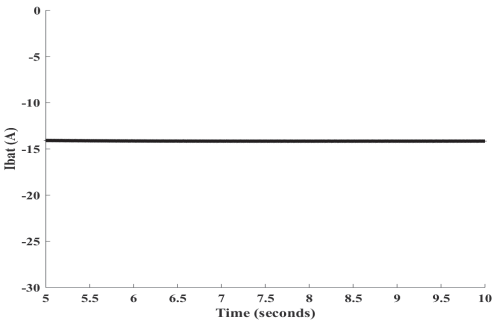


Fig.5. (d) Battery Current (I_{bat})

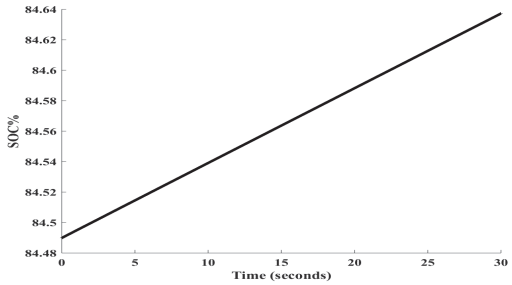


Fig.5. (e) SOC%

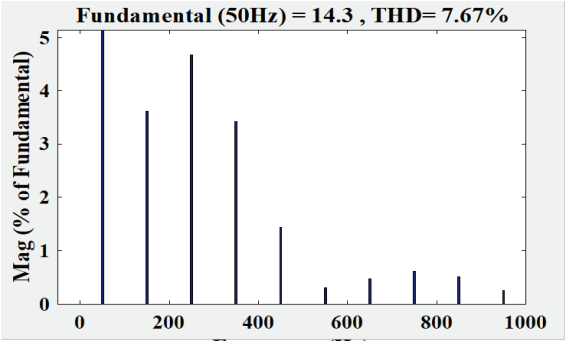


Fig.6. THD of input current

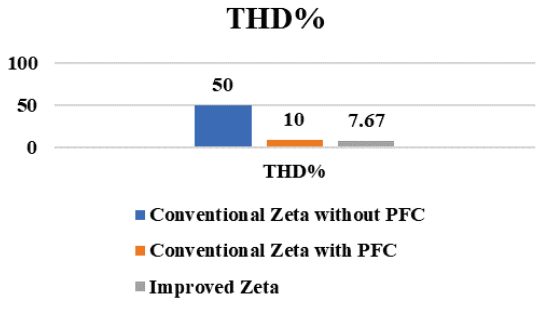


Fig.7. Comparison of THD

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:40:15 UTC from IEEE Xplore. Restrictions apply.

6. CONCLUSION

Modified Zeta converter is efficient as compared to conventional Zeta converter as results of PFC is found out improving in former's case and THD has been reduced to an acceptable range. Advantages of model includes lower stress on switches of DBR and lower losses thereby reducing harmonics and making it much more efficient and reliable in working as PFC unit in charger which makes it easy to use in low power application devices and also making it more cost efficient. This also ensures the cost cutting on requirement of a smaller number of sensors in the circuit due to DCM mode being used in both the converters. Overall, this model gives enhanced power factor closer to unity and low THD properties.

ACKNOWLEDGMENT

The author thanks to the electrical engineering department (EED), Delhi Technological University and SPIN to providing me the opportunity and platform to present my research findings and learnings.

REFERENCES

[1] B. Singh, B. N. Singh, A. Chandra, K. Al-Haddad, A. Pandey and D. P. Kothari, "A review of single-phase improved power quality AC-DC converters," in *IEEE Transactions on Industrial Electronics*, vol. 50, no. 5, pp. 962-981, Oct. 2003.

[2] S. Singh, B. Singh, G. Bhuvaneswari, Vashist Bhist, "Improved power quality switched mode power supply using buck-boost converter", *IEEE Trans. on industry application, International Conference on Power Electronics, Drives and Energy Systems (PEDES),* pp.1-6, Dec 2014.

[3] R. Kushwaha and B. Singh, "An Improved Battery Charger for Electric Vehicle with High Power Factor," 2018 *IEEE Industry Applications Society Annual Meeting (IAS)*, pp. 1-8, Sep 2018.

[4] R. Kushwaha and B. Singh, "An Improved Battery Charger for Electric Vehicle with High Power Factor,*"IEEE Industry Applications Society Annual Meeting (IAS)",* pp. 1-8, Sep 2018.

[5] R. Kushwaha and B. Singh, "An Isolated Bridgeless Cuk-SEPIC Converter Fed Electric Vehicle Charger*," IEEE Industry Applications Society Annual Meeting*, pp. 1-7*,* Oct 2020.

[6] R. Kushwaha and B. Singh, "A Modified Bridgeless Cuk Converter based EV Charger with Improved Power Quality," *IEEE Transportation Electrification Conference and Expo (ITEC)*, Detroit, MI, USA, pp. 1-6, June 2019.

[7] J. Gupta and B. Singh, "Single Stage Isolated Bridgeless Charger for Light Electric Vehicle with Improved Power Quality,"*IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES),* Dec 2020,

[8] R. Kushwaha and B. Singh, "Bridgeless Isolated Zeta–Luo Converter-Based EV Charger with PF pre regulation," in IEEE Transactions on Industry Applications, vol. 57, no. 1, pp. 628-636, Jan.-Feb. 2021.

[9] R. Kushwaha, B. Singh and V. Khadkikar, "An Improved PQ Zeta Converter with Reduced Switch Voltage Stress for Electric Vehicle Battery Charger*," IEEE Energy Conversion Congress and Exposition (ECCE)*, pp- 858-863, Oct 2020.

[10] B. Singh and R. Kushwaha, "A PFC Based EV Battery Charger Using a Bridgeless Isolated SEPIC Converter," in *IEEE Transactions on Industry Applications*, vol. 56, no. 1, pp. 477-487, Jan.-Feb. 2020

[11] A. Kumar, U. Sharma and B. Singh, "PMBLDC Motor Based Ceiling Fan Using an Isolated PFC Zeta Converter*," IEEE International Conference on Power Electronics, Smart Grid and Renewable Energy* (PESGRE2020), pp. 1-6, Jan 2020.

[12] K. V. Ravi Kishore, B. F. Wang, K. N. Kumar and P. L. So, "A new ZVS full-bridge DC-DC converter for battery charging with reduced losses over full-load range*," Annual IEEE India Conference (INDICON)*, pp. 1-6, Dec 2015.

[13] Nagesha C, A. Sreedevi and M. Gopal, "Simulation and hardware implementation of 24 watt multiple output Flyback converter," *International Conference on Power and Advanced Control Engineering (ICPACE)"*, Bengaluru, pp. 366-370, August 2015.

[14] A. Jha and B. Singh, "Portable Battery Charger for Electric Vehicles," *International Conference on Sustainable Energy and Future Electric Transportation (SEFET),* pp. 1-6, Jan 2021.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:40:15 UTC from IEEE Xplore. Restrictions apply.

2021 Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 28-29, 2021

A Neural Network Based Image Steganography Method using Cyclic Chaos and Integer Wavelet Transform

Manoj Kumar
Department of Computer Science
Delhi Technological University
New Delhi, India
mkumarg@dce.ac.in

Tahera Hussaini
Department of Computer Science
Delhi Technological University
New Delhi, India
thr.husaini@gmail.com

***Abstract*--- Steganography is the technique of hiding confidential information inside a cover in order to deceive attackers. Neural Networks (NN) become very popular and effective to train a model and get accurate result. In this article we used unsupervised NN to select and train model for cover image selection. Chaotic structures and transformation techniques like the integer wavelet transform in Steganography have good result due to minimum data loss. In this paper, we propose a cyclic chaos algorithm that produces Pseudorandom Number Generator (PRNG), use seeds for embedding purpose. We also used the Integer Wavelet Transform (IWT) to minimize data loss and return a high-quality encoded image. Our proposed system can embed colored image. Furthermore, evident and easily understandable as well as executable. We found that our approach outperforms other recent steganography approaches in terms of improved embedding time, reduced complexity, and improved decoded image quality with increased hiding capability.**

Keywords— Steganography, neural network, cyclic chaos, LSB embedding, IWT.

I. INTRODUCTION

Steganography is an information security technique used to mask data so that human eyes cannot detect stego-data. No one can determine whether it's stego-data other than authorized and confirmed user. Steganography has a long tradition dating back to ancient Greece. When it was known as steganographia. Steganographia mainly consists of Steganos, meaning concealed or secret, and Graphia, meaning drawing. In either case, Steganography for ancient Greece conveys the essence of writing in a hidden and disguised style. Johannes Trithemius recorded the first steganographia in Latin in the late 1490s. In his book, he explored Steganography and cryptography. In addition, hiding data was widely used in the Second World War by German [20, 21, 22].

There are some misperceptions among cryptography and Steganography, since both techniques are used to protect data. Both are professionally distinct and developed for various purposes. In cryptography, the data is available but in a distinct and unfamiliar format that draws interest. Cryptography aims to encrypt data despite of its format. Cryptanalysis lookups for plain text or private keys [20].

In Steganography, on the other hand, secret data is covered within another data, and the presence of secret data is preserved. The main goal of steganography is to prevent unauthorized access to alter and recover the secret messages. Since the manner in which data has been obscured is undetectable by the naked eye. Steganalysis attempts to locate a contact path, complete steganography process shown in fig. 1 [20, 21].

Let's talk about the steganography method. While various writers and researchers had given their thought. We'll try to figure out the major points that appear in most articles. There are some stenographic terms that we need to know for proper comprehension.

- Secret data
- Cover or host
- Embedding/encoding
- Extraction/decoding

Secret data is a private data with confidential information that has to be concealed that needs to transfer in a secure channel or media.

Cover data which covers the payload message and mask it in the background using any steganography-embedding algorithm. The technique of embedding is a process in that the secret message in cover data have to be covered. We'll need two input data called as secret data and cover data. The result of steganography embedding algorithm is stego-data. Stego-data is the embedded form of secret data and cover data. In the end, extraction is the opposite of embedding where we retrieve secret data from stego data.

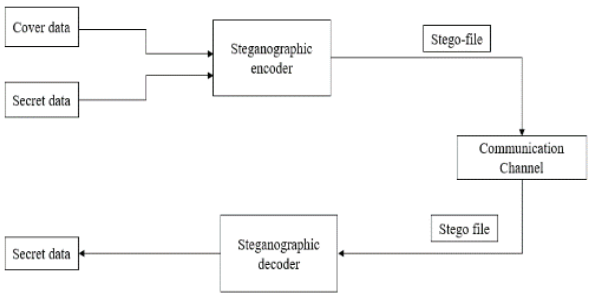


Fig. 1. Steganography process

Different Steganographic methods were used relying on the nature of cover data:

- Plain Text Steganography
- Audio Steganography
- Video Steganography
- Digital Image Steganography

2021 Asian Conference on Innovation in Technology (ASIANCON) | 978-1-7281-8402-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ASIANCON51346.2021.9544831

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:49:32 UTC from IEEE Xplore. Restrictions apply.

Out of these methods our attention is on image steganography. Image steganography has a comparatively large number of bits to hide secret message. It can perform by the frequency transformation domain and the spatial domain.

In the frequency domain either bits of the cover image first translate into another domain or, the secret data is encoded in the cover image's coefficient values. For instance, Discrete Wavelet, Discrete Cosine, and Discrete Fourier Transform, respectively, DWT, DCT, and DCF; additionally, Integer Wavelet Transform (IWT)[1].

DWT is mainly used to disintegrate a pixel in four sub bands (LL, LH, HH, and HL; which will be explored more in depth). It's better than DFC since apart from the frequency information, it offers timing details. When compared to DCT, it offers more compression energy [25, 26].DCT is a technique that splits an image into different parts. The signal is often converted from space to frequency domain. DCT offers low, high, and middle-frequency components like DFT, where it generates superior energy than DFT [24]. Furthermore, DFT converts the spatial domain to the frequency field and this is a framework for the image to break down into its sinuses and cosines [25].The strength of IWT over DWT is to provide integer values, DWT converts bits to double values, while IWT gives exact integer value corresponding to bits. IWT can transform bits back to the near original format during decoding. Although, DWT has degraded the decoded image quality [26]. Authors of [2] used IWT; embedded secret image using cover image coefficients to boost PSNR values.

The bits of secret data, on the other hand, are encoded explicitly into pixel values of cover media in the spatial field [1]. For instance, Least Significant Bits (LSB), this algorithm codes secret data bits in the cover image's least significant bits [27, 28, 29]. [3] Suggests using k least significant bits (K-LSB) to encode a secret image using the desired K into the image's LSB.Whereas the frequency domain (DCT) is more stable and efficient, the spatial domain (LSB) has more potential for data hiding, is a sampler, and is easier to implement.

In section 2 we covered literature studies. Our methods are explained in section 3. The result and comparison in the fourth section are noticed in strengthening our point and we concluded our concept.

II. LITERATURE REVIEW

In [8], the authors proposed a free cryptic LSB steganography using cyclic chaos. The algorithm compared two other efficient steganography methods with better stego-image quality and more hiding capability in terms of PSNR. Also, they evaluate their algorithm against attacks, including chi-square, visual, byte, and RS attacks.

In this article [9], the authors tried to send data over an insecure network in a secure manner. They used a matrix method to conceal confidential data in a colorful image. The algorithm first determined the YCbCr value of the color image, and the parts of each image were placed in the block. Although these secret data are saved and the Y matrix pattern is produced with the components Cb and Cr. The findings revealed that when stego-image quality is taken into account, the model can hold hidden data more effectively.

Hiroshi and Qiangfu discovered third-party gap insecurity that allows them to detect and extract hidden messages. Consequently, they recommend using GAN (Generative Adversarial Network) to uniquely cover each piece of hidden data. There have been two neural networks used to train GANs: the first requires a generator, which generates a virtual image. The second requires a discriminator that evaluates and checks the neutrality of the generated image. It suggested a steganography method using generators and discriminants to provide an infinite natural coverage. While removing suspicion through discriminatory controls [10]. In an attempt to improve the PSNR, [11] IWT has used the secret image to be inserted into the cover photo.

The LSB is used to embed the IWT coefficient values for text into image encoding. In order to help analyze invisibility and statistical distortion of payload and cover images, the authors used MATLAB to lead to effective and use the PSNR and Mean Square Error (MSE) to evaluate. Besides, NormalisedCross-Correlation (NCC) was used for the degree of similarity and closeness of both images. They have demonstrated that the IWT and LSB techniques are more secure, invisible, and have a higher secret text capability, as demonstrated by the PSNR and NCC values [12].

The rectangle prediction error distribution proposed by [13] was used to hide data using histogram shifting. The highest bin was set to 0 and the other bins were arranged symmetrically on the right and left sides. The first peak bins move to the right and then left to leave some space for secret data bits to be embedded. This will improve the capability of embedding while preserving the visual quality of hidden data.

In [14] the authors explain how to hiding secret data with embedded additional data; For example, the message authentication, tag, time stamp, and origin are all embedded directly with a JPEG encrypted image based on zero-run invariant pairs. This algorithm's encryption and embedding process is much easier and has less computational complexity and higher embedding capacity.

The key aim of using this 3-dimensional chaotic map is to improve safety, as suggested in [19]. Because of such a feature, it is resistant to salt and pepper attacks.

In their research, Shaikh and Thanikaiselvan[31] conducted shifting histograms by considering threshold-based transformation domain revisable data hiding (RDH). They used many integer transformations that were very effective, e.g., 9/7, 5/11, 5/3, 13/7-C... They used a flag array as an overflow for the averaging. They compared their performance with other RDH methods, including spatial, DWT-based, and other RDH methods. Their results indicate which smaller Haar values, such as 2/10 and 2/6, result in improved PSNR at the bottom bit, even as enhanced scaling factors, including 9/7 and 13/7-T, result in lower PSNR, also regarded as medium performance. The authors used various wavelets to describe the connection of stego-image quality with essential capacity.

Kumar and Biswas [32] used a genetic algorithm for frequency-domain message embedding. This algorithm can withstand brutal attacks without being harmed. They used multiple random bits to embed the secret message to enhance payload and security. The use of a hash function based on random pixel selection, Genetic Algorithm (GA), encrypting the secret message, and domain transformation are all novel

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:49:32 UTC from IEEE Xplore. Restrictions apply.

aspects of this paper. Confusion Matrix, StirMark 4.0, and other steganalysis tools are used to evaluate the proposed work.

Reference [33] proposes a statistical framework in which the stego message and cover are modeled as a multivariate Gaussian random variable. By introducing this algorithm, they enhanced error detection and extended it to cost-based Steganography. This methodology for a continuous concealing message remains the same for any encoding scenario.

Zhang, Peng, and Min Long proposed DCT and LDA classification in Steganography [34]. DCT, Latent Dirichlet Allocation (LDA), and a coverless image are commonly used in this algorithm to combat steganalysis. LDA is used first for image classification, and then 8x8 block DCT is applied to the result of the classification. Then, between adjacent blocks, a robust sequence feature is implemented. Eventually, an inverted index is created, including the sequence identifier, image path, and location coordinator. The original image is not altered in the proposed method.

Image steganography using FC-DenseNet is proposed to enhance traditional Steganography's low capacity, which is Fully Convolutional Dense Connection Network (FC-DenseNet) [35]. This methodology has several features; for example, the authors proposed cascading the secret and carrier images for better image reconstruction. They used FC-DenseNet to avoid gradient explosion and dissipation. This study's findings show a high PSNR and SSIM value, as well as a large capacity for hiding messages.

III. PROPOSED SYSTEM

The art of dissimulating the encoded message is steganography. As we all know that image steganography means hiding an image in another image. In this study, we introduced a technique using NN for training and cover image selection, cyclic chaos for an exhaustive range of keys for security purpose, the IWT for decreasing data loss, and the LSB for embedding. The IWT is a method for transforming data; it maps data into an integer data set, which corresponds to the exact values and sizes. At the same time, the decoding result is going to be much closer to the original secret image. The LSB method is a traditional method that is still widely used in data hiding. The most significant bits of hidden image (MSB) is encoded into the cover image's LSB. Our focus is on enhancing the capacity of embedding, security, decreasing process time, and secret images' invisibility in embedded images. We described the Neural Network (NN), Chaos function and the IWT method for having a suitable context in the following paragraphs.

A. Neural Network

The artificial neural network has been developed based on the human brain. It can have thousands of neurons to operate properly. In NN connection between neurons have weight, which positive value shows excitatory connection. A simple NN consist of input, hidden, and output layers. NN is used in this article for selecting the cover image. In this article, we are using Self Organizing Map (SOM) for cover selection. SOM is an unsupervised kind of NN, which means a human has no mediation during learning.

B. Chaos function:

Chaos is a mathematical theory due to its sensitivity to the initial values and changes we propose. This property increases security because the keys' domain is large and exhaustive, finding it challenging for attackers to find the keys. For extra protection, the Chaos function generates seeds and uses a pseudorandom number generator (PRNG).

As shown in equation (1), the chaos equation is a logistic map:

$$Xk+1= \mu xk(1-xk) \tag{1}$$

Since there are a few changes in the chaotic parameters, there will be a significant difference in the result. [1] denotes the number of N-dimensional chaotic functions in K dimensions.

$$Xin+1 = f(Xin, \lambda i)n \tag{2}$$

Where Xi = (xi1, xi2,.....xik) e Rk in Eq. (2) shows cyclical chaotic function[4]. Mostly, the three-dimensional signal is used in the cyclic chaos system as follows:

$$\begin{aligned} xn+1 &= \lambda 1\ xn - x3n - \gamma \mid yn \mid m\ xn \\ yn+1 &= \lambda 2\ yn - y3n - \gamma \mid zn \mid m\ yn \\ zn+1 &= \lambda 3\ zn - z3n - \gamma \mid xn \mid m\ zn \end{aligned} \tag{3}$$

Where x0, y0, and z0 are initial conditions that will take any desired value within the range [-1.6, 1.6] and shared as equation keys (3). λ1, λ2, and λ3 are internal parameters, m is the control parameter that takes values between 0 and 0.5, and y will be between [2.7 and 3.0]. Because each parameter has a wide range of values, the key [5] is quite large.

C. Integer Wavelet Transform (IWT):

Relevant information from different types of data is extracted using wavelets, images, audio, and video can be used as files. Multiple wavelets function comes together to compact and retrieve data to minimize data loss [6]. The priority of IWT over DWT is: in DWT, it maps floating-point data set that after extracting the original data we face a large data loss. Also, the LL (Low – Low) sub-band in DWT is reduced while in IWT it is similar to the small size of the original data and very close to it.

In addition, with the comparison of Continuous Wavelet Transform (CWT) and DWT, IWT is not only memory-efficient and takes less computational time but also has good performance to prevent loss of data compression [30]. The IWT HH, HL, LH, and LL coefficients, which also represent the diagonal, horizontal, vertical, and approximation, are very close to the original data's dynamic range, making our job much easier in the encoding [7,12], as fig. 2 shows all IWT sub bands of Mandrill.

Embedding and Extraction process briefly described as follow.

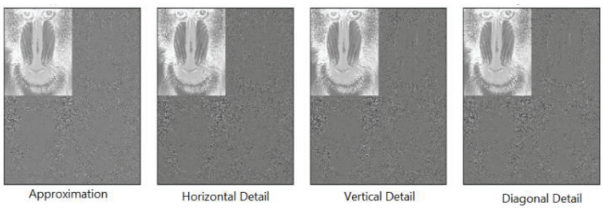


Fig. 2. IWT coefficients of Mandrill

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:49:32 UTC from IEEE Xplore. Restrictions apply.

1) Embedding

Here we describe how to embed hidden image in the cover image as follows.

Step1: Load the secret image, quantize it to the corresponding palette, and save the palette length. Next, we calculate the two-dimensional IWT of the quantified image and apply some addition to the [0, 511] range of the coefficient. We defined annotation and symbols for decoding and converting 8 bits to 9 bits. Concatenate the palette and annotation of these coefficients to the bases to hide in order. As a result, extract 3 bits from the base to hide with a cover image. Now, this secret image is ready to be encoded.

Step 2: Histogram value of cover image becomes input for SOM. SOM classify the histogram intensity as ([0001],[0010],[0100],[1000]) and apply backpropagation to get desired output. Next, cyclic chaos is used to produce seeds and PRNG. Then, generate random columns and rows. For each column, apply cyclic chaos and get seed and PRNG. Sequentially create random pixel columns. Finally, use LSB to embed a secret image.

Embedding algorithm: has two steps

Step 1: Prepare the secret image

1. Load the secret image
2. Quantize it and get its palette
3. Convert quantized image to IWT
4. Declare annotation (flag) and symbols to help extract easily

Step2: Prepare the cover image

1. Cover image selection using NN
2. Apply cyclic chaos algorithm

3.1 Assign chaos parameters to get seed and PRNG

3.2 Generate random rows and columns

3.2.1 Extract three bit integer

3.2.2 Take two lowest bit and compute exclusive-OR of them

$A = a1 \oplus a2 \oplus b1$

$B = b1 \oplus b2 \oplus c1$

$C = c1 \oplus c2 \oplus a1$

3.2.3 Apply LSB embedding

4. Save coded image

2) Extraction

Load the stego image; using cyclic chaos parameters we can obtain seeds which will further generate PRNG. After that, we will create a random column and row with the help of PRNG and the seeds. Next, extract 3 bits from the generated column of random pixels and concatenate 3x3 bits for 9 bits.

Here you can subtract 255 to the contained columns and lines of data if symbols and annotations are correct; also the size of hidden data reached. Apply reverse of IWT, remove the palette from the other bits. The decoded image is generated by recombining the image pallet. Save the image that was decoded.

Extraction algorithm is as follow:

1. Load coded image
2. Use chaotic algorithm

2.1 declare embedding base

2.2 set flag and annotation

2.3 get chaotic coefficients to generate seed and PRNG

2.4 generate random column and row using PRNG

2.5 take two lowest bit and compute exclusive or of them

$A = a1 \oplus a2 \oplus b1$

$B = b1 \oplus b2 \oplus c1$

$C = c1 \oplus c2 \oplus a1$

3. Compute inverse of IWT
4. Get palette from quantized image
5. Convert to RBG
6. Save reconstructed image

IV. RESULT AND COMPARISON

We used the USC SIPI, CALTECH, and NRC datasets [18] in this article to cover 512x512 images and 256x256 secret images in various formats. We assessed our results using PSNR, MSE, and the Structural Similarity Index (SSIM). Though higher PSNR values show improved visual quality with less MSE, higher PSNR implies higher visual quality, and lower MSE means minimum errors on the stego image. A lower PSNR value indicates greater visual quality. Furthermore, the SSIM result specifies the cover image's and stego image's similarity structure.

For comparing compressed images, MSE and PSNR have been defined. MSE is the total error between the original image and the stego image. PSNR is responsible for the maximum error representation. To obtain PSNR, we must first evaluate the worth of the MSE. The accuracy of the initial hidden image is decreased and the deterioration quality with SSIM can be identified if a hidden image is taken from a stego image.

These algorithms are used to enhance our idea. As we proposed Cyclic Chaos and IWT, we compare our frameworks with old ones separately. The MSE, PSNR and SSIM mathematical formulas are as follows.

$$\text{MSE}= \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (I(i,j) - (I'(i,j)) \quad (4)$$

$$\text{PSNR}= 10 \log_{10}(\frac{255x255}{MSE}) \quad (5)$$

In equ. (4) (5) the cover picture and stego image are I and I' and MxN is the picture size, respectively (15).

$$\text{SSIM}= \frac{(2\mu I \mu I' + k1)(2\sigma I,I' + k2)}{(\mu I2 + \mu I'2 + k1)(\sigma I,2 + \sigma I'2 + k2)} \quad (6)$$

I, I', I2, and I'2 are variants of the cover and stego image in equation (6). I and I' demonstrate the covariance between cover and stego images. In addition, k1 and k2 are used to avoid exceptions such as zero error divide. SSIM takes value in the [0, 1] range, 0 is the worst, and 1, the better [16] range.

A comparison of PSNR with research of Mamatha et al reveals that our procedure is improved by 11, means that the image recovered has better quality. In addition, the accuracy of the picture retrieved as Table 1, was increased for SSIM value by 0.10.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:49:32 UTC from IEEE Xplore. Restrictions apply.

The proposed and the Valandar's et.al approach were compared in Table2, respectively, in PSNR, SSIM, and capacity.

We used table 3 and table 4 to compare our algorithm to an existing chaos algorithm. The proposed algorithm produced better value and better image quality by analyzing and comparing all existing method.

TABLE I. PSNR AND SSIM VALUE COMPARISON OF OURS AND (17) REGARDING TO IWT IMPLEMENTATION

Color Image	PSNR		SSIM	
	Mamatha et al	Proposed	Mamatha et al	Proposed
Lena	43.73	55.0264	0.8914	0.9982
Baboon	41.35	55.0166	0.9610	0.9994
Boat	42.84	54.0861	0.9280	0.9987

TABLE II.

Color image	Mandrill			Lena		
Gray scale	Aziz et al	Ours		Aziz et al	Ours	
	PSNR	PSNR	MSE	PSNR	PSNR	MSE
Airplane	44.301	54.0962	0.2531	44.235	54.9535	0.2532
Lena	44.213	54.1014	0.2528	44.191	54.1001	0.2529
Boat	44.109	55.0166	0.2048	44.081	55.0264	0.2043

TABLE III.

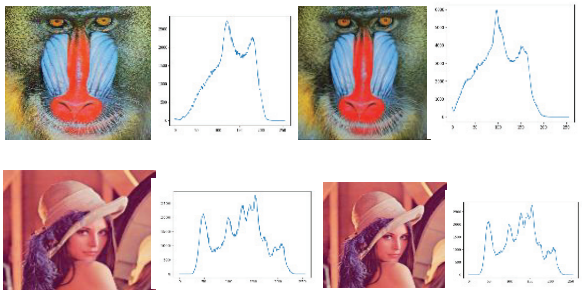
Color image	Airplane			Peppers		
Gray scale	(8)	Ours		(8)	Ours	
	PSNR	PSNR	MSE	PSNR	PSNR	MSE
Airplane	44.421	54.0994	0.2530	44.221	53.8863	0.2657
Lena	39.159	54.1067	0.2525	44.403	53.8990	0.2649
Boat	44.125	55.0263	0.2043	44.277	54.7853	0.2160

Table .2 and .3, PSNR values of proposed and Aziz et al (cyclic chaos comparison)

TABLE IV. COMPARISON OF PSNR AND SSIM

Reference with capacity	Valandar et al, with capacity 77244 bits		Proposed, with capacity 274896 bits	
Color image	PSNR	SSIM	PSNR	SSIM
Lena	51.376	0.9984	54.9535	0.9982
Baboon	52.422	0.9991	54.0937	0.9994
Pepper	52.390	0.9986	53.8701	0.9980
Airplane	51.984	0.9980	54.1016	0.9977

Histogram comparison of the cover image and stego image are as fig. 3, which histograms changes are not easily detectable.



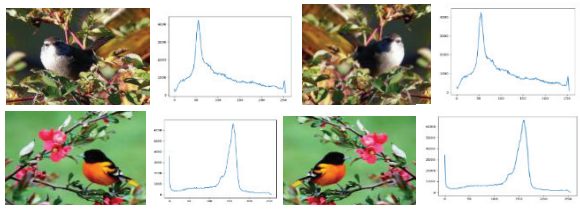


Fig. 3. Cover image, histogram of cover, stego image, and histogram of stego image which are respectively from USC SIPI, CALTECH and NRC datasets

V. CONCLUSION

Data hiding or steganography has developed in the last years. Furthermore steganography became exciting research area. We investigated steganography and finally we proposed an easy method to implements, by using NN for best cover image selection, IWT, and cyclic chaos algorithm for lower computational complexity and minimum data loss. Moreover, hiding ability, increased visual quality, and safety. Histogram of cover images and embedded images also added, which show a satisfying result. In order to confirm our hypothesis, the evaluation result demonstrate our superior theory. We have compared our technique with three powerful methods using, PSNR, SSIM, and MSE values.

REFERENCES

[1] R. Munir, "Chaos-based modified “EzStego” algorithm for improving security of message hiding in GIFimage," 2015 International Conference on Computer, Control, Informatics and its Applications (IC3INA),Bandung, 2015

[2] N. Jothy and S. Anusuyya, "A secure color image steganography using integer wavelet transform," 2016 10th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, 2016

[3] O. Elharrouss, N. Almaadeed and S. Al-Maadeed, "An image steganography approach based on k-leastsignificant bits (k-LSB)," 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), Doha, Qatar, 2020

[4] M. S. Subhedar, V. H. Mankar, Current status and key issues in image steganography: A survey, 2014

[5] N. I.Wu, K. C.Wu, C. M.Wang, Exploring pixel-value differencing and base decomposition for low distortion data embedding, Applied Soft Computing 12 (2) (2012)

[6] Huang J., Kim H. J., Xiang. S., (2008), ‘Audio watermarking robust against time-scale modification and MP3 compression,’ IEEE Journal on Signal Processing, Vol.88, No.10, pp. 2372–2387.

[7] Emmanuel S., Kankanhalli M. S., Wang. L., (2010), ‘EMD and psychoacoustic model based watermarking for audio,’ IEEE International Conference on Multimedia and Expo, pp. 1427–1432.

[8] Aziz, M., Tayarani-N, M.H. & Afsar, M. A cycling chaos-based cryptic-free algorithm for image steganography. Nonlinear Dyn 80, 1271–1290 (2015).

[9] M. Mowafi, O. Oudat, E. Taqieddin and O. Banimelhem, "Image Steganography using YCbCr Color Space and Matrix Pattern," 2019 2nd International Conference on Signal Processing and Information Security (ICSPIS),Dubai, United Arab Emirates, 2019

[10] H. Naito and Q. Zhao, "A New Steganography Method Based on Generative Adversarial Networks," 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), Morioka, Japan, 2019

[11] N. Jothy and S. Anusuyya, "A secure color image steganography using integer wavelet transform," 2016 10th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, 2016

[12] E. Emad, A. Safey, A. Refaat, Z. Osama, E. Sayed and E. Mohamed, "A secure image steganography algorithm based on least significant bit and integer wavelet transform," in Journal of

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:49:32 UTC from IEEE Xplore. Restrictions apply.

Systems Engineering and Electronics,vol. 29, no. 3, pp. 639-649, June 2018

[13] Weiqing Wang, Junyong Ye, Tongqing Wang, Weifu Wang, A high capacity reversible data-hiding scheme based on right-left shift, Signal Processing, Volume 150, 2018

[14] J. He, J. Chen, W. Luo, S. Tang and J. Huang, "A Novel High-Capacity Reversible Data Hiding Scheme for Encrypted JPEG Bitstreams," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no.12, pp. 3501-3515, Dec. 2019

[15] M.N. Wu, C.-C. Lin and C.-C. Chang, "An embedding technique based upon block prediction", J. Syst.Software, vol. 81, no. 9, pp. 1505-1516, 2008.

[16] W.C. Du and W.-J. Hsu, "Adaptive data hiding based on VQ compressed images", Proc. Inst. Elect. Eng. Vision Image and Signal Processing, vol. 150, no. 4, pp. 233-238, Aug. 2003.)

[17] G. Mamatha and A. Shaik, "Reversible Data Hiding Based on Integer Wavelet Transforms and Prediction Error Expansion," 2018 International Conference on Computational Science and Computational Intelligence (CSCI),Las Vegas, NV, USA, 2018

[18] Volume 3: Miscellaneous, Ming Hsieb Department of Electrical Engineering, Signal and Image Processing Institute

[19] Valandar, M.Y., Barani, M.J., Ayubi, P. et al. An integer wavelet transform image steganography method basedon 3D sine chaotic map. Multimed Tools Appl 78, 9971–9989 (2019).

[20] Y. Ke, J. Liu, M. Zhang, T. Su and X. Yang, "Steganography Security: Principle and Practice," in IEEE Access,vol. 6, pp. 73009-73022, 2018

[21] Steganographia, Reference: Peniarth MS 423D, Llyfrgell Genedlaethol Cymru- The National Library of Wales,Aberystwyth, Ceredigion SY23 3BU

[22] J. Qin, Y. Luo, X. Xiang, Y. Tan and H. Huang, "Coverless Image Steganography: A Survey," in IEEE Access,vol. 7, pp. 171372-171394, 2019

[23] M.Anusha, K. N. Bhanu and D. Divyashree, "Secured Communication of Text and Audio using Image Steganography," 2020 International Conference on Electronics and Sustainable Communication System (ICESC), Coimbatore, India, 2020

[24] Biswas, R., Bandyapadhay, S.K. Random selection based GA optimization in 2D-DCT domain color image steganography. Multimed Tools Appl 79, 7101–7120 (2020)

[25] Kumar, V., Kumar, D. A modified DWT-based image steganography technique. Multimed Tools Appl 77, 13279–13308 (2018)

[26] E. Emad, A. Safey, A. Refaat, Z. Osama, E. Sayed and E. Mohamed, "A secure image steganography algorithm based on least significant bit and integer wavelet transform," in Journal of Systems Engineering and Electronics, vol. 29, no. 3, pp. 639-649, June 2018

[27] C. K. Chan and L. M. Cheng, "Hiding data in images by simple LSB substitution", Pattern Recognit., vol. 37, no. 3, pp. 469-474, 2004.

[28] R. R. Ch, "The Process of Encoding and Decoding of Image Steganography using LSB Algorithm", Ijcset, vol.2, no. 11, pp. 1488-1492, 2012.

[29] R. S. Ganesh, V. Nagaraj, S. A. Sivakumar and B. M. Shankar, "An Intelligent and Hybrid Method of Combining Spatial and Frequency Representation for Digital Image Steganography," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 2020

[30] LABVIEW 2010 advanced signal processing ,Integer Wavelet Transform (Advanced Signal Processing Toolkit), zone.ni.com, Jun 2010

[31] Ahmad Shaik, V. Thanikaiselvan, Comparative analysis of integer wavelet transforms in reversible data hidingusing threshold based histogram modification, Journal of King Saud University - Computer and Information Sciences, 2018

[32] Biswas, R., Bandyapadhay, S.K. Random selection based GA optimization in 2D-DCT domain color image steganography. Multimed Tools Appl 79, 7101–7120 (2020)

[33] M. Sharifzadeh, M. Aloraini and D. Schonfeld, "Adaptive Batch Size Image Merging Steganography and Quantized Gaussian Image Steganography," in IEEE Transactions on Information Forensics and Security, vol.15, pp. 867-879, 2020

[34] X. Zhang, F. Peng and M. Long, "Robust Coverless Image Steganography Based on DCT and LDA Topic Classification," in IEEE Transactions on Multimedia, vol. 20, no. 12, pp. 3223-3238, Dec. 2018

[35] X. Duan et al., "High-Capacity Image Steganography Based on Improved FC-DenseNet," in IEEE Access, vol.8, pp. 170174-170182, 2020

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:49:32 UTC from IEEE Xplore. Restrictions apply.

Wireless Personal Communications
https://doi.org/10.1007/s11277-021-09176-7

A Review of Blockchain-Based Applications and Challenges

Pratima Sharma[1] · **Rajni Jindal**[1] · **Malaya Dutta Borah**[2]

Accepted: 16 September 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract
The deployment of blockchain technologies for multiple use cases has been widely investigated in the academic and business sectors over the last few years. The blockchain model has attained considerable attention due to its decentralized, persistent, anonymous, and auditable features. This review does a comprehensive literature analysis of broad blockchain implementations across several domains. The study's key objective is to present a thorough overview of the widespread deployments of blockchain technology and demonstrate how particular aspects of this innovative technology can change the business community's activities. Several papers are addressing the feasibility of using blockchain technologies in various fields. However, we include a description of blockchain concepts and comparative analysis of the application in six main fields: the Internet of Things, artificial intelligence, supply chain, cloud, healthcare, and multimedia networks. For each area, we analyze in-depth the approaches proposed by the research community and industry. This paper also discussed the different problems involved in each area. Finally, we explore the critical issues needed for the broader implementation of blockchain technologies in these sensitive areas.

Keywords Blockchain technology · Applications · IoT · Healthcare · Cloud storage · Supply chain · Multimedia

1 Introduction

Nearly a decade ago, the anonymous person behind Bitcoin, Satoshi Nakamoto, explained how blockchain technology, a decentralized peer-to-peer connected system, can be utilized to address the issue of preserving the transactions sequence and avoiding the problem of double-spending attack [1]. Bitcoin combines the transactions in a system of limited capacity called blocks that share the identical timestamp. The network entities can connect the blocks in sequential order to each other, with each block containing the preceding block hash to create a chain of blocks [2]. Therefore the framework of the blockchain continues to provide a secure and correct record of all transactions. Blockchain presented severe

✉ Pratima Sharma
sharmapratima9818@gmail.com

[1] Delhi Technological University, Delhi, India

[2] National Institute of Technology Silchar, Silchar, Assam, India

disruptions to traditional procedures as applications and transactions that need centralized structures or trusted third parties to confirm them can now work with the same level of certainty in a decentralized manner. The blockchain architecture can support inherent features such as transparency, reliability, auditability, and security [3, 4]. A blockchain is a decentralized registry structured in ordered blocks, where the published blocks are unchangeable. It is appropriate in the banking industry since banks can collaborate using the single blockchain structure and submit users' transactions. Thus, blockchain allows the verification of transactions beyond transparency. Organizations collaborate in this technology as they view the capability to decentralize their architectures and minimize their transaction costs as they inherently turn to be secure, reliable, and transparent.

The cryptocurrencies represent the significance of blockchain, which is currently more than 1900 and still growing [5]. Due to the heterogeneity of cryptocurrency implementations, such a growth rate could eventually cause interoperability [6] [7]. Also, the ecosystem is changing rapidly, as blockchain is utilized in many areas besides cryptocurrency, with smart contracts play a crucial part. Szabo gave the smart contract concept in 1994 as: "a computerized transaction procedure executing the terms of the agreement" [8]. It enables the transformation of the contract agreements into embeddable code [9], thereby minimizing external engagement and threat. Therefore, a smart contract is a contract between participants who do not trust each other, automatically enforces the contract terms. Smart contracts are codes running in a decentralized manner within the blockchain context and saved in the blockchain [4] without depending on any third authority. In particular, blockchain supports smart contracts that allow complicated processes and connections to create a new paradigm for many applications.

Thus blockchain technology has become increasingly relevant [10]. Nearly 33 percent of executives are already involved or consider blockchain technology for implementation [11] [12]. Analyzers and researchers are now aware of the latest technology's potential and research the growing applications across various industries [4]. Depending on the target audience, three types of blockchain can be identified [10]: blockchain 1.0 defines applications that allow digital crypto-currency transactions; blockchain 2.0 covers smart contracts and a range of services expanding outside crypto-currency transactions; and blockchain 3.0 covers applications in fields outside the previous two, like IoT, health, supply chain, and government.

Although many studies on blockchain technologies [13], we contend that there is little focus on surveys based on blockchain-enabled applications. Blockchain applications are not completely explained in many review papers [14]. Indeed, some studies aimed at the blockchain application and include the privacy and security challenges faced by the applications. These reviews consider only a few applications [16] and mainly focused on the security aspect [15] [18]. Other studies lack the technical details of the blockchain applications [17]. Some surveys also consider the particular area of blockchain and include the details of decentralized IoT applications [21] and big data management [20]. Other reviews consider the blockchain's privacy problems [22] [23] and their potential for trust in service systems [24] and P2P networks [25]. In [26] and [27], some technical details of blockchain structure such as its consensus methods [26], smart contract vulnerabilities [27], and other features like size, bandwidth, usability, data integrity, and scalability were also studied. Furthermore, other studies, such as [28–32], are based on the currency dimension of blockchain and the provided protection and security concepts.

Table 1 presents a comparison of existing literature reviews and survey papers. The existing studies lack a detailed and comprehensive analysis of the state-of-the-art applications allowed by blockchain and its challenges. Hence, this is the motivation behind this

Table 1 Comparison of related work

Review papers	Year	Focus area	Remarks
[15]	2019	General study and privacy analysis	The survey provides an analysis of blockchain applications and discusses the associated security and privacy issues
[16]	2019	General study	Present an analysis of three blockchain-based applications: IoT, healthcare, and business
[17]	2018	General study	Conducts a systematic literature review of blockchain applications
[18]	2018	Privacy	Survey the privacy issues of blockchain technology
[19]	2018	Trust model	A systematic review of literature, areas of blockchain technology based on the trust model
[20]	2017	Big data	Review the blockchain solutions and challenges in the field of big data
[21]	2017	IoT security	Present the survey of IoT security and identify the blockchain solutions and challenges
Proposed Work	2020	General study and comparative analysis	Conducts a systematic literature review of six blockchain applications. The proposed work performs a comparative analysis of each application category-wise, identifies the challenges of blockchain architecture in each domain, and generates future trends

work. In particular, we review the current literature and explore the numerous benefits and problems of blockchain use. In fact, by addressing the following three questions, we attempt to resolve this: (1) How do blockchain-based applications evolve? (2) What are the different working areas of blockchain-based applications? (Categories identification). (3) What are the challenges of the blockchain architecture in different procedures/processes of applications? Which limitations are those? Our research leads to a comprehensive analysis of blockchain technology applications. Based on a systematic review approach, we outline the academic community's interest and define three main work fields. First, the description of blockchain technology across a wide variety of sectors. Then further category-wise analysis of blockchain-based applications, taking into account the various limitations posed by the applications. In the end, we are guiding researchers through a roadmap of promising research areas, challenges, and opportunities needed for future research. It is important to note that this study cannot be considered exhaustive since blockchain technology is continuously increasing.

The proposed work covers the following new contributions:

1. The authors provide a complete review of six areas of blockchain, including blockchain core research area, the IoT, healthcare, cloud storage, supply chain, and the latest research domains like AI and multimedia.
2. The proposed review includes the latest findings, articles, and documents in the areas mentioned above. It provides a systematic approach to studying the domain areas based on identified categories such as security and privacy, information storage, access control, consensus mechanism, and smart contract.
3. The various open challenges posed by the blockchain applications are thoroughly discussed, followed by analyzing recent advances, solutions, and future improvements.

The rest of this work is organized according to the following. A brief description of blockchain architecture shall be given in Sect. 2. The method followed for the systematic review of literature is highlighted in Sect. 3. The literature review descriptive analysis is presented in Sect. 4, while various blockchain applications are presented in Sect. 5. Related challenges and different lines of study are explored in Sect. 6. Section 7 highlights the recent advances and solutions for the open challenges. Finally, the paper concludes in Sect. 8.

2 Blockchain Overview

The developments in cryptography and distributed computing have introduced a modern computer technology called a blockchain. Blockchain is a distributed ledger that replicates and exchanges data through peer-to-peer (P2P) networks. Satoshi Nakamoto initially introduced blockchain, which created Bitcoin to directly trade digital currencies without third parties [33, 34]. Nakamoto has developed this paradigm of a network of nodes working to maintain a decentralized and secure database. As the name suggests, blockchain is an ordered list of blocks. By referencing the previous block's hash, each block is distinguished by the hash sequence and ties to the preceding block. The only anomaly is the first block, called the genesis block, which does not have the previous block's hash value [35]. Blockchain is the main backbone of cryptocurrencies. It can be regarded as both a technical breakthrough and financial advancement. It provides a solution to any problem, where

Table 2 Types of blockchain

	Public blockchain	Private blockchain	Consortium blockchain
Access	No access restrictions	Invitations only by network administrators	Restricted to selected members
Transact	Anyone can transact	Only designated individuals	Selected consortium members only
View	Anyone can view	Restricted	Restricted to selected members
Type	Large, decentralized, e.g., Bitcoin, Ethereum platform	Middle ground platform: record keeping and accounting	Participating companies equally involved in consensus, e.g., R3, Consensys

a trustworthy ledger is required in a decentralized setting, and there is no trust among the entities. Various literature reviews review the blockchain technology are application-specific like business applications [36], e-governance [37], healthcare [38], etc., but this paper reviews the six blockchain technology applications without being specific to particular applications, thus addressing its current trends, classifications, and challenges. Furthermore, blockchain technology is an information technology that can be used in software, business, and trade sectors [39]. The blockchain is viewed mainly as an accounting book or digital distributed database [40–45]. The architecture of the blockchain is illustrated in Fig. 1.

Blockchain technology is a distributed network that combines distributed data storage, cryptographic algorithms, and decentralized consensus mechanisms. Hence, this technology enables individuals to concur on a specific situation and record that ascension safely and undeniably in a different type of blockchain structure (Table 2) [46–48].

3 Research Methodology

This section presents scientific and systematic literature to give a transparent and reproducible review of applications based on the blockchain. The literature review presented by [49] has been suggested to understand their limitations. The presented methodological approach consists of the following essential steps:-

1. A review protocol has been developed by identifying the need for blockchain users and presented a review proposal.
2. Identification of various studies from the scientific journals by accessing their quality in terms of data extraction and synthesizing the data.
3. The results of the developed review protocol have been reported.

4 Research Studies

The primary research question has been addressed by initiating research in August 2019 without any time restrictions, and results were subsequently updated during March 2020. Google Scholar is considered the main database in which blockchain and associated terms have been searched in the main title block. Firstly, recent papers of 2019 have been

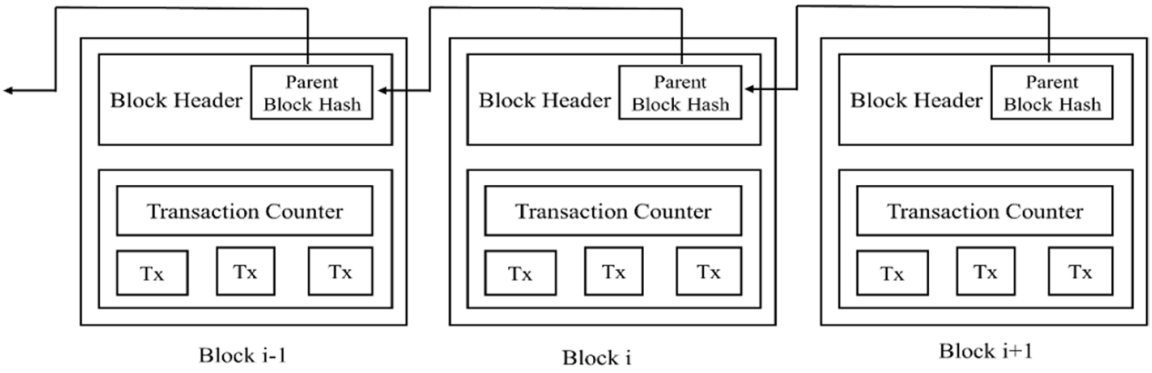


Fig. 1 Blockchain architecture

searched to determine the research gap. Additionally, referenced work in the searched article is also considered called a snowball effect. Consequently, grey literature was composed of unpublished work and reports commissioned through the public institutions identified through the electronic media. Other words, such as 'application' and 'blockchain challenges' was used to enhance the search procedure. The past grey literature in reports and policy draft reviews from the private and government sectors has been reviewed.

The strategy implemented to enhance the search paradigm is depicted in Fig. 2. Also, refinement studies from IEEE, Elsevier, and Springer are extensively described in the presented research. The abstract and conclusion of the past research have been studied, and when the abstract was not described, the full article has been retrieved as per its relevancy, such as m = 161 in the presented research. The articles having full text irrelevant to the research have been eliminated, such as m = 25. The articles having more citations and published in high-impact journals have been considered to enhance the research.

4.1 Assessment of Quality and its Evaluation

The study was conducted following the elimination criteria in which two exclusion principles were used to retrieve the meaningful data in the presented research. One was entirely based on the titles, and the other was based on the abstract and conclusion. The papers related to non-engineering concepts were eliminated. For instance, papers that discuss the ethical issues and economic aspects of the blockchain and cryptocurrencies were eliminated. Finally, the remaining papers from different domains of the blockchain have been included in the list. The selected articles were evaluated based on their quality and finally included in the introduction and literature section—the retrieval and elimination process of the research articles given in Table 3.

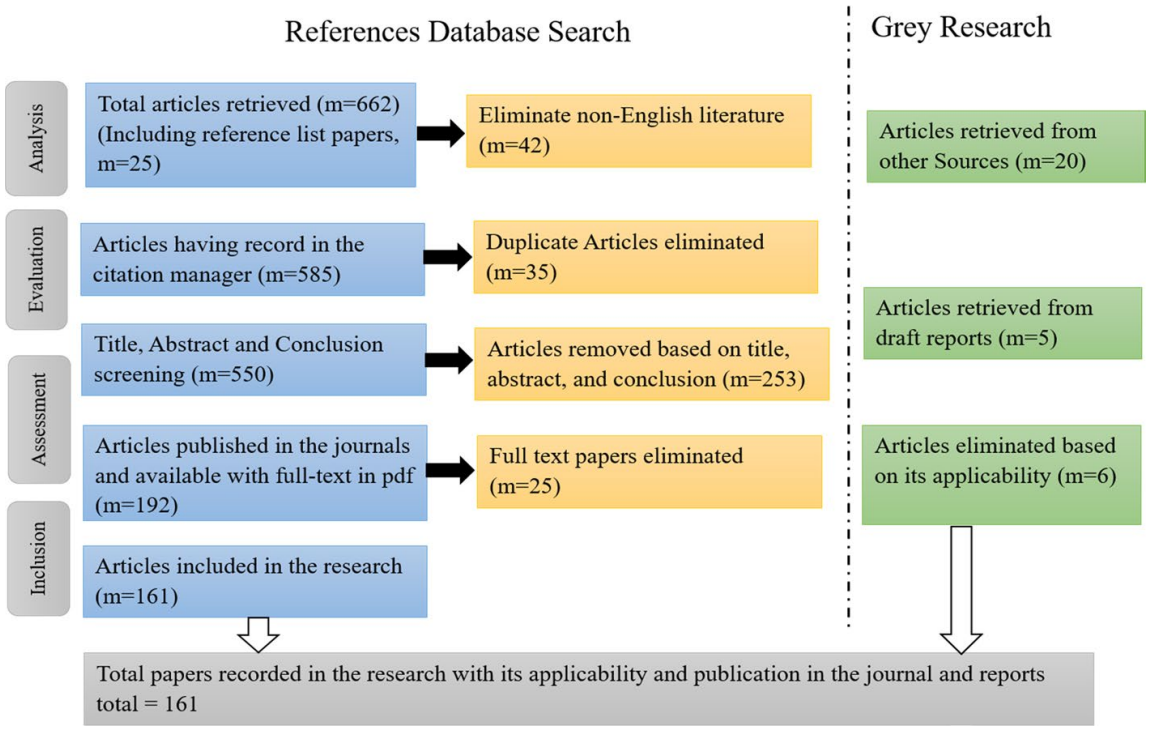


Fig. 2 Search mechanism

Table 3 Retrieval and elimination process

Selection strategy	Scientific journal papers	Grey research
Assessment	Articles related to blockchain in different domains (Journal papers, conference papers, reviews, surveys, book chapters, editorial notes, etc.)	Draft reports
Analysis	Cited in the citation manager	Reports related to blockchain are analyzed firmly
Evaluation	During abstract and title screening	
Elimination	During abstract and full-text screening	Generic reports having no relevant data eliminated
Inclusion	Remaining screened articles	

4.2 Data Synthetization and its Analysis

All the research papers and articles fitted in the inclusion criteria were analyzed using analysis software such as MAXQDA11. The thematic analysis was performed by the authors independently. Afterward, three groups were constructed in which previous review articles related to blockchain were placed in the first group, articles of blockchain, with its application placed in the second group. Finally, the third group consists of challenges faced by the researchers related to blockchain.

5 Descriptive Research Analysis

This section analyses the various research articles to understand the blockchain technology research trends worldwide with the proposed research analysis discussion.

5.1 Timeline of Blockchain Research

We have studied the various bibliometric analysis of blockchain technology to identify the current research status [161–165]. Figure 3 presents the year-wise distribution of publications to illustrate the proliferation of blockchain research in the last 5 years [162]. As shown in Fig. 3, the number of blockchain papers is increasing rapidly in recent years. As the number of published papers in blockchain increases, we can also see an increasing trend in citations year-wise [163–165]. We can see there are only 57 citations in 2016. This is followed by drastic changes in the citations numbers in 2017 and 2018, increasing from 556 to 1246 citations. Further, the number of citations from 2018 to 2019 increases at the double rate for the blockchain articles indexed by Web of Science (WoS) [162].

Similarly, in Fig. 4, we have analyzed the various domain-wise bibliometric studies to identify the current trends in blockchain technology applications [163–165]. According to the research study by Taylor et al. [161], IoT is the most trendy topic in blockchain technology, and most of the research is done in this area [163]. Then, blockchain in cloud storage and sharing is the second most popular topic, followed by healthcare, supply chain, AI,

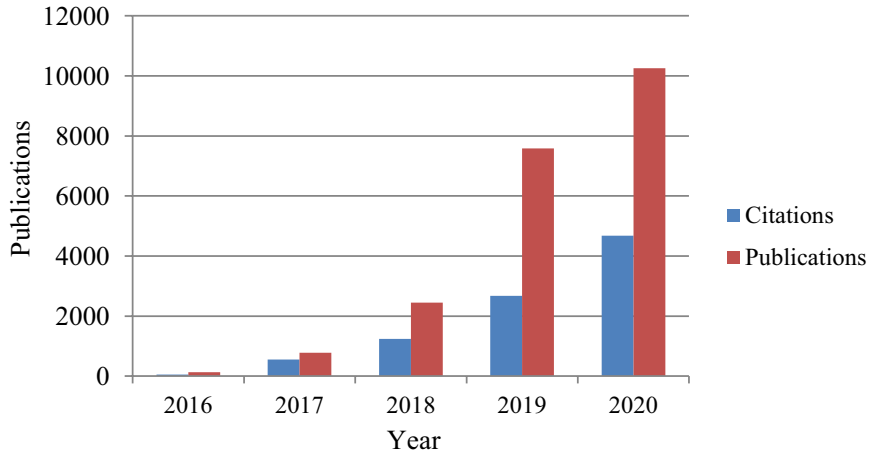


Fig. 3 WoS indexed publications and citations of blockchain

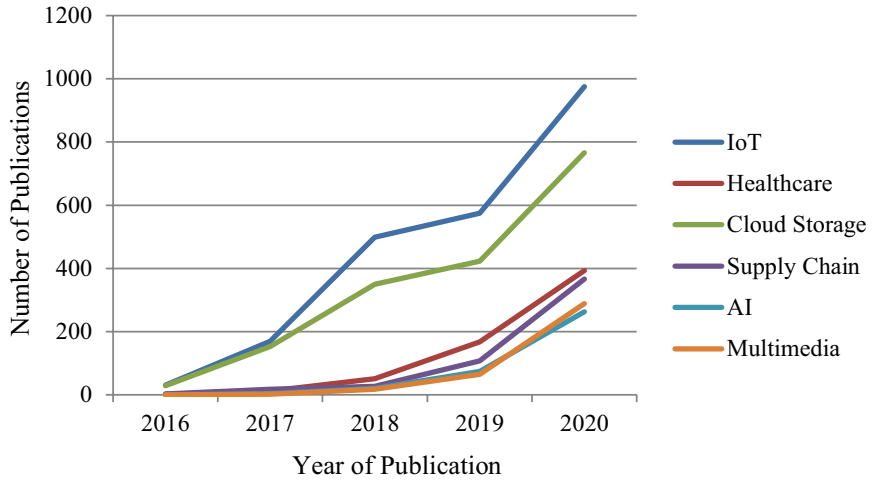


Fig. 4 Application-wise worldwide number of publications

and multimedia. As per the studies, 45% of the blockchain technology research papers are based on IoT, whereas 16% of papers use cloud data storage and sharing system followed by the other domain fields [161–165]. Figure 4 indicates the potential and opportunities of this advanced technology in various application areas. It is showing tremendous growth in IoT, cloud, and healthcare to indicate further improvements. Whereas, for other domain areas such as supply chain, AI, and multimedia, it indicates a baseline for researchers to initiate or bring new blockchain-based projects.

5.2 Proposed Work Research Analysis

The authors have analyzed the research papers related to blockchain and its applications published between 2015 and 2019—the grey research, including draft reports and generic reports extracted from descriptive analysis. The research analysis of descriptive types provides interesting facts about the blockchain, applications, and challenges. The analysis provides depth knowledge about blockchain implementation in different areas. The thematic content was prepared based on two criteria: the content published by type and content published over time. In this paper, year-wise, papers published in various international journals

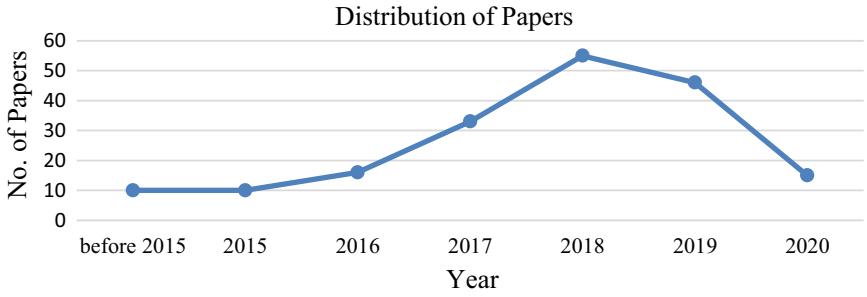


Fig. 5 Number of publications from 2015 to 2020

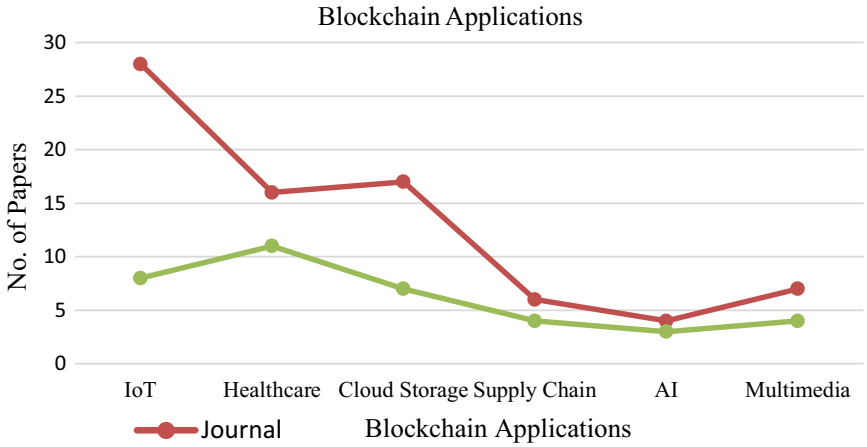


Fig. 6 Blockchain applications wise paper distribution

and conferences are illustrated in Fig. 5. The graph obtained shows that interest in blockchain techniques has risen from the year 2015 onwards. It is also observed that from the year 2015 to date, there is a rise in blockchain applications publications in vivid fields. Figure 6 summarizes the important references that have been referred to in the current review study. It lists the references that published the blockchain applications between the years 2015 and 2019.

6 Blockchain-Based Applications

Blockchain technology has greatly influenced the trade, business, and research sectors. The blockchain-based applications are analyzed based on the following categories, security, data storage, access control, consensus, and smart contracts. The following are some of the solutions offered by the blockchain applications:

Security and privacy: To researchers, the greatest fear is having the year-long study and findings hacked. Similarly, in every sector, the security of online data attracts attention. Thus, the question is how to share the data related to various transactions while preserving data safety. Blockchain answers these questions by adjoining the security of common transactions in social and academic sectors. Blockchain offers a great platform and provides solutions with full attention to security and privacy issues.

Information Storage: Data form an essential part of any organization and enterprise. Reliability issues with data storage can be successfully answered with blockchain

technology. The technology decentralizes the data and provides solutions to manage the data digitally. In the decentralization process, the data are not handled by a single agency and are distributed among numerous blocks. This system of storage is also useful in data management.

Access Control: Access control is a method that regulates access to resources at the system in the computer domain. The existing control systems suffer from many problems, such as third-party interference, inefficiency, and lack of privacy. Blockchain can address these issues, which have received considerable attention in recent years and has several potential benefits.

Consensus: A consensus mechanism is a method by which all blockchain network members reach a shared agreement on the distributed ledger's present state. Thus, consensus algorithms attain durability in the blockchain network and build trust in a distributed computing system between unknown peers. The consensus protocol essentially ensures that every new block added to the blockchain is the only version of the truth accepted by all of the blockchain members.

Smart Contract: A smart contract is a computerized transaction protocol that helps users translate contractual clauses into a coded form, minimizing external risks. It is defined as decentralized scripts that are processed and stored without any requirement of trusted parties and provide a solution for limitless applications. Thus, the applicability of blockchain has rapidly increased [50]. National and international trading requires open collaborations for business growth. Blockchain technology offers a solution to inter-organizational business processes by developing contracts where enterprises can automate the transaction processes without manual confirmation [51].

Figures 7 and 8 list various blockchain applications studied in the present review along with the corresponding references to publications in those fields. These applications have been discussed category-wise in the following section.

6.1 Blockchain and Internet of Things (IoT)

The Internet of Things is an interface of heterogeneous smart devices. The different application areas of IoT are based on a centralized architecture. The centralized system has problems like failure of a single point, trust management, and privacy problems. The blockchain-based architecture for the IoT system connects all devices in a distributed manner and provides a more secure method to share resources or data. As shown in Fig. 9, the centralized architecture of IoT is converted into distributed

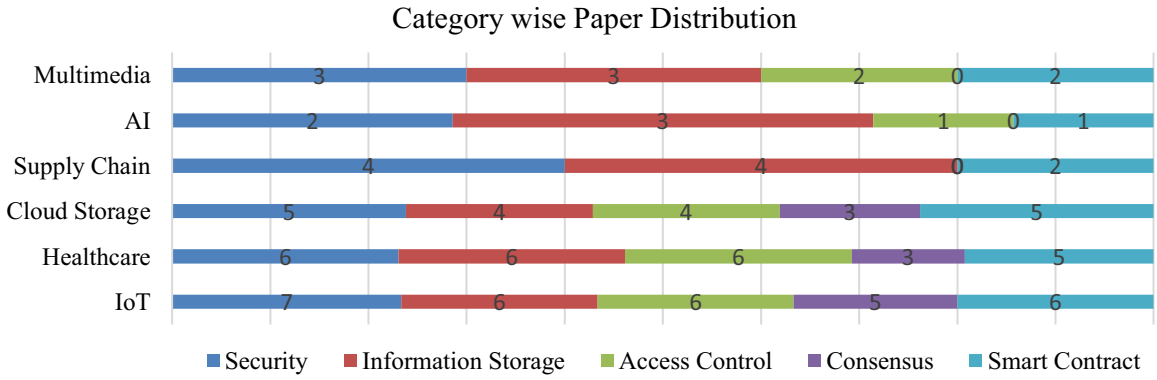


Fig. 7 Category-wise paper distribution

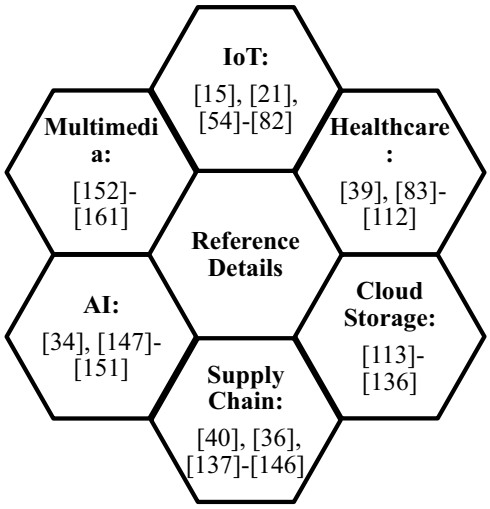


Fig. 8 References details

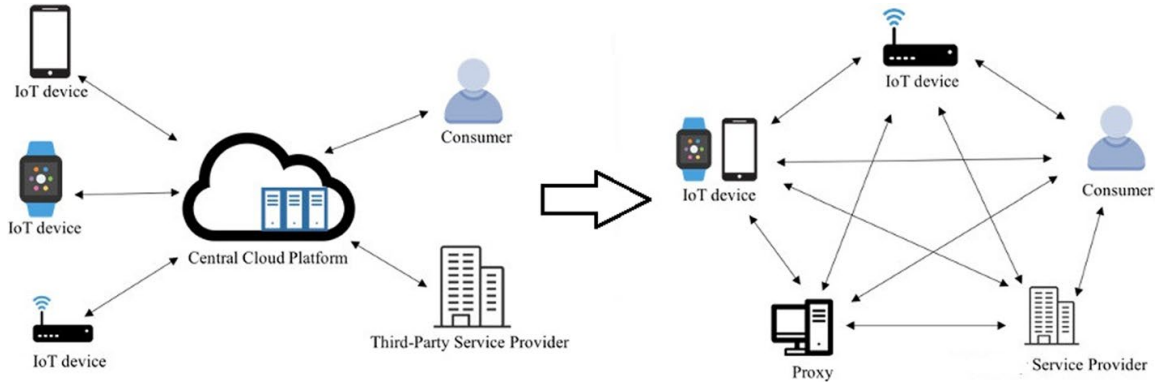


Fig. 9 Centralized and blockchain-based IoT architecture

blockchain-based architecture to resolve the security and single failure problems. The integration of IoT and blockchain technology makes the device robust and tamper-proof. The paper [52] clarified the blockchain-based trust protocol in the IoT scheme. The detailed description of IoT applications and security problems are summarized in [53]. It explained the IoT architecture and described the details of the thread model. In [54], the authors proposed a blockchain-based authentication method for IoT devices that generated zero-knowledge proof to validate the device and used the signature concept for the verification process. Similarly, [55] described the encryption and digital signature methods with IoT model-based applications. In [56], the authors surveyed the industry IoT architecture concepts and analyzed the implementation process's security requirements. The paper [57] described the primary architecture models, communication models, security, and privacy problems of the various IoT applications.

There are many research application areas of IoT by using blockchain. This paper categorized the blockchain-based IoT application into five categories, as shown in Table 4. It presents the category-wise comparative analysis of various applications in the field of blockchain-based IoT systems.

Table 4 Blockchain-based IoT applications

Factor	References	Year	Method	Contribution
Security	[58]	2019	Lightweight, scalable blockchain	Proposed a blockchain-based lightweight, scalable architecture for the IoT system. It used overlay nodes to ensure scalability and reduce overhead. It also proposed a distributed time-based consensus algorithm with blockchain to provide security
	[59]	2020	Lightweight integrated blockchain	Designed an integrated blockchain model deployed in a smart home environment. It utilized overlay networks, certificate less cryptography, and distributed management schemes to verify resource-constrained IoT devices' security and privacy
	[60]	2018	Blockchain-based security management	Introduced blockchain-based security architecture to resolve the security issues of application, network transmission, and perception layer of IoT framework. It utilized a device identification-based key algorithm to guarantee security and reliability
	[61]	2018	Blockchain mechanism for IoT security	Highlighted various IoT environments where the blockchain mechanism provides an important role to ensure IoT security
	[62]	2018	Blockchain-based authentication	Suggested a bubble of decentralized trust architecture to provide robust identification and authentication of IoT devices. It created secure virtual zones (bubbles) to identify the devices and maintain trust between them
Information storage	[63]	2018	IoT devices ownership management	Proposed a blockchain-based architecture for managing the ownership details of IoT devices without a trusted party. It defined the ownership rules with the help of smart contracts
	[64]	2019	Trust management system	A blockchain-based trust management system is deployed to establish trust between IoT devices. It is divided into two layers: the IoT layer and distributed fog layer. The IoT layer provides communication between devices, whereas the fog layer maintains trust values between the devices
	[65]	2020	Blockchain-based data allocation	Suggested a new context-aware data allocation method determining the ranking value of each IoT data request to determine its allocation on-chain

Table 4 (continued)

Factor	References	Year	Method	Contribution
	[66]	2020	IoT device location storage system	Designed a blockchain-based location chain storage system for IoT devices. To provide location privacy service, blockchain technology maintains location information
	[67]	2018	Blockchain-based secure storage	Proposed a blockchain-based threshold system for IoT service called BeeKeeper. The system server used homomorphic computations to process the user's data
Access control	[68]	2015	Attribute-based access control	Proposed a novel access control mechanism for the IoT system. Each device is described using predefined system attributes and assigned by attribute authority as per the identity or ability. Blockchain is utilized to store attributes distribution in the form of a ledger
	[69]	2018	Distributed access control	Designed blockchain-based distributed access control architecture to assign roles and permissions in IoT. It addresses the issues with managing constrained IoT devices by not including them in the blockchain network, thus making integration easier
	[70]	2019	Attributed-based access control	A lightweight and robust access control mechanism is designed to address privacy and security issues for the IoT system. The integration of FairAccess and Privacy-preserving access control provided a more secure environment for the IoT environment
	[71]	2019	Access right delegation	Developed a secure access control model to address delegation issues of the IoT system using blockchain technology. It supports two blockchain networks for transferring the access rights in the IoT framework and also ensures security
	[72]	2017	Hierarchical access control	Proposed a blockchain-based privacy-oriented key management system to achieve hierarchical access control. It combines the blockchain concept with cloud and fog computing to achieve decentralization, scalability, privacy, and access control in the IoT system

Table 4 (continued)

Factor	References	Year	Method	Contribution
Consensus	[73]	2020	Proof of X consensus protocol	Designed a Proof-of-X consensus protocol for the IoT system. It used attributes of the participants of the network
	[74]	2019	Storage compression consensus (SCC)	Proposed an SCC algorithm to compress blockchain in each IoT device to ensure storage capacity
	[75]	2014	Distributed consensus algorithm	Proposed a distributed decision-making consensus algorithm for service detection, processing, classification, and data composition for the IoT system
	[76]	2019	Two-layer consensus algorithm	Designed a blockchain-based two-layer consensus algorithm to optimize the IoT device resource requirements
Smart contract	[77]	2018	IoT-edge framework using blockchain and smart contract	Proposed a blockchain and smart contract-based edge-IoT framework to address security and scalability challenges. It uses a credit-based resource management method to control the resource requirement of IoT devices
	[78]	2016	Blockchain and smart contract for IoT	Provided a detailed description of the integration of blockchain-based smart contracts and IoT system. It identified the potential use cases of the IoT framework with blockchain
	[79]	2018	Smart contract based monetization of IoT data	Proposed a smart contract-based monetization framework in which payment and token access issues automatically to the IoT owner without any third party
	[80]	2019	Auditing scheme with smart contract for IoT	Proposed a decentralized auditing scheme that supports dynamic auditing, public auditing, and batch auditing. Smart contracts are used to store deposits which can pay for auditing and punishment
	[81]	2019	Contract learning approach for resource sharing and task offloading	Introduced two-layer architecture for resource sharing and task offloading in the IoT system. It combined the contract concept with computational intelligence. The first layer uses an incentive method, and the second layer uses a task offloading algorithm to minimize delay in the IoT system

6.2 Blockchain-based Healthcare System

The healthcare system is an information-intensive medical environment where large amounts of data are routinely generated, obtained, and disseminated. Due to the sensitive nature of data and restricting factors such as protection and privacy, storing and distributing this vast volume of data is crucial and significantly challenging [82]. In healthcare, secure information protection has been innovated throughout the last decade through a vast number of platforms, software, and communication technologies, all aimed at improving the security of healthcare records. The first health records of paper were translated into Electronic Health Records (EHRs) [83]. EHRs must be regularly distributed and exchanged by various healthcare centers, doctors, pharmacy manufacturers, and administration to provide a realistic way for a patient's health background to prompt treatment. In the case of a conventional client–server data management healthcare system, each hospital/healthcare center retains its database of sick person medical records; the delivery of EHRs becomes a slow and costly task. Web-based health information monitoring methods [84–86] have been presented previously to solve the accessibility, data usage, single failing point, safety, and security issues in the client–server architecture. There is still a single failure in cloud-based systems, information protection, and the patient's security threat occurring in the platform. Blockchain is a recent development in computer technology. Blockchain technology offers transparent, shared, and digitized ledger. All the entities participating offer shared resources without a single failing point, thereby eliminating the possibility of central point bottlenecks, as shown in Fig. 10.

Many study results have employed the technology to fix the weaknesses in existing EHR. Many research publications [87–91] utilized blockchain to solve health documents' privacy and security issues by maintaining cloud information hash within the blockchain. Several studies either introduce new data encoding/decoding techniques [92–94], or a more modern digital signature method [95], or a protected scheme of information transmission [96, 97] or keys generator method [98] used by the blockchain for health information. The various blockchain-based healthcare applications are summarized in Table 5.

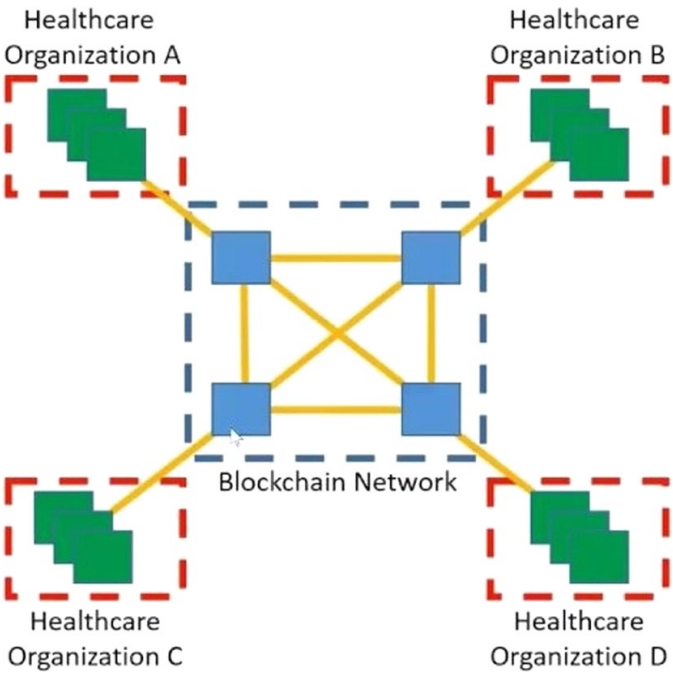


Fig. 10 Blockchain-based healthcare architecture

6.3 Blockchain-based Cloud Storage Service

Cloud computing outsourcing service provides potential benefits to the cloud users. This service addresses the limitations of computationally weak devices by outsourcing their data in the cloud with the help of a pay-per-use approach [112]. The user can rent and pay the storage services or utility computation depending on cloud infrastructure requirements. For the data, confidentiality is vital because of the privacy requirement, and the service provider is not trusted. The researchers' major challenge explored and highlighted is the processing and storing of data into the cloud. The cloud environment and the technology of the blockchain are adapted for this usability. Therefore, for the improvement in existing applications' performance, these two approaches are combined [113]. Hence, blockchain is one of the decentralized or secure networking environments containing several computers called nodes. Furthermore, this technology to improve increases accuracy.

A lot of new technologies and frameworks have been introduced with the current keen interest in blockchain technology. Numerous review articles were published to demonstrate the advantages of blockchain for existing applications. For example, reference [114] provides a detailed overview of privacy and security concerns in cloud computing, covering potential threats and detection methods based on blockchain. Reference [115] presents key concepts of various sharding mechanisms focused on blockchain technology. In [116], the authors address safety, security, and transaction processing issues regarding the use of blockchain for cloud exchange. Moreover, [117] is dedicated to blockchain for edge computing systems and their potential uses. This paper provides a detailed overview of the usage of blockchain technology in cloud computing. From a research point of view, lots of work has been done in the cloud computing field using blockchain; these are summarized in Table 6.

6.4 Blockchain-based Supply Chain

Supply chains are currently becoming extremely complex in structure, challenging in terms of tasks and diverse stakeholders. Several businesses do not have an integrated view of the entire supply chain. Several global companies have developed their own identities and platforms to retain global operational exposure and have the ability to direct their suppliers. They have to focus on centralized administrative or intermediary bodies. In terms of confidentiality, traceability, authentication, and verification method, this low transparency causes many problems and difficulties in the supply chain process. Blockchain is a revolutionary computer technology that can support many possible operations and supply chain-related applications. It is important to notice that blockchain is well suited to the complexities of supply chains. Blockchain technologies can also contribute to the domain of operations and the supply chain [136]. For example, Fig. 11 represents the blockchain-based supply chain management architecture for tracing product details. The architecture utilized a blockchain ledger to store the product details in the form of blocks with the help of smart contracts, and customers are allowed to trace the product by using product RFID code through the mobile application. Thus it is vital to implement blockchain technology, with its immutability, openness, and trustworthiness functionality [137], to have more transparency and security in the supply chain domain. Table 7 summarizes the analysis of supply chain applications that focused on blockchain technology.

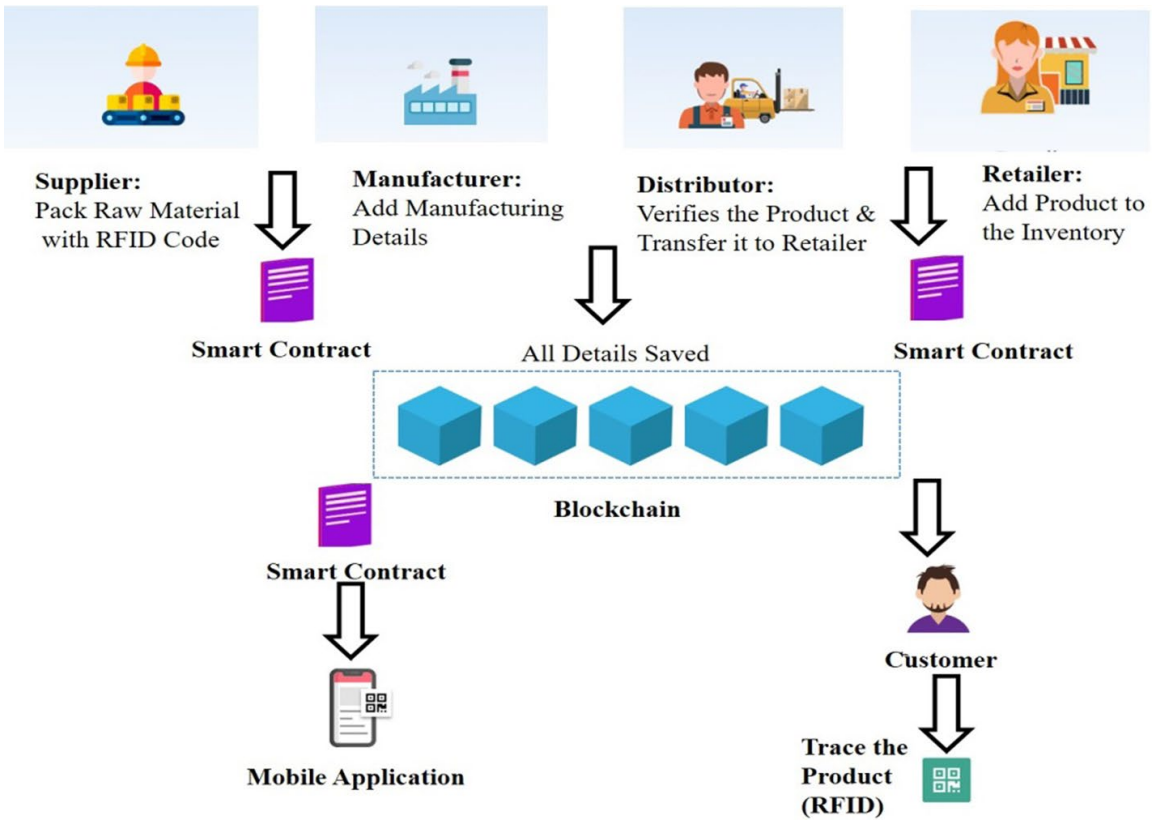


Fig. 11 Blockchain-based product traceability in supply chain system

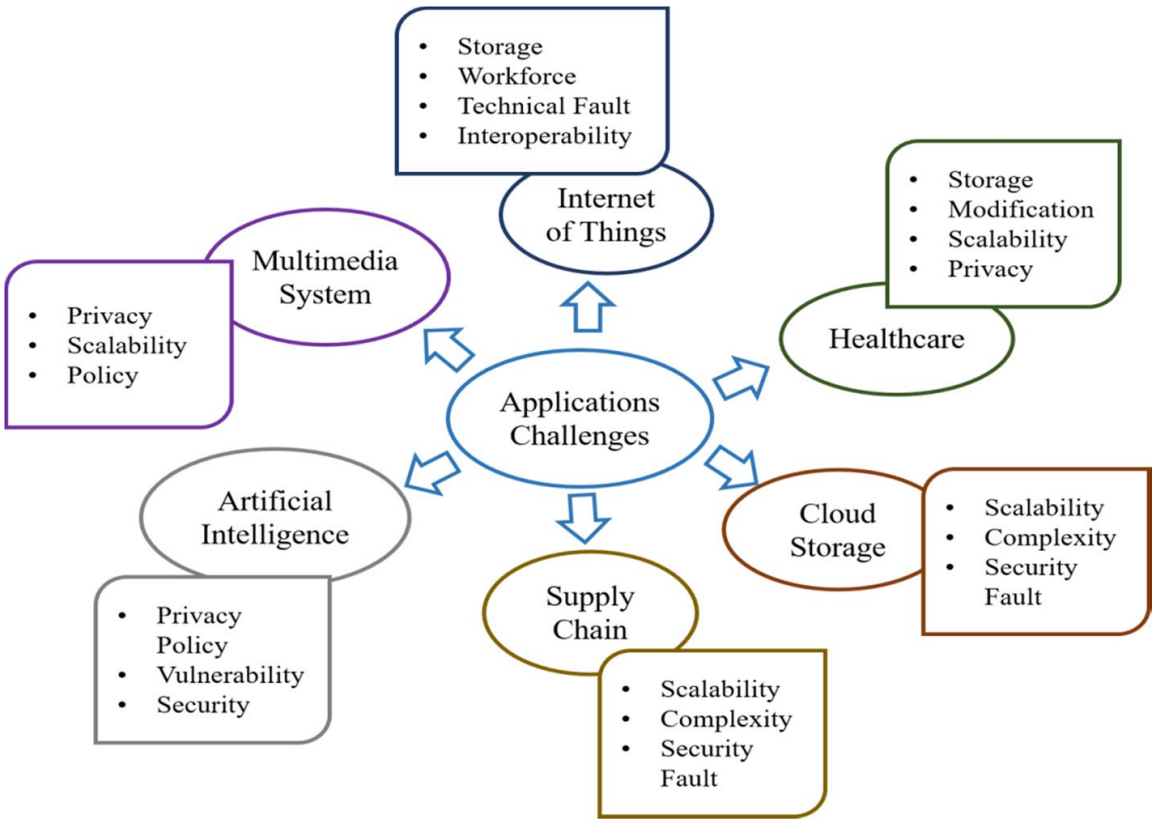


Fig. 12 Challenges of blockchain applications

Table 5 Blockchain-based healthcare applications

Factor	References	Year	Method	Contribution
Security	[99]	2018	Blockchain-based key management scheme	Proposed a solution that integrates the health blockchain and body sensor networks. It designed a lightweight backup and efficient recovery method for the keys of health blockchain
	[100]	2018	Privacy-preserving healthcare data management method	Presented a patient-centric blockchain-based healthcare data management. Cryptographic functions are used to encode the patient's data and ensure privacy
Information storage	[101]	2019	Blockchain-based smart healthcare system	Proposed a secured and smart healthcare system using blockchain to provide security and privacy to preserve the healthcare system
	[102]	2019	Blockchain-based electronic health record	Proposed a Hyperledger based electronic healthcare record sharing system with the use of chain code. It proposed an access control policy algorithm to improve the data accessibility between healthcare providers
	[103]	2020	Off-chain storage of patient diagnostic reports	Proposed a distributed off-chain storage system for patient record management with the help of IPFS storage. It preserves patient privacy and provides accessible healthcare data
	[104]	2019	Mobile cloud-based E-health system using blockchain	Presented the implementation of Ethereum blockchain in a real scenario with a mobile app and Amazon cloud computing. It preserves private health information with smart contracts, access control, security, and privacy features
Access control	[105]	2018	Block-based access control in e-health	Presented a block-based access control method to exchange blockchain-based electronic health records
	[106]	2019	Secure sharing and accessing of medical data	Proposed a MedBloc architecture to allow patients and healthcare providers to access and share data securely. It used encryption and smart contract to regulate access to the records
	[107]	2016	Medical data access and permission management using blockchain	Proposed a MedRec architecture to manage authentication, authorization, and access control method. It used a modular approach to design permission management systems in the healthcare domain

Table 5 (continued)

Factor	References	Year	Method	Contribution
Consensus	[108]	2018	PBFT consensus mechanism in healthcare blockchain domain	Presented a continuous-time Markov Chain model to evaluate the performance of the healthcare blockchain network. It analyzed the factors like mean network delay, primary node delay, and replica node delay
	[109]	2019	IoT healthcare consensus	Proposed a blockchain-based IoT healthcare architecture called IoBhealth that collects the data from IoT sensors. It provided secure access and management of patient data and summarized the various consensus algorithm used in the healthcare domain
Smart contract	[110]	2018	Secure remote healthcare system for hospitals using smart contract	Proposed a remote healthcare system that provides privacy and security of health information with the help of a smart contract. It implemented a processing mechanism to filter the data from the sensors
	[111]	2018	Blockchain and smart contract in decentralized health structure	Presented a decentralized model for the healthcare system. It used a scheme to distribute registered data for the creation of electronic medical records. It also proposed an algorithm for the use of a smart contract in the model

Table 6 Blockchain-based cloud storage applications

Factor	References	Year	Method	Contribution
Security	[118]	2017	Blockchain-based integrity framework	Proposed an integrity checking system using blockchain technology. The designed system allows consumers and data owners to perform trustworthy integrity verification without depending on a trusted third party
	[119]	2019	Privacy and auditing scheme	Suggested an architecture based on blockchain and rank-based Merkle AVL tree structure to provide a privacy-preserving cloud storage environment
	[120]	2019	Encryption based searching	Developed keyword searching technique using public-key encryption method called SEPSE that provides security against keyword guessing attacks. The system allows users to encrypt keywords using the threshold from specific key servers
	[121]	2019	Medical data storage using cloud and blockchain	Designed a blockchain-based architecture to store healthcare data using cloud storage techniques. The system provides security and privacy to the stored healthcare data using blockchain technology
	[122]	2019	Secure healthcare framework using cloud and blockchain	Suggested secure eHealth architecture to provide integrity, correctness, and confidentiality features to the outsourced electronic health records. The healthcare records generated by the doctor during the period were recorded as transactions in the blockchain structure

Table 6 (continued)

Factor	References	Year	Method	Contribution
Information storage	[123]	2019	Decentralized storage	Proposed a provenance information storage system using blockchain technology and IPFS (InterPlanetary File System) with validity checking options
	[124]	2019	Blockchain-based on-chain and off-chain storage	Designed an auction-driven reward system with on and off-chain storage facility using blockchain technology. The proposed system was designed using hyperledger fabric to provide security to the on-chain data, whereas the auction system assessed the off-chain data performance
	[125]	2017	Distributed storage	Suggested a reliable distributed BlockDS system to provide searching and storage services using blockchain technology
	[126]	2018	Distributed cloud storage	Introduced a distributed blockchain-based cloud storage architecture that allows users to divide their data into multiple encrypted chunks and store them randomly to a P2P network
Access control	[127]	2018	Role-based access control	Implemented a role-based access control technique using smart contracts to provide a trans-organizations information sharing system
	[128]	2019	Access control based decentralized method	Designed a blockchain-based distributed system to secure cloud data using the access control method. The proposed work utilized the conventional attribute-based encryption technique using Ethereum smart control concept. It provides a distributed environment that is independent of central authority and a trusted third party
	[129]	2018	Fine-grained access control	The proposed system combines the attribute-based encryption technique with the Ethereum blockchain network. The system using the decentralized storage system IPFS to store the encrypted user data in an off-chain manner. The data owner shared the secret key with the user and specified the access policy rules using the fine-grained access control approach

Table 6 (continued)

Factor	References	Year	Method	Contribution
Consensus	[130]	2019	Proof of stake consensus method	Designed blockchain-based provenance architecture for the cloud computing environment called as BlockCloud. The suggested system implemented a PoS consensus mechanism to reduce the computational overhead compared to the PoW method
	[131]	2019	Synchronous Byzantine Fault Tolerance consensus algorithm	Suggested a Synchronous Byzantine Fault Tolerance (SBFT) algorithm to provide higher efficiency and maintain data consistency than Byzantine Fault Tolerance (BFT) and Practical Byzantine Fault Tolerance algorithm
	[132]	2019	Proof of Work algorithm using maximum factorization statistics	Discussed a PoW consensus approach for cloud and fog computing by using a statistical method. It used the expectation–maximization and matrix factorization algorithm to reduce energy and memory consumption
Smart contract	[133]	2018	Ethereum smart contract	Proposed Ethereum-based blockchain architecture to control the access and provide storage facility with the help of smart contracts deployed functions
	[134]	2018	Decentralized auditing method	Designed a decentralized public auditing system for the cloud storage environment using blockchain technology. The designed system eliminated the need for third-party auditors required for the auditing service with the help of smart contract functionality
	[135]	2018	Deduplication method using smart contract protocols	Proposed a smart contract-based deduplication approach to store files on multiple servers using blockchain technology without a central authority

Table 7 Blockchain-based supply chain applications

References	Year	Method	Contribution
[138]	2018	Double chain based agriculture supply chain system	Proposed a double chain-based blockchain architecture for the agriculture supply chain system. It supports adaptive rent-seeking and matching between supply and demand of resources
[139]	2019	Blockchain in operations and supply chain	Demonstrated the architecture of the blockchain-based logistics monitoring system using the Ethereum platform
[140]	2019	Blockchain architecture for containerized food chain	Demonstrated the architecture using Hyperledger fabric to provide secure information sharing, enhance process access, and prevent risks in containerized food supply chain systems
[141]	2020	Blockchain-based Indian agricultural supply chain	Investigated the barriers to the adoption of blockchain in the Indian agriculture supply chain. Then, the identified barriers are modeled in the proposed architecture and evaluate the performance
[142]	2017	Blockchain-based supply chain management for information sharing	Suggested blockchain-based information sharing architecture that brings benefits to supply chain management. It also used a homomorphic encryption solution without a third party
[143]	2019	Blockchain-based construction supply chain system for key protection	Designed a blockchain-based private-key distribution framework to preserve key security and recovery. It provides payment security in the construction supply chain by using a key management scheme
[144]	2020	Blockchain-based safety management system in grain supply chain	Designed a multi-storage architecture for the grain supply chain system. It is characterized by data security, reliability, information interconnection, sharing, and whole process tracing features
[145]	2020	Secured information sharing in supply chain management	Proposed a cryptographic key distribution mechanism using a smart contract. It designed a blockchain-based information sharing framework in the pharmaceutical supply chain using a smart contract and consensus mechanism

6.5 Blockchain and Artificial Intelligence (AI)

AI and blockchain prove to be a powerful integration that changes almost every sector in which they are applied. Blockchain and AI collaborate to improve all fields, from food procurement distribution to an automated system sharing health care information. Integrating AI and Blockchain includes many fields like secure AI, decision making, and prediction models. AI can efficiently mine and develop new scenarios through a huge dataset and find patterns based on data behavior. Blockchain effectively supports the elimination of bugs and malicious data sets. New AI-generated classifiers and patterns can be verified using a distributed blockchain network. It may be utilized in any user-based company, such as retail purchases. Information captured from users can be used to create marketing automation through AI and the blockchain infrastructure. Perhaps AI and blockchain's integration produces the most robust technology that enabled the decision-making mechanism in the world, which is practically tamper-proof, offering vital perspectives and decisions. It has several advantages, such as improved business models, transparent systems, smart models, etc. Table 8 lists the various application in the domain of artificial intelligence and blockchain technology.

6.6 Blockchain in Multimedia

Multimedia protection, material rights, music files, pictures, and video/movies become a problem for distributors, creators, and artists [152]. Media processing and dissemination are typically supported by CDs, DVDs, pen drives, or email. One of the most known examples of a security violation being the illegal electronic distribution of multimedia before the scheduled release date, which resulted in a major loss for the show's creators and distributors [155]. To avoid these incidences and to identify and convict the individual culprits, it is critical that the media file owner should have full access, permission, and a safe channel to exchange his data [151]. If data were leaked, the owner would be able to claim his rights over the content. Digital technology, such as online cloud storage such as Dropbox or peer-to-peer networking, is the best possible platform for quick and secure data exchange [152].

Multimedia violations over the last few years have raised issues about privacy in this area. Each time a problem is raised, innovations and work are pursued to mitigate it. Therefore, multimedia breaches allow developing alternative solutions that allow customers to save and exchange the data securely. The system should support the owner's authorization and protection controls to track a third party's operation [153]. There are presently many studies underway in this field, and various views and approaches have originated. Many research works have concluded with the owner deliberately disclosing specific information [154]. Even though cloud infrastructure is extremely powerful and more secure than computers, there are many external and internal threats to data integrity. Occasionally there are occurrences of cloud data leakage. It is important to notice that blockchain is well adapted to tackle multimedia issues. Thus, it is significant to implement blockchain technology with its immutability, openness, and trustworthiness functionality to have further protection in this field. Table 9 presents the multimedia applications using blockchain.

Table 8 Artificial intelligence applications using blockchain

References	Year	Method	Contribution
[146]	2019	Integrated blockchain and artificial intelligence solution for autonomous cars	Proposed an integrated solution in which each car is connected to the public ledger and share the experience to learn when to stop. Thus, it eliminates the task of training each car separately
[147]	2019	Blockchain-based intelligent IoT with AI	Implemented an effective big data analysis using blockchain-enabled intelligent IoT architecture. It supports both qualitative and quantitative analysis
[148]	2020	Blockchain and AI for decentralize healthcare system	Proposed a blockchain-based decentralized healthcare system, and AI provided the roadmap for data analysis to discover drug and preventive healthcare
[149]	2019	Decentralized AI model network	Introduced a decentralized AI model to improve the accuracy of machine learning models. It maintained an append-only ledger to store the log of the details of who trains the model and improvement in accuracy
[150]	2019	Securing data using blockchain and AI	Proposed a model to enable secure data sharing, computing, and large-scale sharing with three key components, i.e., blockchain-based data sharing, AI-based computing, and trusted exchange mechanism for security

Table 9 Multimedia applications using blockchain

References	Year	Method	Contribution
[156]	2019	Detection of tempered images using blockchain	Proposed a blockchain-based architecture to register information about ownership and rights of the author and descriptive details of the image to detect a violation
[157]	2019	Blockchain-based video integrity verification	Suggested a blockchain-based approach to combine hash-based message authentication and elliptic curve cryptography to verify the video's integrity
[158]	2018	Multimedia privacy and provenance protection using blockchain	Proposed a blockchain-based multimedia protection framework that enables the users to take full control of data without trusted authority
[159]	2019	Music streaming using blockchain technology	Introduced a blockchain model using Ethereum and IPFS storage to remove delays and transparency problems in the exiting centralized methods
[160]	2019	Global music industry using blockchain	Proposed a novel actor enabled the model to capture the unique properties of digital entrepreneurship

7 Open Issues and Challenges

In this section, we discuss and identify the current challenges associated with combining blockchain technology with different domains as shown in Fig. 12. Despite the exciting advantages and the incredible foreseen potential of blockchain, there are major obstacles in implementing and delivering existing and proposed networks that would require further investigation. Some of the potential challenges faced by blockchain applications are explained in the following subsections.

7.1 Internet of Things

Despite the exciting advantages and the incredible foreseen potential of Blockchain and IoT, there are significant obstacles that would require further investigation:

Restriction with Storage Facility: In the IoT environment, the storage space needed for sensors and devices is less than that of the blockchain ledger. In IoT, strong, unified server storage is enabled, whereas each ledger needs to be stored at each node in the blockchain. Compared to the current capacity in IoT devices, this increases the data size over time.

Workforce Scarcity: Trained workforce for the blockchain platform is very low, and the figure is incredibly small when blockchain is integrated into the IoT concept. This implies that there is a very less-skilled worker who knows about the blockchain-integrated IoT concept.

Variation of Computational Capabilities: IoT networks are distributed and linked over a large network; this structure becomes more complicated when the blockchain is combined with IoT. It is essential that all things connected to the blockchain-based IoT system run the encryption. In such cases, not all of the algorithms used to run the encryption may have similar computing capabilities, thus results in variation problems.

Complex Technical Challenges: The challenges related to scalability, safety, cryptographic implementation, and reliability specifications of novel blockchain-based IoT applications still need to be addressed. Additionally, blockchain systems face design drawbacks in communication efficiency, authentication protocols, or smart contract implementation, which require further investigation.

Interoperability and Standardization: To achieve complete interoperability and the implementation of blockchain in IoT would require the cooperation of all stakeholders and the integration of existing systems. It will require collaborative applications and international standards to protect access control, authentication, and authorization.

7.2 Healthcare

While there are various advantages offered by blockchain technology, some inherent disadvantages pose some challenges. Storage, scalability, alteration, safety, and policies are the primary healthcare issues facing this technology.

Storage: Healthcare and medical records generate massive volumes of data from the patient and wearable IoT devices. While the blockchain architecture enables limited data storage, various operations on blockchain data are also expensive if the data size is larger. Thus, this factor should be considered while designing a blockchain application.

Modification: On the one hand, blockchain data immutability features to protect the system, but on the other hand, it does not provide the data updating and deleting operation. Either there is a need to build a new block out of all nodes by consensus or create a new

chain. These two approaches are expensive and not beneficial. Therefore, the development of blockchain applications must be the lowest need to modify the data.

Scalability: The scalability problem is less severe because of the decentralized architecture. Nevertheless, there are millions of customers of various networks, including healthcare centers, business research agencies, insurance providers, patients, IoT businesses, etc. It is highly unlikely that they will all be able to maintain the same blockchain distributed structure. Blockchain also requires high computing capacity that needs high network equipment electricity consumption. To make blockchain popular, the healthcare scalability issue needs to be considered seriously.

Privacy and Regulations: Blockchain significantly increases the confidentiality of the data. Cryptographic, open, autonomous, and immutable blockchain features can ensure complete protection of the information. Healthcare data is all about storing confidential patient information. In contrast, it can be dangerous to have a copy of such data in each node which requires further consideration.

7.3 Cloud Storage

Although blockchain provides disrupting services to a cloud environment, its development still suffers from various vital complexity, scalability, and security issues. Here are a few reasons why blockchain is not quite ready for large-scale usage cloud application needs.

Scalability: The existing blockchain networks have extreme bottlenecks in scalability, with minimal availability, power, and cost. Due to block size constraints, many blockchains have lengthy processing periods for transactions to be appended onto the chain. Hence the block creation time is rapidly increasing, limiting the overall throughput of the system. Additionally, if all transactions are stored in a chain, this increases the blockchain's capacity. Considering complicated cloud computing environments, consumer data is massive, leading to exponential growth in blockchain size, making it impossible to handle large data volumes. Because of these constraints, many application developers do not consider blockchain a viable option for large-scale industries.

Complexity: In the blockchain-based cloud environment, cloud users execute the consensus algorithm to process the complicated mathematical puzzles that need powerful computing hardware to implement transaction validation. However, due to cloud infrastructure limitations, this is difficult to fulfill these specifications. Also, the blockchain process's complexity can cost expensive energy and human capital resources for devices with relatively small processing capabilities.

Security Fault: The main blockchain disadvantage is an unavoidable security flaw. When more than half of the computers operate as blockchain nodes to monitor processing resources, attackers can change consensus structures and keep new transactions from having permission for malign access. It is also called an attack of 51 percent. Blockchain can be at risk of a data breach and system damage without having comprehensive management of transactions.

7.4 Supply Chain

The following lists some of the potential challenges of a blockchain-based supply chain system:

Privacy: Concerning technological problems and specifications, a primary concern is data protection and consumer security. This, along with the drawbacks of IT infrastructure,

particularly for medium-sized businesses, are the main technical factors that cause resistance among market participants and make them hesitant to apply blockchain technology to their use cases. The researchers explore the features and criteria of the possible blockchain systems specifically for supply chain integration, mainly between companies. They point to the inefficiency of the current status of businesses and analyze the expected requirements and functionalities of supply chains offered by the blockchain network.

Scalability: The major concern is the scalability problem of blockchain technology, which reduces the transactions and data that can be handled and controlled in a limited time with the available technology. Because most businesses are part of large and complicated structures, the present state of blockchain is creating concern about the scale and scope of use cases that can be applied at this time. The need to implement the smart contracts in the system and the constant updates of the digital profiles were identified as critical challenges that impede the usage of blockchain in the supply chain system.

Policy and Regulations: Because of the technology's distributed nature, its success depends heavily on conformance and basic agreement among the various supply chain parties. It also suggests a heavy reliance on blockchain operations and a need for trust between different parties and between businesses and technologies themselves. From a technical point of view, these concerns are appropriate requirements rather than challenges, and it is clear that all of these constraints affect both the industry and the financial sector.

7.5 Artificial Intelligence

This section addresses and describes the emerging problems of integrating AI and Blockchain technology. The following lists some of the conceivable challenges of merging and incorporating both technologies:

Privacy Policy: Public blockchain ledgers facilitate safe and accurate data collection, but the data obtained is publicly available and open to all readers. That can be evasion and a privacy concern. Data privacy can be secured by promoting encryption and providing managed access to private blockchain ledgers. Nonetheless, these private blockchain systems may limit access and transparency to the vast amount of data AI might need to manage and perform reliable and correct decision-making and analytics.

Blockchain Security: Blockchain's distributed ability may suffer from exploitation and misuse. Although blockchain offers effective predictive processing schemes, the blockchain networks are vulnerable to cyber-attacks, such as 51 percent attacks. Depending on the miner's hashing power, the consensus algorithm can be corrupted and can work around a few mining entities that control consensus as a centralized platform. For shared blockchains like Ethereum and Bitcoin, this security issue is more apparent. Private blockchain systems suffer less from this problem, as protocols of consensus between parties are predefined.

Vulnerability of Smart Contracts and Deterministic Implementation: Ensuring that the deployment of a smart contract is bug-free, security breaches, and secure from attacks is crucial. Protecting the coding and details on the network is critical, as they can be susceptible to attack. This can present a crucial problem for distributed AI, in which decision-making algorithms and machine learning are carried out by the mining nodes using smart contracts, in which the execution result is typically not deterministic but somewhat arbitrary and unreliable. This results in an innovative approach for working with provisional calculation and designing consensus protocols for mining nodes to agree on outcomes with a limited degree of reliability, precision or consistency, and extremely fluctuating data content.

Governance: Implementing, designing, and controlling a blockchain structure among various stakeholders and members is complicated. For a private cooperative blockchain, significant problems emerge relating to the blockchain type to be implemented, who administers and troubleshoots the blockchain, which codes the smart contracts, conflict resolution, protocols for off-chain operations, implementation of side channels, and control of side channels. This requires research aimed at developing governance models.

7.6 Multimedia

Blockchain has generated some hype, and it's still under discussion on how the technology can be implemented for mass adoption in the entertainment industry. There is a need to set regulations and standards to maximize the value blockchain can create. The process of introducing new technology is costly and time-consuming.

Blockchain Architecture: A robust platform or network that can satisfy all the criteria for using blockchain in multimedia systems must be built. Many research methods address commitment, privacy, and security depend on policy and controls across domains. The governments, for example, should establish a blockchain structure to provide public interest cases.

Storage: Large quantities of records to be kept at the blockchain nodes due to the number of "transactions" (images, streams, videos, etc.) across all channels of the multimedia system. Conversely, the blockchain model enables minimal on-chain storage of data. Blockchain's distributed and hashed architecture is too expensive to store the data. Similarly, if the size of multimedia is larger, access, maintenance, and functions on blockchain data can also be costly. Hence blockchain applications must be designed with these factors in mind.

Regulations: Absence of legislation, vague advantages, and communication gaps between technical experts and policy-makers are major problems that do not promote technology implementation. Common standards for blockchains still need to be agreed upon. Many potential stakeholders in the multimedia system must approve common blockchain platforms and interoperable blockchain standards.

8 Recent Advances

This section identifies the solutions and recent advances to the identified open challenges and also explores the chances of further improvements.

Storage: To resolve the storage requirement, researchers utilize off-chain storage solutions like bigchainDB, Swarm, IPFS, Filecoin, Sia, Storj, etc. [166]. BigchainDB expands the blockchain features as a big data distributed system increases the throughput with immutability and decentralized blockchain system [167]. Other necessary decentralized storage solutions manage a large amount of data since it is difficult to store them inside the blockchain structure. The decentralized storage system provides a peer-to-peer distributed file-sharing system that increases blockchain-based systems' efficiency and storage capability [166]. The decentralized storage network utilizes the blockchain structure to store the meta details and allow the users to access the information at any time.

Workforce: The demand for professionally qualified blockchain developers is increasing dramatically, whereas the existing blockchain-based domains suffer from a shortage of trained or skilled people. According to [168], the demand for blockchain developers has

increased by 2000% from 2017 to 2020. Therefore it is a major concern in the current scenario for the organizations. The blockchain is still evolving. Thus it requires time for the educational institutions to introduce relevant courses to alleviate the market demand.

Scalability: The recent Blockchain platforms have severe bottlenecks of scalability. Many blockchains have long waiting times to add new transactions, making it difficult to process high volumes of data due to block size restrictions. Also, transaction validation is the key component of blockchain technology that uses consensus protocol to validate the transaction of each block. The computational power required to generate a block depends on the number of transactions in a block, which affects the transactions confirmation times. Thus, the consensus algorithms have a direct impact on blockchain technology scalability. Recent advances such as GHOST [171] utilize the chain selection rule to improve the scalability of the bitcoin network. Bitcoin-NG [169] proposes a new consensus protocol to improve the latency of the transaction. Similarly, the off-chain storage solutions [172] perform the transactions offline and increase the blockchain network bandwidth. Another proposal Litecoin [170], confirms the transactions faster and improves the storage efficiency due to the reduction of block generation time.

Privacy: To increase privacy, the Hawk [173] compiler is designed that stores encrypted transactions. The Hawk translates the generic code into cryptographic primitives to enable transaction anonymity. Similarly, the Enigma [174] solution splits the data into multiple chunks and distributes them into peer-to-peer networks so that no node accesses the data. It uses the distributed hash table to store data references details in a decentralized offline manner. Also, the privacy problem in private blockchain networks resolves by using the private permissioned blockchain structures such as Quorum [175], Hyperledger Fabric [176], a blockchain that utilizes access control and membership functionality.

Security/Technical Fault: The most common 51% attack or majority attack [177] is a significant security concern in blockchain technology. If more than half of the Blockchain nodes act maliciously, they may risk a data breach and prevent new transaction confirmations. The researchers propose many consensus algorithms for blockchain technology to prevent majority attacks. In reference [178], the authors discuss the solo mining incentive or peer-to-peer mining concept to alleviate this problem. Still, the proposed consensus mechanism is susceptible to 51% attacks, specifically those that utilize the centralized consensus among a limited number of users.

Policy and Regulations: The absence of central authority makes blockchain technology suspected of promoting and facilitating illegal conduct. Due to the lack of governance, many countries are developing new regulations to legalize virtual currency and distributed technology [179–181]. The legal implication of virtual currency is a significant concern as it directly or indirectly affects blockchain applications. Thus, the key to increasing confidence in blockchain technology could be requiring governments' or companies' policies and regulations in their development. Current initiatives such as Alastria [182] aim to consider multiple authorities to develop a blockchain-based national regulation system for public notaries, universities, and private organizations. These initiatives allow a legal wallet or account for each person that represents digital proof of possession. This information is used for many services like storing healthcare, cloud, or IoT data legally or trustworthy.

Interoperability: Blockchain applications want to implement interoperability in their platform to create an ecosystem to enable communication between different blockchains. The interoperable blockchain relies on several functionalities like integration with existing systems, request transactions in other systems, conduct transactions with other blockchains, integrating with applications, and provide easy switch between the platforms. Specifically, one blockchain network cannot meet a given transaction's needs in the supply chain, IoT,

healthcare, finance, etc. In reference [183], the authors propose a chain of things platform to provide identity, security, and interoperability in the blockchain-based IoT system. It presents three case studies to focus on interoperability, identity, and security features. The majority of interoperability solutions are based on public blockchain; therefore, there is a need to provide a blockchain solution that provides cross-communication between public networks, private networks, and private and public networks.

Smart Contract Vulnerability: To address the smart contract vulnerability, some tools should be deployed to test smart contracts' vulnerabilities or bugs before deployment to evaluate the smart contracts' security states [184]. In reference, [185], the authors present Hawk a privacy-preserving smart contract that ensures on-chain privacy and cryptographically hides the contract details from the public's view. Similarly, in [186], the authors propose a ChainSpace that gives privacy-preserving extensibility in the smart contract platform. The suggested platform provides a highly scalable and secure platform by using a sharding mechanism with the help of the S-BAC distributed atomic commit protocol.

Modification: In reference [187], authors develop building information modeling (BIM) that allows manipulation of information, data and management flow. It combines BIM as a cloud service to the blockchain network for the big data sharing process. The timestamp concept is utilized to trace and allow modification in history records. Similarly, in [188], the authors design a smart provenance system using distributed blockchain technology that prevents malicious modification in the captured data and utilizes the smart contract concept to record immutable data trials.

Complexity: The blockchain network relies on complex algorithms to provide security and establish consensus on the distributed network. Complex algorithms must run to prove that the user can update the chain, which requires much computing power and cost. Thus, to reduce the complexity of the consensus algorithms utilized in blockchain technology, some alternative consensus algorithms are developed, such as Proof of Stake, Proof of Burn, Proof of Importance, Proof of Capacity, Paxos, etc. That reduces the computational power requirement [189].

Blockchain technology is still a new technology in many domains, and new ways to employ it can still be found and researched. Each blockchain solution or system is formed for a specific platform and designed for a specific goal. Each recent advance or solution to overcome the existing challenges of each domain can be causal for the future direction. For further improvement, certain key issues are found in the above solutions that need to be focused on while using blockchain. These include the issues related to throughput and speed while maintaining the large volume of data, node synchronization issues, risk management problems, security challenges, encryption algorithms security issues, and regulatory-related problems. Although many solutions exist for the above issues still there is a need to provide more optimized, efficient, and reliable solutions for blockchain technology.

9 Conclusions

Blockchain technology is a decentralized, peer-to-peer network and the distributed ledger that is available to all nodes. Since 2008 Bitcoin and Blockchain have been the two most significant innovations in the information system. Numerous applications use blockchain to carry out operations in a trustworthy environment without any trusted authority. This study initially discusses the research method, followed by the introduction of blockchain. Then six applications where blockchain provides a solution to the existing centralized

architecture problems in a decentralized manner are analyzed. Some of the most potential domains have been outlined, such as IoT, healthcare, cloud computing, supply chain, etc. Also, we described the individual features that are mostly required for each application area. This allows the selection of blockchain and the appropriate structuring methods to the application's actual needs.

References

1. Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system.
2. Crosby, M., Pattanayak, P., Verma, S., & Kalyanaraman, V. (2016). Blockchain technology: Beyond bitcoin. *Applied Innovation 2*, 6–10.
3. Greenspan, G. (2015). Ending the bitcoin versus blockchain debate, http://www.multichain.com/blog/2015/07/bitcoin-vs-blockchain-debate
4. Barker, J. F., Chow-White, P. A., Adams, P. R., Mentanko, J., Ha, D., & Green, S. (2019). Blockchain as a disruptive technology for business: A systematic review. *International Journal of Information Management 51*, 102029.
5. Coin Market Cap, (2017). Cryptocurrency market capitalizations, https://coinmarketcap.com/.
6. Tschorsch, F., & Scheuermann, B. (2016). Bitcoin and beyond: A technical survey on decentralized digital currencies. *IEEE Communications Surveys & Tutorials 18*(3), 2084–2123.
7. Haferkorn, M., & Quintana Diaz, J. M. (2015). seasonality and interconnectivity within cryptocurrencies–an analysis on the basis of Bitcoin Litecoin and Namecoin. In A. Lugmayr (Ed.), *Enterprise applications and services in the finance Iidustry* (pp. 106–120). Springer.
8. Szabo, N. (1994). Smart contracts.
9. Szabo, N. (1997). The idea of smart contracts.
10. Zhao, J. L., Fan, S., & Yan, J. (2016). Overview of business innovations and research opportunities in blockchain and introduction to the special issue. *Financial Innovation, 2*(1), 28.
11. IBM. (2017). Three ways blockchain explorers chart a new direction, https://www-935.ibm.com/services/studies/csuite/pdf/GBE03835USEN-00.pdf.
12. IBM. (2017). 10 Key marketing trends for 2017, https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WRL12345USEN.
13. Tama, B. A., Kweka, B. J., Park, Y., & Rhee, K. H. (2017). A critical review of blockchain and its current applications. In: 2017 International Conference on Electrical Engineering and Computer Science (ICECOS) IEEE. pp. 109–113.
14. Zheng, Z., Xie, S., Dai, H.-N., & Wang, H. (2016). Blockchain challenges and opportunities: a survey. *International Journal of Web and Grid Services, 14*(4), 352–375.
15. Mohanta, B. K., Jena, D., Panda, S. S., & Sobhanayak, S. (2019). Blockchain technology: a survey on applications and security privacy challenges. *Internet of Things.* https://doi.org/10.1016/j.iot.2019.100107
16. Ali, T., Alzahrani, A., Jan, S., Siddiqui, M. S., Nadeem, A., & Alghamdi, T. (2019). A comparative analysis of blockchain architecture and its applications: problems and recommendations. *IEEE Access* https://doi.org/10.1109/access.2019.2957660
17. Casino, F., Dasaklis, T. K., & Patsakis, C. (2018). A systematic literature review of blockchain-based applications: current status, classification and open issues. *Telematics and Informatics* https://doi.org/10.1016/j.tele.2018.11.006
18. Feng, Q., He, D., Zeadally, S., Khan, M. K., & Kumar, N. (2018). A survey on privacy protection in blockchain system. *Journal of Network and Computer Applications*. https://doi.org/10.1016/j.jnca.2018.10.020
19. Brandão, A., Mamede, H. S., & Gonçalves, R. (2018). Systematic review of the literature, research on blockchain technology as support to the trust model proposed applied to smart places. In Á. Rocha, H. Adeli, L. P. Reis, S. Costanzo (Eds.), *Trends and advances in information systems and technologies. WorldCIST'18 2018. Advances in Intelligent Systems and Computing*, vol 745, (pp. 1163–1174). Cham: Springer. https://doi.org/10.1007/978-3-319-77703-0_113
20. Karafiloski, E., & Mishev, A. (2017). Blockchain solutions for big data challenges: a literature review. In: IEEE EUROCON 2017–17th International conference on smart technologies IEEE. pp. 763–768.
21. Khan, M. A., & Salah, K. (2017). IoT security: Review, blockchain solutions, and open challenges. *Future Generation Computing System, 82*, 395–411.

22. Meng, W., Tischhauser, E. W., Wang, Q., Wang, Y., & Han, J. (2018). When intrusion detection meets blockchain technology: Areview. *IEEE Access, 6*, 10179–10188.
23. Seebacher, S., & Schüritz, R. (2017). Blockchain technology as an enabler of service systems: A structured literature review. In Za S., Drăgoicea M., Cavallari M. (Eds.), *Exploring Services Science. IESS 2017. Lecture Notes in Business Information Processing,* vol 279, (pp. 12–23). Cham: Springer. https://doi.org/10.1007/978-3-319-56925-3_2
24. Hawlitschek, F., Notheisen, B., & Teubner, T. (2018). The limits of trust-free systems: A literature review on blockchain technology and trust in the sharing economy. *Electronic Commerce Research Applications, 29*, 50–63.
25. Yli-Huumo, J., Ko, D., Choi, S., Park, S., & Smolander, K. (2016). Where is current research on blockchain technology?–A systematic review. *PloS One, 11*(10), e0163477.
26. Sankar, L.S., Sindhu, M., & Sethumadhavan, M. (2017). Survey of consensus protocols on blockchain applications. In: 2017 4th International conference on advanced computing and communication systems (ICACCS) IEEE. pp. 1–5.
27. Atzei, N., Bartoletti, M., & Cimoli, T. (2017). A survey of attacks on Ethereum smart contracts (SoK). In: International conference on principles of security and trust Springer. pp. 164–186.
28. Bonneau, J., Miller, A., Clark, J., Narayanan, A., Kroll, J.A., & Felten, E.W. (2015). Sok: Research perspectives and challenges for bitcoin and cryptocurrencies. In: 2015 IEEE Symposium on security and privacy (SP) IEEE. pp. 104–121.
29. Tsukerman, M. (2015). The block is hot: A survey of the state of Bitcoin regulation and suggestions for the future. *Berkeley Tech LJ, 30*, 1127.
30. Mukhopadhyay, U., Skjellum, A., Hambolu, O., Oakley, J., Yu, L., & Brooks, R. (2016). A brief survey of cryptocurrency systems. In: 2016 14th Annual conference on privacy, security and trust (PST) IEEE. pp. 745–752.
31. Khalilov, M.C.K., & Levi, A. (2018). A survey on anonymity and privacy in bitcoin-like digital cash systems. IEEE Communications surveys &tutorials. pp. 1–1.
32. Conti, M., Sandeep Kumar, E., Lal, C., & Ruj, S. (2018). A survey on security and privacy issues of bitcoin. *IEEE Communications Surveys & Tutorials, 20*(4), 3416–3422.
33. Maxmen, A. (2018). AI researchers embrace Bitcoin technology to share medical data. *Nature, 555*(7696), 294.
34. Baynham-Herd, Z. (2017). Enlist blockchain to boost conservation. *Nature, 548*(7669), 523–523. https://doi.org/10.1038/548523c
35. Ahmed, S., & ten Broek, N. (2017). Blockchain could boost food security. *Nature, 550*(7674), 43–43. https://doi.org/10.1038/550043e
36. Konstantinidis, I., Siaminos, G., Timplalexis, C., Zervas, P., Peristeras, V., & Decker, S. (2018). Blockchain for business applications: A systematic literature review. In: International Conference on Business Information Systems. Springer, pp. 384–399. https://doi.org/10.1007/978-3-319-93931-5_28
37. Batubara, F. R., Ubacht, J., & Janssen, M. (2018). Challenges of blockchain technology adoption for e-government: A systematic literature review. In: 19th Annual International Conference on Digital Government Research: Governance in the Data Age. pp. 1–9. https://doi.org/10.1145/3209281.3209317
38. Agbo, C. C., Mahmoud, Q. H., & Eklund, J. M. (2019). Blockchain technology in healthcare: A systematic review. *Healthcare, 7*(2), 56. https://doi.org/10.3390/healthcare7020056
39. Lee, J. H., & Pilkington, M. (2017). How the blockchain revolution will reshape the consumer electronics industry [future directions]. *IEEE Consumer Electronics Magazine, 6*(3), 19–23. https://doi.org/10.1109/MCE.2017.2684916
40. Tama, B. A., Kweka, B. J., Park, Y., & Rhee, K. H. (2017). A critical review of blockchain and its current applications. In: International Conference on Electrical Engineering and Computer Science (ICECOS). pp. 109–113. https://doi.org/10.1109/ICECOS.2017.8167115
41. Zheng, Z., Xie, S., Dai, H., Chen, X., & Wang, H. (2017). An overview of blockchain technology: Architecture, consensus, and future trends. In: Big Data (BigData Congress). IEEE International Congress on IEEE. pp. 557–564. https://doi.org/10.1109/BigDataCongress.2017.85
42. Mosakheil, & Hayat. J. (2018). Security threats classification in blockchains. Culminating Projects in Information Assurance. 48. https://repository.stcloudstate.edu/msia_etds/48
43. Sankar, L. S., Sindhu, M., & Sethumadhavan, M. (2017). Survey of consensus protocols on blockchain applications. In: Advanced Computing and Communication Systems (ICACCS). 2017 4th International Conference on IEEE. https://doi.org/10.1109/ICACCS.2017.8014672

44. Lei, A., Cruickshank, H., Cao, Y., Asuquo, P., Ogah, C. P., & Sun, Z. (2017). Blockchain-based dynamic key management for heterogeneous intelligent transportation systems. *IEEE Internet of Things Journal, 4*(6), 1832–1843. https://doi.org/10.1109/JIOT.2017.2740569
45. Vyas, C. A., & Lunagaria, M. (2014). Security concerns and issues for bitcoin. National Conference cum Workshop on Bioinformatics and Computational Biology, NCWBCB-2014.
46. Wright, A., & Filippi, P. D. (2015). *Decentralized blockchain technology and the rise of lex cryptographia.* https://doi.org/10.2139/ssrn.2580664
47. Lin, C. I., & Liao, T. C. (2017). A survey of blockchain security issues and challenges. *International Journal of Network Security, 19*(5), 653–659. https://doi.org/10.6633/IJNS.201709.19(5).01
48. Li, X., Jiang, P., Chen, T., & Luo, X. (2017). A survey on the security of blockchain systems. *Future Generation Computer Systems* https://doi.org/10.1016/j.future.2017.08.020
49. Anascavage, Robert, and Davis, N. (2018). Blockchain technology: A literature review. SSRN Electronics Journal
50. Zhao, J. L., Fan, S., & Yan, J. (2011). Overview of business innovations and research opportunities in blockchain and introduction to the special issue. *Financial Innovation, 2*, 28.
51. Vishnevsky, V. P., & Chekina, V. D. (2018). Robot versus tax inspector or how the fourth industrial revolution will change the tax system: A review of problems and solutions. *Journal of Tax Reform, 4*, 6–26.
52. Shala, B., Trick, U., Lehamnn, A., Ghita, B., & Shiaeles, S. (2019). Novel trust consensus protocol and blockchain-based trust evaluation system for m2m application services. *Internet of Things, 7*, 100058.
53. Aly, M., Khomh, F., Haoues, M., Quintero, A., & Yacout, S. (2019). Enforcing security in internet of things frameworks: A systematic literature review. *Internet of Things., 6*, 100050.
54. Prada-Delgado, M. Á., Baturone, I., Dittmann, G., Jelitto, J., & Kind, A. (2020). PUF-derived IoT identities in a zero knowledge protocol for blockchain. *Internet of Things*. https://doi.org/10.1016/j.iot.2019.100057
55. Lee, In. (2019). The internet of things for enterprises: An ecosystem, architecture, and IoT service business model. *Internet of Things*. https://doi.org/10.1016/j.iot.2019.100078
56. Wang, Q., Zhu, X., Ni, Y., et al. (2019). Blockchain for the IoT and industrial IoT: A review. *Internet of Things*. https://doi.org/10.1016/j.iot.2019.100081
57. Grammatikis, P. I. R., Sarigiannidis, P. G., & Moscholios, I. D. (2019). Securing the Internet of Things: Challenges, threats and solutions. *Internet of Things*. https://doi.org/10.1016/j.iot.2018.11.003
58. Dorri, A., Kanhere, S. S., Jurdak, R., & Gauravaram, P. (2019). LSB: A lightweight scalable blockchain for IoT security and anonymity. *Journal of Parallel and Distributed Computing, 134*, 180–197.
59. Mohanty, S. N., Ramya, K. C., Rani, S. S., Gupta, D., Shankar, K., Lakshmanaprabu, S. K., & Khanna, A. (2020). An efficient lightweight integrated blockchain (ELIB) model for IoT security and privacy. *Future Generation Computer Systems, 102*, 1027–1037.
60. Qian, Y., Jiang, Y., Chen, J., Zhang, Y., Song, J., Zhou, M., & Pustišek, M. (2018). Towards decentralized IoT security enhancement: A blockchain approach. *Computers & Electrical Engineering, 72*, 266–273.
61. Minoli, D., & Occhiogrosso, B. (2018). Blockchain mechanisms for IoT security. *Internet of Things., 1–2*, 1–13.
62. Hammi, M. T., Hammi, B., Bellot, P., & Serhrouchni, A. (2018). Bubbles of trust: A decentralized blockchain-based authentication system for IoT. *Computers & Security 78*, 126–142. https://doi.org/10.1016/j.cose.2018.06.004
63. Alblooshi, M., Salah, K., & Alhammadi, Y. (2018). Blockchain-based ownership management for medical IoT (MIoT) Devices. In: International Conference on Innovations in Information Technology (IIT). https://doi.org/10.1109/innovations.2018.8606032
64. Frahat, R. T., Monowar, M. M., & Buhari, S. M. (2019). Secure and scalable trust management model for IoT P2P network. In: 2nd International Conference on Computer Applications & Information Security (ICCAIS). https://doi.org/10.1109/cais.2019.8769467
65. Yanez, W., Mahmud, R., Bahsoon, R., Zhang, Y., & Buyya, R. (2020). Data allocation mechanism for internet of things systems with blockchain. *IEEE Internet of Things Journal* https://doi.org/10.1109/jiot.2020.2972776
66. Li, D., Hu, Y., & Lan, M. (2020). IoT device location information storage system based on blockchain. *Future Generation Computer Systems, 109*, 95–102. https://doi.org/10.1016/j.future.2020.03.025
67. Zhou, L., Wang, L., Sun, Y., & Lv, P. (2018). BeeKeeper: A blockchain-based IoT system with secure storage and homomorphic computation. *IEEE Access.* https://doi.org/10.1109/access.2018.2847632

68. Ding, S., Cao, J., Li, C., Fan, K., & Li, H. (2019). A novel attribute-based access control scheme using blockchain for IoT. *IEEE Access.* https://doi.org/10.1109/access.2019.2905846
69. Novo, O. (2018). Blockchain meets IoT: An architecture for scalable access management in IoT. *IEEE Internet of Things Journal, 5*(2), 1184–1195. https://doi.org/10.1109/jiot.2018.2812239
70. Ouaddah, A. (2018). A blockchain based access control framework for the security and privacy of IoT with strong anonymity unlinkability and intractability guarantees. *Advances in Computers* https://doi.org/10.1016/bs.adcom.2018.11.001
71. Pal, S., Rabehaja, T., Hill, A., Hitchens, M., & Varadharajan, V. (2019). On the Integration of blockchain to the Internet of Things for enabling access right delegation. *IEEE Internet of Things Journal* https://doi.org/10.1109/jiot.2019.2952141
72. Ma, M., Shi, G., & Li, F. (2019). Privacy-oriented blockchain-based distributed key management architecture for hierarchical access control in the IoT scenario. *IEEE Access.* https://doi.org/10.1109/access.2019.2904042
73. Wang, E. K., Sun, R., Chen, C.-M., Liang, Z., Kumari, S., & Khurram Khan, M. (2020). Proof of X-repute blockchain consensus protocol for IoT systems. *Computers & Security* https://doi.org/10.1016/j.cose.2020.101871
74. Kim, T., Noh, J., & Cho, S. (2019). SCC: Storage compression consensus for blockchain in lightweight IoT network. In: IEEE International Conference on Consumer Electronics (ICCE). https://doi.org/10.1109/icce.2019.8662032
75. Li, S., Oikonomou, G., Tryfonas, T., Chen, T. M., & Da Li, Xu. (2014). A distributed consensus algorithm for decision making in service-oriented internet of things. *IEEE Transactions on Industrial Informatics, 10*(2), 1461–1468. https://doi.org/10.1109/tii.2014.2306331
76. Bai, H., Xia, G., & Fu, S. (2019). A two-layer-consensus based blockchain architecture for IoT. In: IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), pp. 1–6, https://doi.org/10.1109/ICEIEC.2019.8784458.
77. Pan, J., Wang, J., Hester, A., Alqerm, I., Liu, Y., & Zhao, Y. (2018). EdgeChain: An edge-IoT framework and prototype based on blockchain and smart contracts. *IEEE Internet of Things Journal* https://doi.org/10.1109/jiot.2018.2878154
78. Christidis, K., & Devetsikiotis, M. (2016). Blockchains and smart contracts for the internet of things. *IEEE Access., 4*, 2292–2303. https://doi.org/10.1109/access.2016.2566339
79. Salah, K., Suliman, A., Husain, Z., Abououf, M., & Alblooshi, M. (2018). Monetization of IoT data using smart contracts. *IET Networks*. https://doi.org/10.1049/iet-net.2018.5026
80. Fan, K., Bao, Z., Liu, M., Vasilakos, A. V., & Shi, W. (2019). Dredas: decentralized, reliable and efficient remote outsourced data auditing scheme with blockchain smart contract for industrial IoT. *Future Generation Computer Systems* https://doi.org/10.1016/j.future.2019.10.014
81. Zhou, Z., Liao, H., Gu, B., Mumtaz, S., & Rodriguez, J. (2019). Resource sharing and task offloading in IoT fog computing: A contract-learning approach. *IEEE Transactions on Emerging Topics in Computational Intelligence* https://doi.org/10.1109/tetci.2019.2902869
82. Griebel, L., Prokosch, H. U., Köpcke, F., Toddenroth, D., Christoph, J., Leb, I., Engel, I., & Sedlmayr, M. (2015). A scoping review of cloud computing in healthcare. *BMC Medical Informatics and Decision Making* https://doi.org/10.1186/s12911-015-0145-7
83. Jamoom, E., Yang, N., & Hing, E. (2016). Adoption of certified electronic health record systems and electronic information sharing in physician offices: United States, 2013 and 2014. US department of health and human services, Centers for disease control and prevention, National Center for Health Statistics, NCHS Data Brief. 1–8.
84. Bahga, A., & Madisetti, V. K. (2013). A cloud-based approach for interoperable electronic health records (EHRs). *IEEE Journal of Biomedical and Health Informatics, 17*(5), 894–906.
85. Fernández-Cardeñosa, G., de la Torre-Díez, I., López-Coronado, M., & Rodrigues, J. J. P. C. (2012). Analysis of cloud-based solutions on EHRs systems in different scenarios. *Journal of Medical Systems, 36*(6), 3777–3782.
86. Zangara, G., Corso, P. P., Cangemi, F., Millonzi, F., Collova, F., & Scarlatella, A. (2014). A cloud-based architecture to support electronic health record. *Studies in Health Technology and Informatics 207*, 380–389.
87. Saravanan, M., Shubha, R., Marks, A. M., & Iyer, V. (2017). SMEAD: Asecured mobile enabled assisting device for diabetics monitoring. In: IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS). pp. 1–6.
88. Liang, X., Zhao, J., Shetty, S., Liu, J., & Li, D. (2017). Integrating blockchain for data sharing and collaboration in mobile healthcare applications. In: IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC). pp. 1–5.

89. Patel, V. (2018). A framework for secure and decentralized sharing of medical imaging data via blockchain consensus. *Health Informatics Journal, 25*(4), 1398–1411.
90. Juneja, A., & Marefat, M. (2018). Leveraging blockchain for retraining deep learning architecture in patient-specific arrhythmia classification. In: IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Las Vegas, Nevada, USA, 393–397.
91. Griggs, K. N., Ossipova, O., Kohlios, C. P., Baccarini, A. N., Howson, E. A., & Hayajneh, T. (2018). Healthcare blockchain system using smart contracts for secure automated remote patient monitoring. *Journal of Medical Systems* https://doi.org/10.1007/s10916-018-0982-x
92. Wang, H., & Song, Y. (2018). Secure cloud-based EHR system using attribute based cryptosystem and blockchain. *Journal of Medical Systems* https://doi.org/10.1007/s10916-018-0994-6
93. Zhang, X., & Poslad, S. (2018). Blockchain support for flexible queries with granular access control to electronic medical records (EMR). In: IEEE International Conference on Communications (ICC). pp 1–6.
94. Badr, S., Gomaa, I., & Abd-Elrahman, E. (2018). Multi-tier blockchain framework for IoT-EHRs systems. *Procedia Computer Science, 141*, 159–166.
95. Guo, R., Shi, H., Zhao, Q., & Zheng, D. (2018). Secure attribute-based signature scheme with multiple authorities for blockchain in electronic health records systems. *IEEE Access., 6*, 11676–11686.
96. Zhang, J., Xue, N., & Huang, X. (2016). A secure system for pervasive social network-based healthcare. *IEEE Access., 4*, 9239–9250.
97. Brogan, J., Baskaran, I., & Ramachandran, N. (2018). Authenticating health activity data using distributed ledger technologies. *Computational and Structural Biotechnology Journal, 16*, 257–266.
98. Hussein, F., Arunkumar, N., Ramírez-González, G., Abdulhay, E., Tavares, J. M. R., & de Albuquerque, V. H. C. (2018). A medical records managing and securing blockchain based system supported by a genetic algorithm and discrete wavelet transform. *Cognitive Systems Research Journal, 52*, 1–11.
99. Zhao, H., Bai, P., Peng, Y., & Xu, R. (2018). Efficient key management scheme for health blockchain. *CAAI Transactions on Intelligence Technology, 3*(2), 114–118. https://doi.org/10.1049/trit.2018.0014
100. Al Omar, A., Bhuiyan, M. Z. A., Basu, A., Kiyomoto, S., & Rahman, M. S. (2019). Privacy-friendly platform for healthcare data in cloud based on blockchain environment. *Future Generation Computer Systems* https://doi.org/10.1016/j.future.2018.12.044
101. Tripathi, G., Ahad, M. A., & Paiva, S. (2019). S2HS-a blockchain based approach for smart healthcare system. *Healthcare*. https://doi.org/10.1016/j.hjdsi.2019.100391
102. Tanwar, S., Parekh, K., & Evans, R. (2020). Blockchain-based electronic healthcare record system for healthcare 4.0 applications. *Journal of Information Security and Applications 50*, 102407. https://doi.org/10.1016/j.jisa.2019.10240
103. Kumar, R., Marchang, N., & Tripathi, R. (2020). Distributed off-chain storage of patient diagnostic reports in healthcare system using ipfs and blockchain. In: International Conference on Communication Systems & NETworks (COMSNETS). https://doi.org/10.1109/comsnets48256.2020.9027313
104. Nguyen, D. C., Pathirana, P. N., Ding, M., & Seneviratne, A. (2019). Blockchain for secure EHRs sharing of mobile cloud based e-health systems. *IEEE Access.* https://doi.org/10.1109/access.2019.2917555
105. Zhang, X., Poslad, S., & Ma, Z. (2018). Block-Based access control for blockchain-based electronic medical records (EMRs) query in eHealth. In: IEEE Global Communications Conference (GLOBECOM). https://doi.org/10.1109/glocom.2018.8647433
106. Huang, J., Qi, Y. W., Asghar, M. R., Meads, A., & Tu, Y.-C. (2019). MedBloc: A blockchain-based secure EHR system for sharing and accessing medical data. In: 18th IEEE International Conference On Trust, Security and Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). https://doi.org/10.1109/trustcom/bigdatase.2019.00085
107. Azaria, A., Ekblaw, A., Vieira, T., & Lippman, A. (2016). MedRec: Using blockchain for medical data access and permission management. 2016 2nd International Conference on Open and Big Data (OBD). https://doi.org/10.1109/obd.2016.11
108. Zheng, K., Liu, Y., Dai, C., Duan, Y., & Huang, X. (2018). Model checking PBFT consensus mechanism in healthcare blockchain network. In: 9th International Conference on Information Technology in Medicine and Education (ITME). https://doi.org/10.1109/itme.2018.00196
109. Ray, P. P., Dash, D., Salah, K., & Kumar, N. (2020). Blockchain for IoT-based healthcare: Background, consensus, platforms, and use Cases. *IEEE Systems Journal* https://doi.org/10.1109/jsyst.2020.2963840
110. Pham, H. L., Tran, T. H., & Nakashima, Y. (2018). A secure remote healthcare system for hospital using blockchain smart contract. In: IEEE Globecom Workshops (GC Wkshps), Abu Dhabi, United Arab Emirates, 1–6, https://doi.org/10.1109/GLOCOMW.2018.8644164.

111. Novikov, S. P., Kazakov, O. D., Kulagina, N. A., & Azarenko, N. Y. (2018). Blockchain and smart contracts in a decentralized health infrastructure. In: IEEE International Conference quality management, transport and information security, information technologies (IT&QM&IS). doi:https://doi.org/10.1109/itmqis.2018.8524970
112. Wang, C., Ren, H., & Wang, J. (2016). Secure optimization computation outsourcing in cloud computing: A case study of linear programming. *IEEE Transactions on Computers, 65*(1), 216–229.
113. Xu, C., Wang, K., & Guo, M. (2017). Intelligent resource management in blockchain-based cloud data centers. *IEEE Cloud Computing 4*(6), 50–59.
114. Park, J. H., & Park, J. H. (2017). Blockchain security in cloud computing: Use cases, challenges and solutions. *Symmetry*. https://doi.org/10.3390/sym9080164
115. Yu, G., Wang, X., Yu, K., Ni, W., Zhang, J. A., & Liu, R. P. (2019). Survey: Sharding in blockchains. *IEEE Access*. https://doi.org/10.1109/ACCESS.2020.2965147
116. Xie, S., Zheng, Z., Chen, W., Wu, J., Dai, H. N., & Imran, M. (2019). Blockchain for cloud exchange: A survey. *Computers & Electrical Engineering* https://doi.org/10.1016/j.compeleceng.2019.106526
117. Yang, R., Yu, F. R., Si, P., Yang, Z., & Zhang, Y. (2018). Integrated blockchain and edge computing systems: A survey, some research issues and challenges. *IEEE Communications Surveys & Tutorials* https://doi.org/10.1109/COMST.2019.2894727
118. Liu, Bin, Yu, X. L., Chen, S., Xu, X., & Zhu, L. (2017). Blockchain based data integrity service framework for IoT data. In: IEEE International conference on web services (ICWS). 468–475. https://doi.org/10.1109/ICWS.2017.54
119. Wang, H., Wang, X. A., Xiao, S., & Zhou, Z. (2019). Blockchain-based public auditing scheme for shared data. In: International Conference on innovative mobile and internet services in Ubiquitous computing. Springer, pp. 197–206. https://doi.org/10.1007/978-3-030-22263-5_19
120. Zhang, Y., Xu, C., Ni, J., Li, H., & Shen, X. S. (2019). Blockchain-assisted public-key encryption with keyword search against keyword guessing attacks for cloud storage. *IEEE Transactions on Cloud Computing* https://doi.org/10.1109/TCC.2019.2923222
121. Chen, Y., Ding, S., Xu, Z., Zheng, H., & Yang, S. (2019). Blockchain-based medical records secure storage and medical service framework. *Journal of Medical Systems, 43*(1), 5. https://doi.org/10.1007/s10916-018-1121-4
122. Cao, S., Zhang, G., Liu, P., Zhang, X., & Neri, F. (2019). Cloud-assisted secure eHealth systems for tamper-proofing EHR via blockchain. *Information Sciences, 485*, 427–440. https://doi.org/10.1016/j.ins.2019.02.038
123. Hasan, S. S., Sultan, N. H., & Barbhuiya, F. A. (2019). Cloud data provenance using ipfs and blockchain technology. In: 7th International workshop on security in cloud computing. 5–12. https://doi.org/10.1145/3327962.3331457
124. Chen, W., Chen, Y., Chen, X., & Zheng, Z. (2019). Toward secure data sharing for the IoV: A quality-driven incentive mechanism with on-chain and off-chain guarantees. *IEEE Internet of Things Journal* https://doi.org/10.1109/JIOT.2019.2946611
125. Do, H. G., & Ng, W. K. 2017. Blockchain-based system for secure data storage with private keyword search. In: IEEE World Congress on Services (SERVICES). 90–93. https://doi.org/10.1109/SERVICES.2017.23
126. Li, J., Wu, J., & Chen, L. (2018). Block-secure: Blockchain based scheme for secure P2P cloud storage. *Information Sciences, 465*, 219–231. https://doi.org/10.1016/j.ins.2018.06.071
127. Cruz, J. P., Kaji, Y., & Yanai, N. (2018). RBAC-SC: Role-based access control using smart contract. *IEEE Access, 6*, 12240–12251. https://doi.org/10.1109/ACCESS.2018.2812844
128. Wang, S., Wang, X., & Zhang, Y. (2019). A secure cloud storage framework with access control based on blockchain. *IEEE Access*. https://doi.org/10.1109/ACCESS.2019.2929205
129. Wang, S., Zhang, Y., & Zhang, Y. (2018). A blockchain-based framework for data sharing with fine-grained access control in decentralized storage systems. *IEEE Access, 6*, 38437–38450. https://doi.org/10.1109/ACCESS.2018.2851611
130. Tosh, D., Shetty, S., Liang, X., Kamhoua, C., & Njilla, L. L. (2019). Data provenance in the cloud: A blockchain-based approach. *IEEE Consumer Electronics Magazine, 84*, 38–44. https://doi.org/10.1109/MCE.2019.2892222
131. Zhu, Z., Qi, G., Zheng, M., Sun, J., & Chai, Y. (2019). Blockchain based consensus checking in decentralized cloud storage. *Simulation Modelling Practice and Theory*. https://doi.org/10.1016/j.simpat.2019.101987
132. Kumar, G., Saha, R., Rai, M. K., Thomas, R., & Kim, T. H. (2019). Proof-of-work consensus approach in blockchain technology for cloud and fog computing using maximization-factorization statistics. *IEEE Internet of Things Journal* https://doi.org/10.1109/jiot.2019.2911969

133. Sukhodolskiy, I., & Zapechnikov, S. (2018). A blockchain-based access control system for cloud storage. In: IEEE Conference of Russian young researchers in electrical and electronic engineering (EIConRus), Moscow. 1575–1578. https://doi.org/10.1109/EIConRus.2018.8317400
134. Yu, H., & Yang, Z. (2018). Decentralized and smart public auditing for cloud Storage. In: IEEE 9th International conference on software engineering and service science (ICSESS). pp. 491–494. IEEE. https://doi.org/10.1109/ICSESS.2018.8663780
135. Li, Jingyi, Wu, J., Chen, L., & Li, J. (2018). Deduplication with blockchain for secure cloud storage. In: CCF Conference on big data, Springer. pp. 558–570. https://doi.org/10.1007/978-981-13-2922-7_36
136. Francisco, K., & Swanson, D. (2018). The supply chain has no clothes: Technology adoption of blockchain for supply chain transparency. *Logistics, 2*(1), 2.
137. Chen, G., Xu, B., Lu, M., & Chen, N. S. (2018). Exploring blockchain technology and its potential applications for education. *Smart Learning Environments, 5*(1), 1.
138. Leng, K., Bi, Y., Jing, L., Fu, H.-C., & Van Nieuwenhuyse, I. (2018). Research on agricultural supply chain system with double chain architecture based on blockchain technology. *Future Generation Computer Systems, 86*, 641–649. https://doi.org/10.1016/j.future.2018.04.061
139. Helo, P., & Hao, Y. (2019). Blockchains in operations and supply chains: A model and reference implementation. *Computers & Industrial Engineering* https://doi.org/10.1016/j.cie.2019.07.023
140. Bechtsis, D., Tsolakis, N., Bizakis, A., & Vlachos, D. (2019). A blockchain framework for containerized food supply chains. In: 29th European Symposium on Computer Aided Process Engineering. 1369–1374. https://doi.org/10.1016/b978-0-12-818634-3.50229-0
141. Yadav, V. S., Singh, A. R., Raut, R. D., & Govindarajan, U. H. (2020). Blockchain technology adoption barriers in the Indian agricultural supply chain: An integrated approach. *Resources, Conservation and Recycling, 161*, 104877. https://doi.org/10.1016/j.resconrec.2020.104877
142. Dwivedi, S. K., Amin, R., & Vollala, S. (2020). Blockchain based secured information sharing protocol in supply chain management system with key distribution mechanism. *Journal of Information Security and Applications, 54*, 102554. https://doi.org/10.1016/j.jisa.2020.102554
143. Xiong, F., Xiao, R., Ren, W., Zheng, R., & Jiang, J. (2019). A key protection scheme based on secret sharing for blockchain-based construction supply chain system. *IEEE Access, 7*, 126773–126786. https://doi.org/10.1109/access.2019.2937917
144. Zhang, X., Sun, P., Xu, J., Wang, X., Yu, J., Zhao, Z., & Dong, Y. (2020). Blockchain-based safety management system for the grain supply chain. *IEEE Access* https://doi.org/10.1109/access.2020.2975415
145. Nakasumi, M. (2017). Information sharing for supply chain management based on block chain technology. In: IEEE 19th Conference on business informatics (CBI). https://doi.org/10.1109/cbi.2017.56
146. Gandhi, G. M., & Salvi. (2019). Artificial intelligence integrated blockchain for training autonomous cars. In: Fifth International Conference on science technology engineering and mathematics (ICONSTEM). https://doi.org/10.1109/iconstem.2019.8918795
147. Singh, S. K., Rathore, S., & Park, J. H. (2019). BlockIoTIntelligence: A blockchain-enabled intelligent IoT architecture with artificial intelligence. *Future Generation Computer Systems* https://doi.org/10.1016/j.future.2019.09.002
148. Lobo, V. B., Analin, J., Laban, R. M., & More, S. S. (2020). Convergence of blockchain and Artificial Intelligence to decentralize healthcare systems. In: Fourth international conference on computing methodologies and communication (ICCMC). https://doi.org/10.1109/iccmc48092.2020.iccmc-000171
149. Teerapittayanon, S., & Kung, H. T. (2019). DaiMoN: A decentralized artificial intelligence model network. In: IEEE International conference on blockchain (Blockchain). https://doi.org/10.1109/blockchain.2019.00026
150. Wang, K., Dong, J., Wang, Y., & Yin, H. (2019). Securing data with blockchain and AI. *IEEE Access* https://doi.org/10.1109/access.2019.2921555
151. Zyskind, G., Nathan, O., & Pentland, A. S. (2015). Decentralizing privacy: Using blockchain to protect personal data. In: IEEE Symposium on security and privacy. pp. 180–184.
152. Bhowik, D., & Feng, T. (2017). The multimedia blockchain: A distributed and tamper-proof media transaction framework. In: 22nd International conference on digital signal processing (DSP), London. pp. 1–5.
153. Torre, I., Koceva, F., Sanchez, O. R., & Adorni, G. (2016). A framework for personal data protection in the loT. In: 11th International conference for internet technology and secured transactions. pp. 384–391.

154. Farkas, C., & Stoica, A. G. (2004). Correlated data inference. Data and applications security XVII. Springer US. p. 119132.
155. https://en.wikipedia.org/wiki/Game_of_Thrones_(season_7) Retrieved : June, 2018
156. Ghimire, S., Choi, J. Y., & Lee, B. (2019). Using blockchain for improved video integrity verification. . *IEEE Transactions on Multimedia* https://doi.org/10.1109/tmm.2019.2925961
157. Vishwa, A., & Hussain, F. K. (2018). A Blockchain based approach for multimedia privacy protection and provenance. In: IEEE symposium series on computational intelligence (SSCI). https://doi.org/10.1109/ssci.2018.8628636
158. Jnoub, N., & Klas, W. (2019). Detection of tampered images using blockchain technology. In: IEEE International conference on blockchain and cryptocurrency (ICBC). https://doi.org/10.1109/bloc.2019.8751300
159. Chavan, S., Ghuge, S., Warke, P., & Deolek, R. V. (2019). Music streaming application using blockchain. *6th International Conference on computing for sustainable global development (INDIACom)* (pp. 1035–1040). New Delhi.
160. Chalmersa, D., Matthewsb, R., & Hyslopb, A. (2018). Blockchain as an external enabler of new venture ideas: Digital entrepreneurs and the disintermediation of the global music industry. Journal of Business Research.
161. Taylor, P. J., Dargahi, T., Dehghantanha, A., Parizi, R. M., & Choo, K. R. (2019). A systematic literature review of blockchain cyber security. *Digital Communications Network,* https://doi.org/10.1016/j.dcan.2019.01.005
162. Dabbagh, M., Sookhak, M., & Safa, N. S. (2019). The evolution of blockchain: A bibliometric study. *IEEE Access, 7*, 19212–19221. https://doi.org/10.1109/ACCESS.2019.2895646
163. Kamran, M., Khan, H. U., Nisar, W., Farooq, M., & Rehman, S. U. (2020). Blockchain and Internet of Things: A bibliometric study. Computers and Electrical Engineering, 81.
164. Rejeb, A., Trieblmaier, H., Rajeb, K., & Zailani, S. (2021). Blockchain research in healthcare: A bibliometric review and current research trends. *Journal of Data, Information and Management, 3*, 109–124. https://doi.org/10.1007/s42488-021-00046-2
165. Chang, S. E., & Chen, Y. (2020). When blockchain meets supply chain: A systematic literature review on current development and potential applications. *IEEE Access, 8*, 62478–62494. https://doi.org/10.1109/ACCESS.2020.2983601
166. Benisi, N. Z., Aminian, M., & Javadi, B. (2020). Blockchain-based decentralized storage networks: A survey. *Journal of Network and Computer Applications, 162*, 102656.
167. Bigchaindb: The scalable blockchain database powering ipdb., https://www.bigchaindb.com/, 2017.
168. Blockchain trends, https://www.leewayhertz.com/blockchain-trends/
169. Eyal, I., Gencer, A. E., Sirer, E. G., Van Renesse, R. (2016). Bitcoin-ng: A scalable blockchain protocol. In: 13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16), Santa Clara, CA, USA. pp. 45–59.
170. Litecoin, (2018). Retrieved from 2018–02–04 https://litecoin.org/, 2011.
171. Sompolinsky, Y., & Zohar, A. (2013). Accelerating bitcoin's transaction processing, fast money grows on trees, not chains. IACR Cryptology ePrint Archive 881.
172. Decker, C., & Wattenhofer, R. (2015). A fast and scalable payment network with bitcoin duplex micropayment channels. In: Symposium on self-stabilizing systems, Edmonton, AB, Canada, Springer. pp. 3–18.
173. Kosba, A., Miller, A., Shi, E., Wen, Z., & Papamanthou, C. (2016). Hawk: The blockchain model of cryptography and privacy-preserving smart contracts. *Security and Privacy (SP), 2016 IEEE Symposium on* (pp. 839–858). CA, USA, IEEE.
174. Zyskind, G. Nathan, O., Pentland, A. (2015). Enigma: Decentralized computation platform with guaranteed privacy. arXiv preprint
175. Quorum whitepaper, (2016). Available online: https://github.com/jpmorganchase/quorum-docs/blob/master/Quorum%20Whitepaper%20v0.1.pdf.
176. Androulaki, E., Barger, A., Bortnikov, V., Cachin, C., Christidis, K., De Caro, A., Enyeart, D., Ferris, C., Laventman, G., & Manevich, Y. (2018). Hyperledger fabric: A distributed operating system for permissioned blockchains. arXiv preprint
177. Eyal, I., & Sirer, E. G. 2014. Majority is not enough: Bitcoin mining is vulnerable. In: International Conference on Financial Cryptography and Data Security, San Juan, Puerto Rico, Springer. pp. 436–454.
178. Bonneau, J., Felten, E. W., Goldfeder, S., Kroll, J. A., & Narayanan, A. (2016). Why buy when you can rent? Bribery attacks on bitcoin consensus.
179. Bitlegal, (2017). http://bitlegal.io/.

180. R3, (2018). https://www.r3.com/.
181. Trusted IoT alliance, (2017). https://www.trustediot.org/.
182. Alastria: National blockchain ecosystem, (2018). https://alastria.io/,
183. Chain of things, (2017). https://www.chainofthings.com/.
184. Li, X., Jiang, P., Chen, T., Luo, X., Wen, Q. (2017). A survey on the security of blockchain systems. Future Generation Computer Systems. (In press)
185. Kosba, A., Miller, A., Shi, E., Wen, Z., Papamanthou, C. (2016). Hawk: The blockchain model of cryptography and privacy-preserving smart contracts. In: 2016 IEEE Symposium on security and privacy (SP). pp. 839–858. doi:https://doi.org/10.1109/SP.2016.55.
186. Al-Bassam, M., Sonnino, A., Bano, S., Hrycyszyn, D., & Danezis, G. (2017). Chainspace: A sharded smart contracts platform. arXiv preprint
187. Zheng, R., Jiang, J., Hao, X., Ren, W., Xiog, F., & Ren, Y. (2019). bcBIM: A blockchain-based big data model for bim modification audit and provenance in mobile cloud. *Mathematical Problem in Engineering, 2019*, 1–13.
188. Ramachandran, A., & Kantarcioglu, M. (2018). SmartProvenance: A distributed, blockchain based dataprovenance system. In: Eight ACM conference of data and application security and privacy. pp. 35–42.
189. Nguyen, G. T., & Kim, K. (2018). A survey about consensus algorithms used in blockchain. *Journal of Information Processing System, 14*(1), 101–128.

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



Ms. Pratima Sharma is a Research Scholar, Department of Computer Science & Engineering at Delhi Technological University, India. Prior to this she worked as an Assistant Professor, at Inderprastha Engineering College, Ghaziabad, for nearly three years. She received the M.Tech. and B.Tech. degrees in Computer Science and Engineering from Guru Gobind Singh Indraprastha University, Delhi, India, in 2013 and 2015, respectively. She developed HoneyDos Application for preventing system from Malicious Packets and Denial of Service Attack Using Support Vector Machine Technique during the Post—Graduation degree. Her research and publication interests include Blockchain Technology, Honeypot, Network Security, Information Security and Data Mining. She has presented papers at International/National conferences, published articles and papers in various journals.



Dr. Rajni Jindal is working as Professor and Head at Computer Engineering Department, Delhi Technological University, Delhi. She received her M.E. from Delhi College of Engineering. She completed her Ph.D. (Computer Engineering) from Faculty of Technology, Delhi University, Delhi. She also worked as Professor (IT), Dean (Research and Collaboration) at Indira Gandhi Delhi Technical University for women, Delhi for 3 years. She possesses a work experience of around 31 years in research and academics. Her major area of interests are Database Systems, Data Mining, Operating Systems and Compiler Design. She has authored around 140 research papers and articles for various national and international journals/conferences and also 5 books. She is a senior member of IEEE and life member of CSI.



Dr. Malaya Dutta Borah is working as an Assistant Professor in the Department of Computer Science & Engineering at National Institute of Technology (NIT) Silchar, Assam. Her research areas are Blockchain Technology, Data Mining and Cloud Computing. She has authored/co-authored around 45 research papers in International/National Journals/Conferences/Books of repute. She has published 2 patents. She is the Editor of three upcoming edited books (publishers-CRC press, IGI Global and Springer) in the field of Blockchain technology. She is Treasure of IEEE Silchar subsection. She has experience of organising International conferences and workshops. She is a member of IEEE, CSI, ACM and MIR Lab.

2021 8th International Conference on Signal Processing and Integrated Networks (SPIN) | 978-1-6654-3564-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SPIN52536.2021.9566135

A NOVEL LOW POWER DESIGN OF BRENT KUNG ADDER HAVING FAULT TOLERANT CAPABILITY

Sumit Pahuja
Department of Electronics and Communiaction
Delhi Technological University
Delhi, India
spahuja25121992@gmail.com

Prof. Gurjit Kaur
Department of Electronics and Communiaction
Delhi Technological University
Delhi, India
gurjeet_kaur@rediffmail.com

***Abstract*— In today's technology, Parallel Prefix Adders is widely used in VLSI to get higher delay performance. The demand for adders that use reversible logic gates is growing every day as technology advances. However, since reversible logic gates have no information loss, reversible circuits consume less heat than conventional circuits. Reversible logic gates along with extra characteristics that input parity and output parity are equal is termed as PPRLG (parity preserving reversible logic gates). This characteristic allows any device to function normally even if any intermediary nodes fail. The Brent-Kung adder offers a faster response time as compared to many adders. As a consequence, designing an extremely fast, fault-tolerable, and low-power adder immediately results in increased device speed for quicker computing purposes, so Brent Kung adder with the help of PPRLG is presented in our work. The suggested Adder implementation is done using Verilog programming language in the Xilinx software with version 14.7. For the purpose of demonstration of its quality, presented Parity Preserving Reversible Adder, including many sub-parts, we used a number of parameters like Quantum-Cost, CI, GC, and Unused Output, including the schematic (Register Transfer Level) of the Brent-Kung Adder and its sub-parts. So yet, no Brent-Kung adder circuit employing Reversible Logic Gates with Fault-Tolerant capabilities has been suggested, to the best of our knowledge.**

Keywords— RLG, PPRLG, Parallel Prefix adder, Fault Tolerant, Brent-Kung Adder, Parity Preserving, Power, QC, Reversible Logic

I. INTRODUCTION

Not just in ALU, but also in other parts of processors, adders play a crucial role in mathematical computation. There are several distinct types of adder topologies in the research. Parallel prefix adder's efficiency is better as compared to others. We live in a fast-paced technological world where gadget size, speed, and power dissipation are critical, and delays and faults in any operation are not acceptable. Microcontrollers, signal processing applications, and a range of embedded devices all rely on high-speed parallel-prefix adders [1-3]. In the case of PPA (parallel prefix adders), speed is faster as carry signals are generated simultaneously. Parallel prefix adders come in a variety of flavors, the famous one is BKA (Brent Kung adder) and another one is KSA (Kogge-tone adder). BKA's performance is slow as compared to KSA since it has a greater number of prefix stages as compared to KSA prefix stages. The area of BKA is less and also consumes less energy because it has fewer Carry Merge blocks. Moreover, there is significantly less wire traffic as compared to the Kogge Stone adder [4].

Energy is dissipated in the form of bits of information in CMOS devices. This energy dissipation takes the form of power dissipation, and it's crucial in application required low power for its operation. Nowadays majority of the digital circuits is construct using RLG (Reversible Logic Gates). When compared to their combinational equivalents, reversible circuits dissipate less energy. There is generally bit-loss in any conventional combinational logic. One can take example of basic OR gate having two inputs. It will give one output, so there is loss of one bit here which is dissipated as heat. As a result, preserving bits will assist us in preserving energy [5][6]. This is in the case of irreversible computation and according to Landaeur theorem, which states that there is always a minimal amount of energy consumed in any irreversible computation for each bit of information loss which may be expressed as:

$$Q \geq KTln2 \qquad (1)$$

K has a defined value (1.380065×10^{-23} $JK^{-1}(m^2\,kg^{-2}\,K^{-1})$) is called Boltzmann's constant whereas thermodynamic temperature or also called absolute temperature in Kelvin is depicted by T [7]. But according to Bennet, Power loss can be ideally 0, if one will use reversible concepts [8]. The goal of this study is to construct a low-power Brent Kung adder using a reversible logic gate with its special qualities, less area, fault tolerance capability, and high performance. Numerous Adders developed by many experts to design the adder superior in terms of speed, power, and size from period to period, by employing a range of technologies. Various adders are designed using different approaches in the research [9-11]. Overall speed of any system like processor etc. depends on the speed of adders and so with respect of power consumption and speed, it must be superb.

According to paper [12], Computation nodes are less in Brent Kung adder when one compare with KSA and also area is less in case of brent kung adder but have more delay as compared to KSA. Area occupied by KSA and Sklansky is larger than BKA. Structure of BKA is simpler with respect to KSA and BKA's fan-out with respect to Sklansky adder. Because of various features, Brent Kung adder is excellent. and one can reduce its power dissipation by incorporating RLG. We included ability to tolerate fault with reversible concepts. So, there is various application where one can use this adder for better efficiency [13-14]. In this research paper, a 16-bit Brent Kung adder through making use of PPRLG (Parity-Preserving Reversible Logic Gates) is introduced and no Brent Kung adder circuit by using PPRLG has been proposed thus far. Because of PPRLG, BKA design have characteristics that is fault tolerable, so more reliable.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:35:41 UTC from IEEE Xplore. Restrictions apply.

The following is how the structure of the paper is organised. There are total five sections to this paper. Introduction part is included in 1st section and also gives a brief literature review of the earlier work. Section 2 focussing on basic concepts related to RLG and PPRLG. Section 3 explains the proposed BKA implemented with the help of PPRLG. Results of simulation in terms of waveforms are discussed in section number 4. Finally, in 5th section, the conclusion and future work are presented.

II. REVRESIBLE LOGIC BASICS

In this part, we will discuss about essential reversible logic ideas. The classical way of computation is inherently irreversible in nature. Consider the following scenario. Two input AND gate one can say is irreversible in nature. For any input combination, loss of single bit information is unavoidable. because one can get only one output by giving two inputs. And if output is 1, no one will be able to declare that the output value is 1.for of two inputs as 1 and 1 / 1 and 0 / 0 and 1 i.e. there is no possibility to recover inputs from output. There is always dissipation of energy according to [7]. But if one will use RLG then there are no chances loss of information and that's why dissipation of power is less. A system will be called reversible if at all abstraction level, there is reversibility.

A. Reversible Logic Gates

When adopting reversible concept, a circuit or system must have identical number of O/P's & I/P's with one to one correspondence between O/P's and I/P's [15]. Because of this property, one can get inputs uniquely from outputs. When building a reversible circuit, there are a few guidelines to keep in mind:

- The number of I/Ps and O/Ps should be the same, with as minimal reversible logic gates as possible.
- Inputs are uniquely mapped to outputs and so Inputs retrieved from Outputs uniquely.
- Garbage output and Constant Inputs value should be kept as less as possible.
- There will be no fan-out or feedback from output to input side. [16][17].

PPRLG are extended version of RLG with same parity at input and output side.

A few of the Reversible Gates features that design engineers employ to enhance a reversible system are listed below:

- Garbage Outputs:

  There used to remain need to accomplish the functionality of RLG and because of this some outputs remain unused that defines Garbage Outputs [18].

- Quantum Cost (QC):

  With respect of number of elementary gates used (like NOT gate or CNOT gate or Controlled-V+ gate or Controlled-V gate) in realization of Circuits; one can define Quantum Cost [20].

- Gate Count (GC):

  The quantity of reversible design elements essential to build the any design is considered as Gate Count. [18].

- Constant Input (CI):

  Constant Inputs is another name. To accomplish the desired functionality, some inputs needs to remain at constant logic zero or one [18].

Number of simple RLG along with commutative number of I/P's-O/P's and QC is mentioned in figure 1.

B. Parity Preserving Reversible Gates

If any circuit polarity is same both at input and output then one can say circuit have fault tolerant capability. By using reversibility property, one can say there is no bit loss or information loss but one can detect bit error if will preserve parity. But by using PPRLG, circuit implemented will have ability to tolerate fault. In any circuit throughout the calculation, the input data's parity will be retained. Hamming weight of inputs and Outputs are same or one can say EXOR operation on all inputs and EXOR operation on all outputs will be same in case of PPRLG [15]. Some examples of simple PPRLG in literature are mentioned in figure 2:

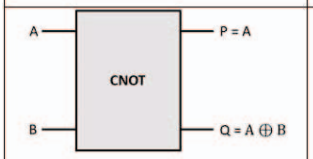
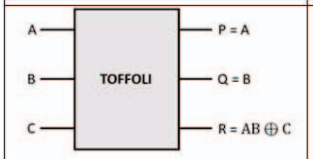
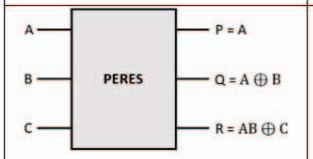
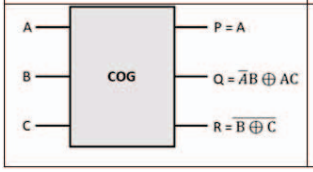
Reversible Logic Gates	No of Inputs & Outputs	Quantum Cost
	2	1
	3	5
	3	4
	3	4

Fig.1. Gates having Reversible Concepts

III. BRENT-KUNG ADDER

In case of Ripple carry adder, first sum bit is not possible until carry input is not present and 2nd sum bit is not available till the time preceding carry is available and so on. One can say, carry have to ripple through each level, so it is time consuming adder as not all sum bits are available at the same time. To overcome this issue, carry look ahead adder comes into picture in which parallel generation carry is possible at each level and so no rippling effect. Larger the size of input

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:35:41 UTC from IEEE Xplore. Restrictions apply.

binary sequence to be added, more the complexity of carry increases. As a result, higher order CLA is harder to implement. To overcome this issue, there come Parallel Prefix adder into picture [19].
In-between phases, Carries would be available prior with the help of CP (Carry propagate) and CG (Carry generate) no matter what Cin is, to add every bit of input binary sequence. The complexity of Parallel Prefix adder is considered to be logarithmic.

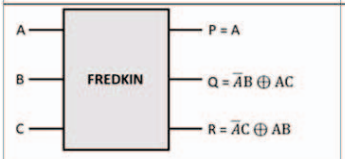
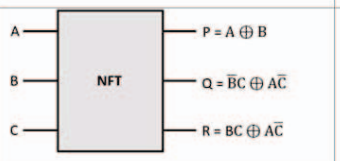
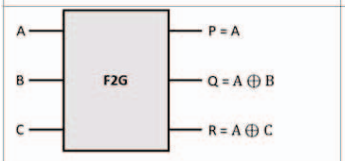
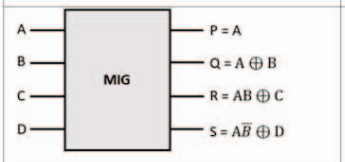
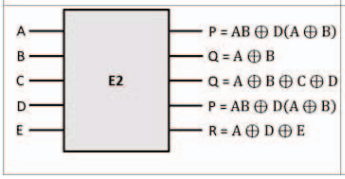
Parity Preserving Reversible Logic Gate	No of Inputs & Outputs	Quantum Cost
	3	5
	3	5
	3	2
	4	7
	5	8

Fig.2. Parity Preserving Reversible Logic Gates

Brent kung adder falls in category of PPA. A schematic of General 16-bit adder where 16 bits inputs and Carry Inputs is shown in left side of figure 3 [20]. Here Sum and Cout are 16-bit output sum and output carry of inputs respectively.

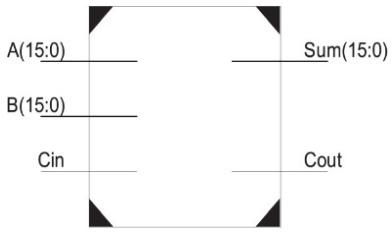


Fig.3. 16 Bit Full Adder block diagram

Figure 4 depicts a 16-bit BKA block diagram. There are mainly 3 stages in parallel Prefix adder's operation are explained as:

- Pre computation stage:

Using the equations given below, this stage generates two signals i.e. Pi and Gi are called Propagate signals and Generate signals respectively.

$$P_i = A_i \oplus B_i \qquad (2)$$

$$G_i = A_i B_i \qquad (3)$$

Rectangular-block shown in figure 4 is responsible for this stage.

- Prefix tree stage:

The prefix signals for each bit are integrated to produce carry signals. Here two blocks are working which are circular block (both differently represented by different size). Two signal merges propagate and generate signals are produces by Large Circular block which are defined by following equations:

$$G = G_i + P_i . G_i' \qquad (4)$$

$$P = P_i . P_i' \qquad (5)$$

Here new generate Block is producing by using two different generate signal represented G_i and G_i' as from two distinct GP block. Similarly, new propagate block producing by different propagate signals P_i and P_i' from two different GP block. Circular block (smaller in size) produces a merged generate signal only represented as:

$$G = G_i + P_i . G_i' \qquad (6)$$

- Summation stage:

This stage produces final sum bit result which defined by following equation:

$$S_i = P_i \oplus C_{i-1} \qquad (7)$$

This stage is implemented by triangular block in figure 4.

A. Proposed Low Power BKA design using PPRLG

Brent Kung adder with the same behavior as indicated above by employing PPRLG is presented in this part. As a consequence, the suggested circuit is fault-tolerable, high-performable and having low power loss which has not been addressed in the literature too far. The Proposed adder design can be used in various components and devices like for energy efficient technologies [21-25]. One can simulate the digital design using various simulators to check its logical functionality, but for our proposed design along with its various blocks we used Xilinx ISE 14.7 version because of its major feature i.e., user friendly.
Figure 5 shows a Brent Kung adder RTL schematic that has been implemented with PPRLG. Various different blocks of figure 5 is implemented by various PPRLG is mentioned in figure 6 along with various parameters and RTL schematic of each block is shown in figure 7, 8, 9 and 10.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:35:41 UTC from IEEE Xplore. Restrictions apply.

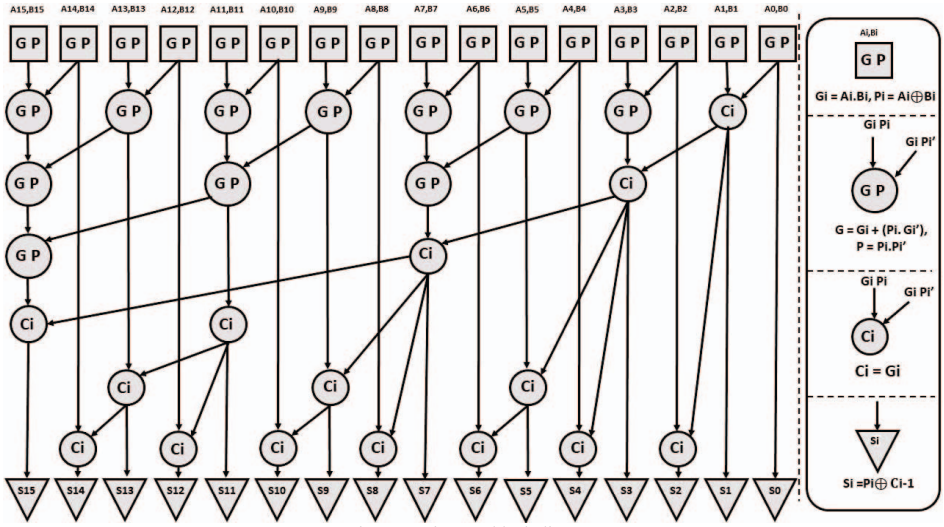


Fig.4. 16 Bit BKA block diagram

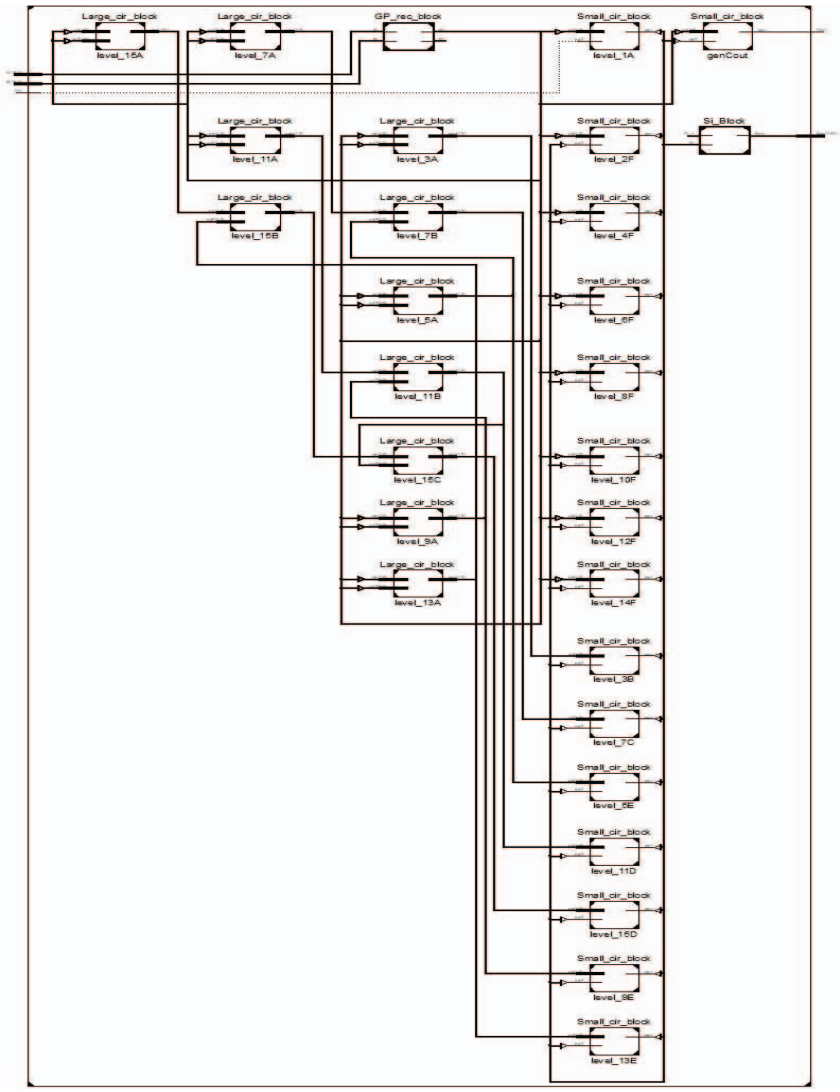


Fig.5. Proposed PPRLG Brent Kung adder's RTL Schematic

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:35:41 UTC from IEEE Xplore. Restrictions apply.

Various Block of Brent Kung adder mentioned in figure 5	Corresponding Parity Preserving reversible logic Gate	Quantum cost	Garbage Output	Constant Inputs	Gate count
Large GP rectangular block	1 MIG gate	7	2	2	1
Large circular GP block	2 NFT & 1 E2 gate	18	8	5	3
Small circular GP block	1 NFT & 1 E2 gate	2	6	4	2
final sum output	1 F2G gate	13	2	1	1

Fig.6. Various Parameters and Reversible gates used in sub – parts of BKA

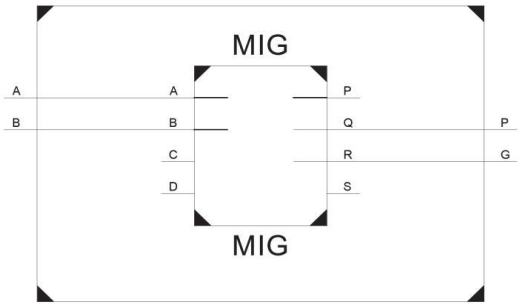


Fig.7. GP Rectangular block's RTL Schematic

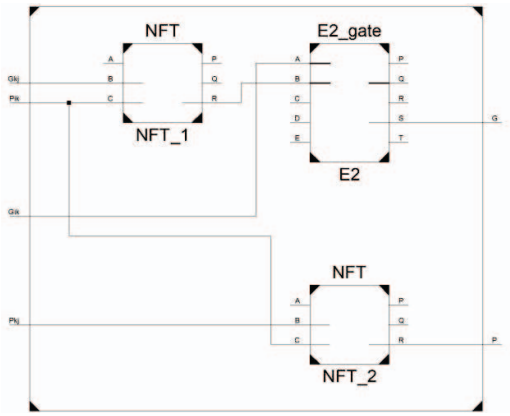


Fig.8. Large circular block's RTL Schematic

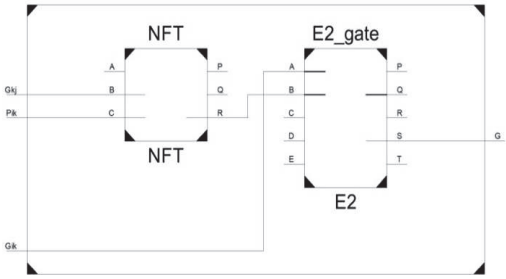


Fig.9. Small circular block's RTL Schematic

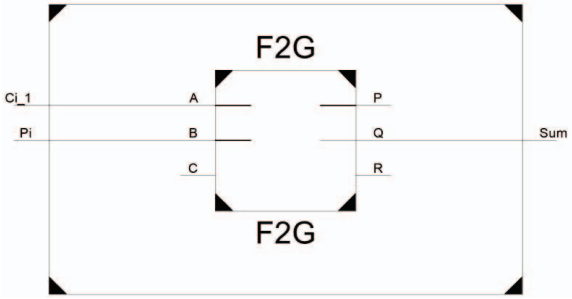


Fig.10. Triangular Si block's RTL Schematic

IV. Simulation & Results

Simulation and Synthetization of Proposed 16-bit Brent Kung adder are done using Xilinx ISE design suite version 14.7 by programming language Verilog. Figures 12 illustrate the results of simulations to confirm the functionality of the suggested adder. Comparison of our Proposed design with existing design of Brent Kung adder in paper [12] is mentioned in Figure 11.

Features Compared	[12]	Proposed Design
Reversible or Irreversible	Irreversible	Reversible
Fault Tolerant Capability	Not fault tolerable	Fault Tolerable
Power Dissipation	More	Less

Fig.11. Comparison of proposed design with existing design in paper [12]



Fig.12. Proposed Brent Kung adder simulation waveforms

V. CONCLUSION & FUTURE SCOPE

In this research, Fault-tolerant Brent Kung Adder is presented which is novel one. The Proposed Adder design described has the capability for fault tolerance, faster in speed, area is less, and wiring load is also less. This adder is faster, less computation nodes, area is less and wiring load is also less. All qualities of the proposed adder design can all be beneficial to design any system which require Performance, size, power, and fault-tolerant capabilities.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:35:41 UTC from IEEE Xplore. Restrictions apply.

References

[1] Wijeratno SB, Siddaiah N, Mathew SK, Anders MA, Krishnamurthy RK, Anderson J, Ernest M, Nardin M. A 9-GHz 65-nm Intel Pentium 4 processor integer execuion unit. IEEE Journal of Solid-State Circuits 2007; 42(1):26–37.

[2] Zhou C, Fleischer BM, Gschwind M, and Puri R. 64-bit prefix adders: power-efficient topologies and design solutions. IEEE Custom Integrated Circuits Conf. Proceedings, 2009; 179–182.

[3] Silberman J, Aoki N, Boerstler D, Burns JL, Dhong S, Essbaum A, Ghoshal U, Heidel D, Hofstee P, Lee KT, Meltzer D, Ngo H, Nowka K, Posluszny S, Takahashi O, Vo I, Zoric B. A 1.0-GHz single-issue 64-bit PowerPC integer processor. IEEE Journal of Solid-State Circuits 1998; 33(11):1600–1608.

[4] C. L. Ayala, N. Takeuchi, Y. Yamanashi, T. Ortlepp and N. Yoshikawa, "Majority-Logic-Optimized Parallel Prefix Carry Look-Ahead Adder Families Using Adiabatic Quantum-Flux-Parametron Logic," in IEEE Transactions on Applied Superconductivity, vol. 27, no. 4, pp. 1-7, June 2017, Art no. 1300407

[5] P. Vasudevan, P. K. Lala, D. Jia, and J. P. Parkerson, "Reversible logic design with online testability", IEEE Transactions on Instrumentation and Measurement, 55(2):406–414, 2006.

[6] B.RaghuKanth, B.Murali Krishna, M.Sridhar, V.G.SanthiSwaroop," A Distinguish Between Reversible and Conventional Logic Gates", International Journal of Engineering Research and Applications (IJERA), ISSN:2248-9622, Vol. 2, Issue 2, Mar-Apr 2012, pp.148-151.

[7] Landauer, R., "Irreversibility and heat generation in the computing process", IBM J. Research and Development, 5(3): pp. 183-191, 1961

[8] C.H. Bennett, Logical reversibility of computation. IBM J. Res. Dev. 17, 525–532 (1973)

[9] Bhavani Koyada,1 N. Meghana,2 Md. Omair Jaleel3 and Praneet Raj Jeripotula "A comparative study on Adders" MGIT, JNTUH, Hyderabad, 2017.

[10] M Morrison, M Lewandowski, R Meana et al., "Design of a novel reversible ALU using an enhanced carry look-ahead adder[C]", Nanotechnology (IEEE-NANO) 2011 11th IEEE Conference on, pp. 1436-1440, 2011

[11] G. V. Nikhil, B. P. Vaibhav, Vishnu G. Naik and B. S. Premananda, "Design of Low Power Barrel Shifter and Vedic Multiplier with Kogge-Stone Adder Using Reversible Logic Gates"", the Proceedings of IEEE International Conference on Communication and Signal Processing, pp. 1690-1694, 2017.

[12] Y. d. Ykuntam, K. Pavani and K. Saladi, "Design and analysis of High speed wallace tree multiplier using parallel prefix adders for VLSI circuit designs," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-6

[13] Pahuja Sumit and Garima Singh, and Prabhjot Singh "Real-Time Crop Monitoring in Agriculture." Artificial Intelligence and IoT-Based Technologies for Sustainable Farming and Smart Agriculture, edited by Pradeep Tomar and Gurjit Kaur, IGI Global, 2021, pp. 230-242.

[14] A Yadav, R., Kaur, G. Design and performance analysis of 1D, 2D and 3D prime sequance code family for optical network. J Opt 45, 343-356 (2016)

[15] S. Parvin and M. Altun, "Perfect Concurrent Fault Detection in CMOS Logic Circuits Using Parity Preservative Reversible Gates," in IEEE Access, vol. 7, pp. 163939-163947, 2019

[16] P. Vasudevan, P. K. Lala, D. Jia, and J. P. Parkerson, "Reversible logic design with online testability", IEEE Transactions on Instrumentation and Measurement, 55(2):406–414, 2006.

[17] B.Raghu Kanth, B.Murali Krishna, M.Sridhar, V.G.Santhi Swaroop," A Distinguish Between Reversible and Conventional Logic Gates", International Journal of Engineering Research and Applications (IJERA), ISSN:2248-9622, Vol. 2, Issue 2, Mar-Apr 2012, pp.148-151.

[18] L. Mahor, A. Chauhan and P. Tiwari, "Parity Preserving Reversible Design Using FinFETs," 2019 1st International Conference on Signal Processing, VLSI and Communication Engineering (ICSPVCE), 2019, pp. 1-6

[19] David Harris, "A Taxanomy of Parallel Prefix Networks",0-7803-8104-1/03/$17.0,2003, IEEE

[20] A. Raju, R. Patnaik, R. K. Babu and P. Mahato, "Parallel prefix adders -A comparative study for fastest response," 2016 International Conference on Communication and Electronics Systems (ICCES), 2016, pp. 1-6.

[21] Tomar, P., Kaur, G., & Singh, P. (2018). A prototype of IoT-based real time smart street parking system for smart cities. In Internet of Things and Big Data Analytics Toward Next-Generation Intelligence (pp. 243-263). Springer, Cham.

[22] Parasher, Y., Singh, P., & Kaur, G. (2019). Green Smart Town Planning. Green and Smart Technologies for Smart Cities, 19-41

[23] Singh, P., Dixit, V., & Kaur, J. (2019). Green Healthcare for Smart Cities. Green and Smart Technologies for Smart Cities, 91-130

[24] Parasher, Y., Singh, P., & Kaur, G. (2019). Green Smart Security System. Green and Smart Technologies for Smart Cities, 165-184.

[25] Parasher, Y., Singh, P., & Kaur, G. (2019). Green Smart Town Planning. Green and Smart Technologies for Smart Cities, 19-41

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:35:41 UTC from IEEE Xplore. Restrictions apply.

An Automated Multi-Web Platform Voting Framework to Predict Misleading Information Proliferated during COVID-19 Outbreak using Ensemble Method

Deepika Varshney[1], Dinesh Kumar Vishwakarma[2]
Biometric Research Laboratory, Department of Information Technology,
Delhi Technological University, Delhi-110042, India
Email: deepikavarshney06@gmail.com[1], dinesh@dtu.ac.in[2]

Abstract

Spreading of misleading information on social web platforms has fuelled huge panic and confusion among the public regarding the Corona disease, the detection of which is of paramount importance. To address this issue, in this paper, we have developed an automated system that can collect and validate the fact from multi web-platform to decide the credibility of the content. To identify the credibility of the posted claim, probable instances/clues(titles) of news information are first gathered from various web platforms. Later, the crucial set of features is retrieved that further feeds into the ensemble-based machine learning model to classify the news as misleading or real. The four sets of features based on the content, linguistics/semantic cues, similarity, and sentiments gathered from web-platforms and voting are applied to validate the news. Finally, the combined voting decides the support given to a specific claim. In addition to the validation part, a unique source platform is designed for collecting data/facts from three web platforms (Twitter, Facebook, Google) based on certain queries/words. This unique platform can also help researchers build datasets and gather useful/efficient clues from various web platforms. It has been observed that our proposed intelligent strategy gives promising results and quite effective in predicting misleading information. The proposed work provides practical implications for the policy makers and health practitioners that could be useful in protecting the world from misleading information proliferation during this pandemic.

Keywords: COVID-19, Information Pollution, Fake News Detection.

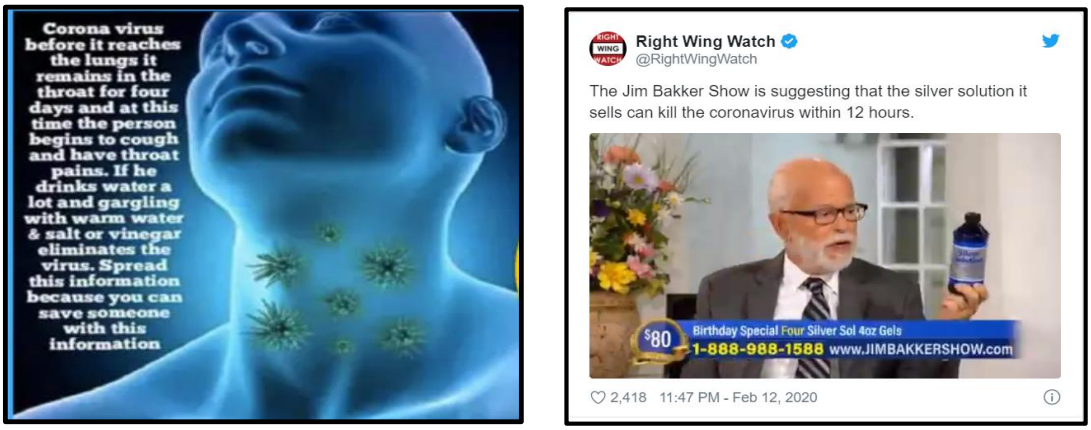
1. Introduction

In the recent scenario, a new coronavirus disease spread around the world. The disease emerges as a respiratory infection with significant concern for global public health hazards.

Initially, it is suspected that the disease is transmitted from animal to humans, later the paradigm is shifted that the infection is transmitted towards human to human via droplets, close contacts creating huge panic with approximate 6,359,182 confirmed cases and 380,663 deaths[1] have been encountered till now and growth rate is still high which has alarmed the global authorities including world health organization (WHO) [1]. The COVID-19 pandemic affects worldwide badly, however, there is no shortage of people who are taking this crisis as an opportunity for malicious activities/gaining profit[2]. A lot of health-related misleading information, some of the fake cures are suggested for COVID-19 has been posted by the malicious users and creates lots of confusion and misconceptions about the disease. During this pandemic, people have their eye on any new announcement from the government official or some news that can help to get rid of COVID-19. As the disease is deadly, the people are also desperate to know some cure and in rush to find a treatment for the new coronavirus disease. Some of the fake cures posted over social media are harmful too and give bad health advice. The recent examples of fake cures are shown in Fig.1, where Fig.1(a) shows an image with a false claim is gone viral that drinking water a lot and gargling with warm water and salt or vinegar eliminates the coronavirus, however, there is no significant evidence has been found concerning to this claim. Another cure in Fig1(b) reported that the silver solution can kill coronavirus within 12 hours. The proliferation of these misleading information creates lots of misconceptions in the mind-set of people related to coronavirus disease and some of the users are spreading it without verification and fuelled panic among people regarding the COVID-19. According to [3], misleading or fake can be defined as any post that shares content that does not faithfully represent the event that it refers to. We followed this definition in our work and define *"*misleading information as the content that does not faithfully represent the event that it refers to and having no significant evidence of proof to validate the claim*"*. From the recent research, it has been observed that numerous misleading content is circulating about the coronavirus and it is becoming difficult to differentiate fake news from the real one [4]. The propagation of misleading content on the virus could also be deleterious to mankind. This has led to the dire need for a system that can differentiate fake from real. Earlier many of the previous research has been reported methods of detecting fake news in online social media considering a variety of applications[5]. Most of the previous research has counter fake news problems mainly in the following type: Image-based algorithm and Text-based algorithms[6][7]. Many previous studies have worked on fake news detection by applying a

[1] https://in.search.yahoo.com/search?fr=mcafee&type=E211IN885G0&p=coronavirus

text-based approach. Text-based approaches mainly use text patterns and match them with already existing patterns of fake news. They are sometimes referred to as the linguistic approach. Along with this lots of researchers have shifted their interest in the credibility detection of posts/tweets using text-based features[8][9]. Like a text-based approach, research has also been done by employing an image-based approach. From the study, it has been seen that researchers have explored images based algorithms for the analysis of fake images or images attached with false claims in mainly following ways, Text additive images, and Manipulated images. The manipulated images termed as an image whose piece/part or certain region is manipulated with respect to visual context. Various image-based features have been explored for the classification of images. The authors of [10], propose 5 visual features and 7 statistical features for the verification of news events. Along with the manipulated images, some of the researchers have also considered text-additive images for the analysis of misleading content. The text additive images termed as images embedded with false claims instead of having any manipulation from visual context. The authors of [11], have incorporated text additive images, where they have applied a rule-based algorithm for the prediction of fake news. From recent research, it has been observed that none of the works have shown and reported fake news prediction analysis propagated during one of the major pandemic "CORONAVIRUS". Many people are sharing fake cures to get rid of coronavirus disease without any verification and create lots of misconceptions. Government and officials have also urged peoples to check the authenticity of the post before sharing [4]. This also motivates us to build an intelligent system for the prediction of fake news spreading during this pandemic. We, therefore, developed a generalized multi web platform framework of detecting misleading content on social media platforms, where we have considered COVID-19 as a special issue which is a huge pandemic and taken as one of the application case studies in this work. However, our model is generalized and works for other applications as well. COVID-19 is an emerging issue and very few research have been reported yet in this context that leads to motivates us to build an efficient framework to predict misleading content spreading during the COVID outbreak. The major key contributions of the work are highlighted in the following points.



(a) (b)

Fig 1. Examples of Misinformation[1]

- The proposed work gives a significant contribution in providing a novel generalized Automated Multi-Web Platform Voting Framework for collecting and validating misleading content in an online social network where considering COVID-19(fake news spreading during Corona outbreak) one of the special case studies from the application perspective.
- To the best of our knowledge, we are first to build a unique platform (Facts collector) for the collection of crucial facts and knowledge from three different prominently used social media and web search platform (Twitter, YouTube and Google) as well as provide different mechanism to search the query (build query) to get efficient and relevant results.
- The four set of novel features based on content, linguistics/semantic cues, similarity and sentiments has been gathered from web-platforms that further fed into ensemble based machine learning model to classify the news as Misleading or real. Finally, voting is applied to validate the news and to check the confidence/support given by different web platforms.
- As the COVID-19 is one of the emerging issues and none of the work has been reported yet to predict the fake news propagating during this phase and gives a major contribution by providing the analysis, which greatly helps researchers for further study.
- We investigate the model performances with different classifiers, and comparative analysis reveals that our proposed method outperforms other states of the art on the same dataset.

The remainder of this paper is organized as follows. In Section 2, we are going to discuss the previous work that has been done related to this field, wherein Section 3, we discuss the

problem statement and unique fact collector platform. In contrast, Section 4, elaborates the strategy/method that we have employed for the misleading information detection, which is followed by a discussion of experimental results in Section 5. Lastly, the paper is concluded with some suggested future work aspects.

2. Related Work

In the current era, spreading false information is one of the crucial problems nowadays, where it is quite difficult for the online user to discriminate fake news from the real one and that's why the development of an intelligent system is required. Most of the methods proposed in earlier states-of-the-art[12] to detect misleading information considered it a classification problem intending to associate label as true or false with a particular claim/post. From the survey analysis, it has been observed that the classification approaches are turn divided into approaches based on machine learning and deep learning. A detailed description is given below.

2.1 Machine Learning

From the study, it has been proven that machine learning algorithms are extremely useful in countering numerous problems in the information engineering field. In particular, many of the machine learning approaches implemented for misleading/ fake information detection applied as a supervised learning strategy[13]. In machine learning classification algorithms support vector machines(SVMs) are one of the widely used methods for classifications. The authors of [14], have proposed a method where they employed a graph-kernel-based SVM classifier to detect rumors using propagation structure and content features with an accuracy of 0.91 on Sina-Weibo dataset. Whereas, in [15] the author reported a set of features to distinguish among fake news, real news, and satire. The SVMs are also employed for clickbait detection in [16]. Like SVM, the random forest has also been exploited in numerous works for fake news and rumor detection. Most of the studies have reported random forests as a strong performer among other machine learning algorithms [17], [18], [19], [20], [21]. In [18], the author has proposed a set of temporal, structural, and linguistic features for the classification of rumors in a tweet graph by employing a random forest with an accuracy of 0.90. The random forest has also been used for stance detection in [17],[22]. The comparative studies of a different approach in the context of rumor and fake news have shown competitive performance for logistic regression [19],[23],[24],[25]]. The authors of [26], employed logistic regression for stance classification of news articles or headlines and claims. Another widely studied family of the algorithm

proposed particularly for misleading content detection is a decision tree [27]. The effectiveness of decision tree algorithms like j48 with respect to other machine learning paradigms including SVMs has been reported in [[8],[19]]. The authors of [8] have used the content and context-based features to perform credibility evaluation of tweets and the model is performing well with an accuracy of 0.86. In [28], to evaluate the trustworthiness of users in social media via decision tree, the author has proposed a series of user trust metrics and reported an accuracy of 0.75.

2.2 Deep Learning

Deep learning is one of the prominent and widely explored research topics in machine learning. The main advantage of deep learning over traditional machine learning approaches is they are not based on manually crafted features and lead to reduce feature extraction time. Along with this, the deep learning framework can learn hidden representations from simpler inputs both in context and content variation [29]. The two prominent and widely used paradigms in Morden artificial neural network are RNN and CNN. In [29], authors have proposed a novel RNN architectures, namely tanh-RNN, LSTM, and Gated Recurrent Unit(GRU) for the detection of rumors. From the results, it has been found that GRU has obtained the best results in both the datasets considered with 0.88 and 0.91 accuracies, respectively. Whereas, in [30], the author has proposed a multi-task learning approach and designed a multi-task learning framework with an LSTM layer shared among all tasks to counter the problem of rumor classification. Like RNN, CNNs have also been explored and widely studied for image recognition and many other fields of computer vision. However, it now gaining popularity in the NLP field as well [31]. The authors of [22], have explored a technique using CNN with single and multi-word embedding to counter problems concerning both stance and veracity classification of tweets. The author has reported accuracy of 0.70 for stance classification problem and 0.53 for the problem of veracity classification. Whereas, Paragraph embedding is explored to learn the representation of a small group of posts in a specific event and used them as input for their CNN model in [32] and achieved an accuracy of 0.93 for Sina Weibo and 0.77 for Twitter. From the study, it has been observed that most of the recent work has explored the combination of RNN and CNN in their model [[33],[34],[35]]. The authors of [35] proposed an architecture applied on the LIAR dataset that encodes text information via a CNN and metadata about the author of the text using an LSTM layer as well as it has been also found that the hybrid model has proved to outperform all other baselines along with a bi-LSTM architecture with an accuracy of 0.27 on the testing dataset. Whereas,

in [[33]] author has proposed an approach based on repost sequence patterns for the detection of false rumours.

All the different approaches discussed above have considered different machine and deep learning methods for the prediction of misleading content. In our study, we have used a multi web platform framework to gather effective clues for predicting false information. The clues can be detected from multiple web platforms to get the strong support as it may happen that one platform may not give effective clues to predict some information but other can. Moving to this concept, instead of relying only on a specific platform for getting information, our proposed model incorporates multi web platform for retrieving clues concerning to specific query. To the best of our knowledge, none of the previous studies has been explored this concept. Along with this, very few studies incorporate the concept of a unique platform that can collect information from various social media and web sources for building data. These all points we will discuss in later sections in detail.

3. Problem Definition and Unique Fact Collector Platform

In this paper, we have considered a binary class classification problem. We assume that the posted source claim $c = \{c1, c2, c3 \dots ck\}$ can be divided into two classes $Class = \{M, R\}$: 1) Real (R), namely the posted claim faithfully represents the event that it refers to, 2) Misleading (M), namely the posted claim does not faithfully represent the event that it refers to. The claims that have been included here are related to coronavirus COVID-19. Take an example of the post related to coronavirus. The post that "chlorine and alcohol products cannot kill viruses within the body" is true information and belongs to the real class since much authoritative and authentic news media have reported relevant news and has also acknowledged by WHO[2]. While one of the posts says that "coronavirus is caused by 5G technology" is absolutely a false rumour, it lacks factual support and deviates from the scientific principles. So, the goal is to learn a classifier from the labelled feature set, that is $f : X^k \mapsto Y^k$, where Y^k takes one of the two fine-grained classes: {R, M}. Given the input feature set X^k, the classifier f can output the classification result for the posted claim C^k. In the coronavirus outbreak, lots of misleading information was posted by users, and most of the misleading information reading fake cures, lockdown, and others have gone viral during the period. The posts have gone viral on different social media platforms like Twitter, Facebook, and others. For analysing post from

[2] https://www.indiatoday.in/world/story/drinking-alcohol-will-not-protect-you-from-covid-19-says-who-1653555-2020-03-08

different social media platform or building dataset, we build a Unique Fact Collector Platform that can take user input query and search on three prominently used web platforms (Google, Twitter and YouTube) simultaneously to collect relevant knowledge concerning to query. From the analysis, it has been found that twitter is not taking long queries as an input that's why it can only collect information by putting keyword based search. Whereas, Google web search and YouTube can take long queries as an input. Our Unique Fact Collector Platform allow users to fetch information from all three platforms concerning to a query. The Fig.2. shows the flow diagram of our collector framework.

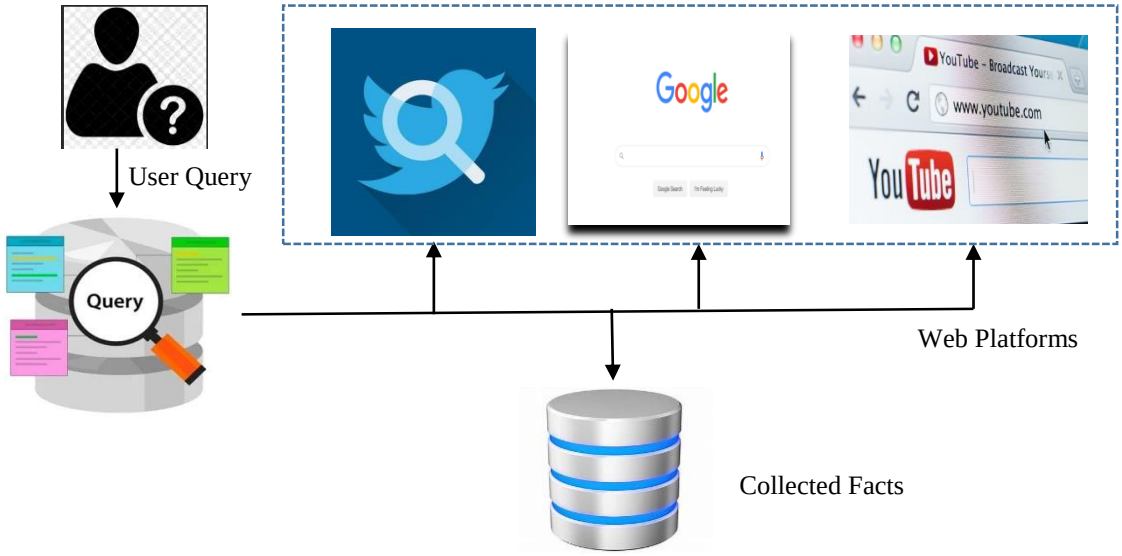


Fig.2. Flow Diagram of Unique Fact Collector Platform

4. Proposed Methodology

In this section, we elaborate our automated Multi-Web Platform Voting Framework to Predict Misleading Information Proliferated during COVID-19 outbreak. The detailed flow-diagram of our proposed model is shown in Fig.3. In the first phase of the process the input query is given by the user that he/she wants to validate. The input query is passed through the text processing phase where the cleaning of data has been performed to make it in a format so that it can be used for further processing, and it includes removal of stopwords, removing duplicates, handling missing data, stemming, punctuation removal, text translation (Google translation API) to English language, Removing URLs, symbol, emoji etc. After, text processing the cleaned data is pass to the next module called as "Fact Collection". In this phase, the input is passes through Multi-Web platform to reterive relevant facts concerning to query. The two prominent social media platform that has been utilized to retrieve the facts are

YouTube platform and google web search. In order to gather efficient and relevant titles the query building is one of the important aspects. What query should be pass to get more relevant responses. We have defined three novel way to build a query, however from all the given build

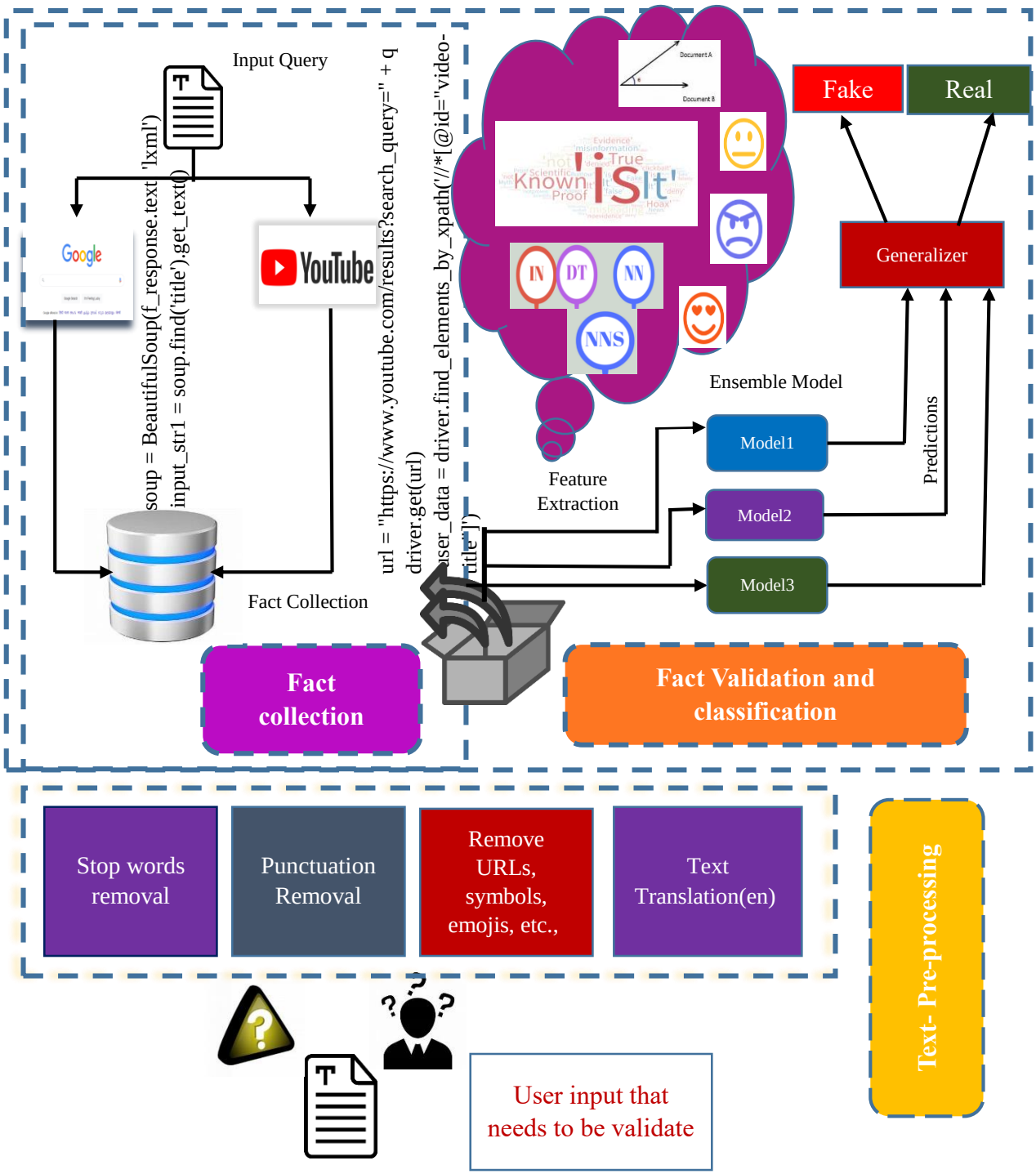


Fig.3. The flow diagram of the proposed approach

cases we have adopted the one which is most effective in reteriving relevant information where we have adopted case 1 as other have some limilation as disccused in Table 1. The Table 1 describes 3 possible cases we have considered for building query.

Table 1 Possible build cases

Case No.	Possible Cases	Limitation
1.	Stopword_removal(Input_query)+ " "+ "fake news"	-
2.	N_grams(Input_query)+ " " + fake news	1. Sometimes the context of the query cannot be come out properly and missed out.
3.	Pos(Input_query)+" "+ Fake news	1. Sometimes the context of the query cannot be come out proprely. 2. In some of the cases giving too many irrelevant facts, goes out of context.

4.1 Fact Collection

In the second phase, the query has been created by removing stopwords to make the query clear and short. As in many cases, the stopwords makes query too long that leads to give irrelevant responses. After removing stopwords, the query is attached with a space concatenated with "*fake news*" keyword. From the analysis, this case of query generation found to be good and considered in this study. Whereas, the other cases includes the N_grams concatenated with the keyword "fake news" and the POS part of speech tagging(proper nouns), we can find all the proper nouns from the input query. Each proper noun is concatenated with the fake news keyword to build a query. However, these case have certain limitations. Sometimes the context of the query cannot become out properly and missed out the relevance.

After this step, the build query is passed through two prominently used platforms, YouTube and Google, to retrieve relevant facts. Finally, the top 10 title headings are scrapped automatically using selenium from both the platform that further be used for analysis. The algorithm for fact collection is shown below in Algorithm 1.

Algorithm 1, shows the process of facts collected from the Multi-web platform. Here, in this study, we have incorporated two social media and web search platforms for reteriving efficient facts/title heading concerning a query that further be used in feature engineering and validate the claim as fake/real. However, other platforms like twitter has also been explored for the collection of facts, but the issue with twitter is it support keyword based searching and long query-based search is not applicable that leads to be a major issue in the collection of relevant facts. Whereas, in the case of google web search and YouTube, we can fetch efficient responses concerning a query.

Algorithm 1.(Facts Collection)

Input(Text/Text additive image) and Output(Collection of Facts)

```
def. main ():
  if(image)
       claim= OCR(image)
  else:
  claim= raw input(claim)
  claim= text_preprocessing(claim)
  build query= Stopword_removal(claim) + " "+ “fake news”
  facts google= google_fact_collect (build query)
  facts_youtube= youtube_fact_collect (build query)
 def. google_fact_collect(query):
     for j in search (query, tld="co.in", num=10, stop=10, pause=2):
           f_response= s.get (j, headers = {'User-agent': 'your bot 0.1'})
            soup = BeautifulSoup (f_response. text, 'lxml')
            input_str1 = soup. Find('title'). get_text ()
            input_str1= input_str1.lower()
            input_str1= clean(input_str1)
             return(input_str1)
  def. YouTube_fact_collector(query):
       q= query +" "+'fake news'
       print(q)
       url = "https://www.youtube.com/results?search_query=" + q
       print(url)
       count=0
       driver.get(url)
       user data = driver. find_elements_by_xpath('//*[@id="video-title"]')
       print(user_data)
       links = []
       titles= []
       for i in user data:
         links. append(i.get_attribute('href'))
         titles. append(i.get_attribute('title'))
        for x in titles:
           v_title = x
           return(v_title)
```

4.2 Fact Validation

The third module is fact validation, this module take the facts collected from the previous step and utilizes them to get some efficient clues to predict the claim as fake/real. The four set of features are employed based on content, linguistics/semantic cues, similarity and sentiments. Each of these feature category has been discussed in detail in the below section

Table 2: Detailed description of proposed features

Features Category	Features	Feature Description
Content-Based	1. Question mark count 2. Fake word count	1. Number of question mark in a title heading. 2. Number of fake words encountered in a title heading.
Linguistics/Semantic cues-Based	1. NLTK POS TAGGING Semantic Similarity	1. The nltk wordnet's synset is used to measure the semantic similarity between user query and title headings.
Similarity-Based	1. Cosine Similarity $\boldsymbol{COS(x,y)} = \frac{x.y}{\lvert x\rvert.\lvert y\rvert}$	1. The cosine similarity is used to measure the similarity between user query and title headings.
Sentiment-Based	1. Query sentiments 2. Clue sentiments 3. Sentiment match count	1. This features return the sentiment of the user query, either positive negative or neutral. 2. This features return the sentiment of the title heading, either positive negative or neutral. 3. This features return the count of how many times the query sentiments matchs with the title heading sentiments.

4.2.1 Content based features:

The content based features has been widely explored in numerous data mining research field. In this paper we have incorporated content based features for the prediction of misleading information including question mark count and fake word count. The question mark count gives necessary clues regarding the confidence reflected from the sentence. If the sentence showing uncertainty, it means that the claim is not sure regarding that event. Question mark count plays a major role in finding the uncertainty in a given sentence. It returns true if any question mark has been encountered in a title/headings retrieved from the web platform, while searching a specific query. Whereas, fake_word_count is also one of the important feature that discriminate fake from real. There are set of false_phrase_corpus that incorporates list of keywords that prominently used to represent fake news. The keyword corpus that we have created including following phrases: { *'false','misleading','inaccurate','rumor','rumour', 'not correct' ,'fake news', 'incorrect', 'wrong', 'confounding', 'deceiving', 'deluding', 'wont', 'did' 'Did', 'funny', 'memes', 'catchy', 'bogus', 'counterfeit', 'fabricated', 'fictitious', 'forged',*

'fraudulent', 'mock', 'phony', 'affected', 'artificial', 'erroneous', 'fake', 'fanciful', 'faulty', 'improper', 'invalid', 'mistaken', 'unfounded', 'unreal', 'untrue', 'untruthful', 'casuistic', 'fishy', 'illusive', 'imaginary', 'inexact', 'lying', 'misrepresentative', 'falsity', 'misreport', 'misstatement', 'deception', 'falsification', 'artificial', 'fabrication', 'falsehood', 'hoax', '?', 'Not Died', 'misinformation', 'not committed', 'not dead', 'death rumour', 'is it true', 'not known', 'no proof', 'no known', 'no scientific evidence', 'no evidence', 'not verified', 'clickbait', 'not proven', 'denied', 'deny', 'unverified', 'falsely', 'myth', 'ridiculous', 'not true'}, if any of these word has been encountered in the retrieved responses corresponding to a query, the fake count incremented by 1. The feature is helpful in identifying fake as the title having these phrases more likely representing news as fake.

4.2.2 Linguistics/ semantic cues- based Features

It is very diffcult to process raw text intelligently as the same words in a different order can mean something completely different, while using lingustic knowledge can be possible to solve some problems starting from only the raw characters. For a given claim it is very important to understand in what context it is used. The python libray nltk.pos_tag is designed to do the same. When a raw text is passed as an input, it returns a doc object that comes with a variety of annotation. The nltk parse and tag a given document, there are some statistical model which enable it to make prediction of which tag or label most likely applies in this context also called as POS part of speech tagging/ grammatical tagging which is used to mark up a word in a text as corresponding to a particular part of speech, based on both its definition and its context. POS tagging also describes the charactersitics of lexical terms within a sentence or text that further be used for making prediction / assumptions about semantics. To compute the semantic text similarity between two sentence we have used POS (Part of speech) text similarity. There are different POS tags that can be given to a each word in a sentence like $(NNS, noun\ plural)(NNP, proper\ noun, singular)(NN, noun, singular) etc$ .NLTK POS tagger is employed to assign grammatical information of each word of the sentence. This feature is useful in computing the semantic text similarity between the user query and the clues reterived from web platforms. The tags generated by nltk.pos_tag are converted to the tag used by wordnet.synsets. The nltk wordnet's synset is used to measure the similarity.

4.2.3 Similarity based Feature

This is another category of feature used in this work based on similarity. This feature is helpful in segregating relevant title/heading from all the given responses, as not all responses

are useful for validation. To get efficient performance of the model we need to remove irrelevant titles from the analysis, only those who cross the threshold value are used for analysis. The one of the prominently used similarity measure "cosine similarity" has been used in this work to compute the similarity between two sentences irrespective of their size. The sentences are considered as two vectors and the cosine similarity between two vectors is measured in 'θ'. If the angle between two sentence is 0 it means they are similar, and if θ = 90° they are dissimilar. The formula of calculating similarity between two sentence x and y can be given as:

$$\boldsymbol{COS}(\boldsymbol{x},\boldsymbol{y}) = \frac{x.y}{|x|.|y|} \quad (1)$$

4.2.4 Sentiment based:

Sentiment based features are the fourth set of features employed for the prediction of fake news. Sentiments plays an important role in identifying polarity of the sentence, whether it is showing positive negative or neutral sentiments, Here, we have considered 3 features under this category.

1) Query Sentiment: Query Sentiment is a sentiment of the input query given by the user.
2) Title/heading sentiment: This is a sentiment of the responses(title/heading) received as a search result concerning a specific query.
3) Sentiment match count: From all the 10 responses retrieved from the web platforms, how many times the sentiments of the query and the titles are matches. It also represents whether the sentiment pose by the input query is equivalent to the responses received. It also means that both query and heading are posing the same sentiments and presented in the same polarity.

All these above discussed features are briefly shown in Table 2 and the Algorithm 2 elaborate the complete process of fact validation, where the functions to evaluate the four set of features are briefly explained that later be fed to ensemble based classifier for analysing the performance of the model.

Algorithm 2.(Facts Validation)

Input(facts) and Output(status(fake/real))

```
def. main ():
  Linguistic features=  Feature_extraction_lingustic(facts)
  Content_feature  =  Feature_extraction_content(facts)
  Sentiments_features  = Feature_extraction_sentiments(facts)
  Similarity_features =  Feature_extraction_similarity(facts)
  Classification_model  =  Ensemble_classifier  (Content_features,  Sentiments_features,  Linguistic_features,
Similarity_features)
   return(status)
def. Feature_extraction_lingustic(facts):
```

```
            r. extract_keywords_from_text(input_str1)
             ti=r.get_ranked_phrases () # To get keyword phrases ranked highest to lowest.
             print(ti)
             Y1= listToString(ti)
             s1 = nltk.pos_tag(nltk.word_tokenize(input_str))
             s2 = nltk.pos_tag(nltk.word_tokenize(input_str1))
             Semantic_similarity= similarity (s1, s2)
             return(Semantic_similarity)
def. Feature_extraction_content(facts):
        r. extract_keywords_from_text(input_str1)
        ti=r.get_ranked_phrases () # To get keyword phrases ranked highest to lowest.
        print(ti)
        Y1= listToString(ti)
        if any (word in Y1 for word in punctuation):
                pun=pun+1
                print ("pun count", pun)
       if any (word in Y1 for word in keyword):
                fake_count=fake_count+1
                print ("The fake count", fake_count)
  def. Feature_extraction_similarity(facts):
          list1 = word_tokenize(input_str)
          list2 = word_tokenize(input_str1)
          X_set = {w for w in list1 if not w in sw}
          Y_set = {w for w in list2 if not w in sw}
          # form a set containing keywords of both strings
          rvector = X_set.union(Y_set)
          for w in rvector:
           if w in X_set: l1.append(1) # create a vector
           else: l1.append(0)
           if w in Y_set: l2.append(1)
           else: l2.append(0)
          c = 0
       # cosine formula
          for i in range(len(rvector)):
           c+= l1[i]*l2[i]
          cosine = c / float((sum(l1)*sum(l2))**0.5)
          print("The similarity: ", cosine)

    def. Feature_extraction_sentiments(facts):
            query_sentiment= get_tweet_sentiment(query)
            print("The sentiment of the query",query_sentiment)
            title_sentiment= get_tweet_sentiment(input_str1)
            print ("The sentiment of the title", title_sentiment)
            if(query_sentiment==title_sentiment):
             senti_count=senti_count+1
             print ("The senti_count", senti_count)
```

5. Experimental Analysis and Results

In this section, the experimental analysis is performed on publicly available datasets, different performance measures are adopted (Precision, Recall, F1-score, Accuracy etc.) to measure the effectiveness of the proposed method and lastly presenting the results showing the performance of the proposed model as well as comparative analysis with other State-of-the-art methods. This section covers each of these points in the following subsections.

5.1 Constraint-2021 COVID-19 Fake News Detection Dataset

In this paper, we have used the constraint-2021 shared task to detect COVID-19 fake news in English. The dataset is collected from various social media like Twitter, Facebook, Instagram, etc. The main objective of this task is to classify a given social media post into Fake/Real. The dataset collects 10,700 manually annotated social media posts and articles of fake and real news on COVID-19 [36]. The dataset is further split into training validation and test sets in the ratio of 3:1:1 as shown in Table 3.

Table 3. The Constraint -2021 task dataset description

Split category	Real	Fake	Total
Training set	3360	3060	6420
Validation set	1120	1020	2140
Test set	1120	1020	2140
Total	5600	5100	10700

The experimental analysis is performed by employing various machine learning algorithms like Logistic Regression(LR), Support Vector Machine(SVM), Random forest, Ensemble based classification model, etc. We employed this dataset to measure the performance of our model with respect to precision, recall, f1-score and accuracy. As we have incorporated multi-web platforms, so analysis has been performed on both the web platform separately as well hybrid (Google +YouTube) of both. The Results concerning to this experiment is shown Table 4, Table 5 and Table 6 respectively. The first study incorporates the clues retrieved from google platform concerning to specific query. The performance has been analysed majorly on four measures Precision, Recall, F1-Score and Accuracy by employing Random Forest, Support Vector Machine, Logistic- Regression and Ensemble based classification algorithm, where it has been observed that ensemble based Voting classifier (RF, LR, KNN) and (LR, LSVM, CART) performs best and outperforms all other classifier with an F1 Score of 0.989. Whereas, the second best run is given by SVM with an F1-Score of 0.987. Another set of experiment has been performed on the clues retrieved from YouTube platform. It has been observed that LR perform best with an F1-Score of 0.869. However, it can be clearly seen that the clues retrieved from the YouTube alone are not well efficient as compare to Google web search platform. Therefore, the concept of Multi Web platform comes into picture, giving support to a claim in making decision as false or true, if some other platform cannot give sufficient clues for the prediction. The hybrid model improves the performance of the model by incorporating clues concerning to both the web platform to improve the efficiency of the model. The best F1-Score on hybrid model is achieved from SVM and ensemble based classifier (LR, LSVM, CART)

having value 0.980 as shown in Table 6. Some earlier studies have also worked on the given problem and reported results concerning the Constraint Task 2021 Covid fake news dataset. The Comparative Study with the other state-of-the-art method on the validation set is shown in Table 7. The authors of [37], proposed a method to predict misleading information proliferated during COVID-19 outbreak, by incorporating various machine learning classifiers and here represented as Model 1. The best run provided by Model 1 by using SVM with an F1-Score of 95.70, whereas our proposed approach on SVM giving the F1-Score of 98.70 and enhanced the performance by 3% as shown in Fig.4. Similarly, the authors of [36] (Model 2) and [38](Model 3) , worked on the same problem task using machine learning and ensemble based classification approach, where they reported best run F1-Score of 93.46 by employing SVM and 98.32 using Ensemble based model respectively as shown in Fig.5. and Fig.6. respectively. It can be clearly that that our model outperforms in all discussed cases and provided the best run using ensemble based model incorporating (LR, LSVM and CART) with an F1-Score of 98.88.

Table 4. Performance of the model incorporating Google web platform

Model	Precision	Recall	F1-Score	Accuracy
Random Forest	0.986	0.986	0.986	0.985
SVM	0.987	0.987	0.987	0.987
LR	0.986	0.986	0.986	0.986
Ensemble learners				
Random Forest	0.986	0.986	0.986	0.985
Voting classifier (RF, LR, KNN)	0.989	0.989	**0.989**	0.989
Voting Classifier (LR, LSVM, CART)	0.989	0.989	0.989	0.987
Bagging Classifier(Decision-Tree)	0.980	0.979	0.979	0.978

Table 5. Performance of the model incorporating YouTube web platform

Model	Precision	Recall	F1-Score	Accuracy
Random Forest	0.866	0.852	0.850	0.851
SVM	0.860	0.860	0.860	0.860
LR	0.870	0.869	0.869	0.869
Ensemble learners				
Random Forest	0.866	0.852	0.850	0.851
Voting classifier (RF, LR, KNN)	0.863	0.863	0.863	0.862
Voting Classifier (LR, LSVM, CART)	0.865	0.865	0.865	0.864
Bagging Classifier(Decision-Tree)	0.853	0.795	0.785	0.795

Table 6. Performance of the model incorporating both the Web Platform (Google + YouTube)

Model	Precision	Recall	F1-Score	Accuracy
Random Forest	0.974	0.973	0.973	0.973
SVM	0.980	0.980	0.980	0.979
LR	0.978	0.978	0.978	0.976
SGD	0.980	0.980	0.980	0.975

Ensemble learners				
Random Forest	0.974	0.973	0.973	0.973
Voting classifier (RF, LR, KNN)	0.979	0.979	0.979	0.979
Voting Classifier (LR, LSVM, CART)	0.980	0.980	**0.980**	0.980
Bagging Classifier(Decision-Tree)	0.972	0.971	0.970	0.975

Table 7 Comparative Study with the other state-of-the-art method on the validation set.

Author	Model	Precision	Recall	F1-Score	Accuracy
[37] (Model1)	SVM	95.71	95.70	95.70	95.70
[37] (Model 1)	LR	95.43	95.42	95.42	95.42
[37] (Model 1)	RF	90.98	90.79	90.80	90.79
[37] (Model 1)	NB	93.33	93.32	93.31	93.32
[37] (Model 1)	MLP	93.62	93.60	93.59	93.60
[36] (Model 2)	DT	85.31	85.23	85.25	85.23
[36] (Model 2)	LR	92.76	92.79	92.79	92.75
[36] (Model 2)	SVM	93.46	93.48	93.46	93.46
[38] (Model 3)	Ensemble based model	98.32	98.32	98.32	98.32
Our proposed model	Ensemble voting classifier(LR,CART,LSVM)	98.88	98.88	**98.88**	98.79
Our proposed model	Random Forest	98.20	98.10	98.10	98.09
Our Proposed model	LSVM	98.70	98.70	98.70	98.70
Our Proposed model	Logistic Regression	98.60	98.60	98.60	98.55
Our Proposed Approach model	NB	95.55	95.53	95.54	95.34

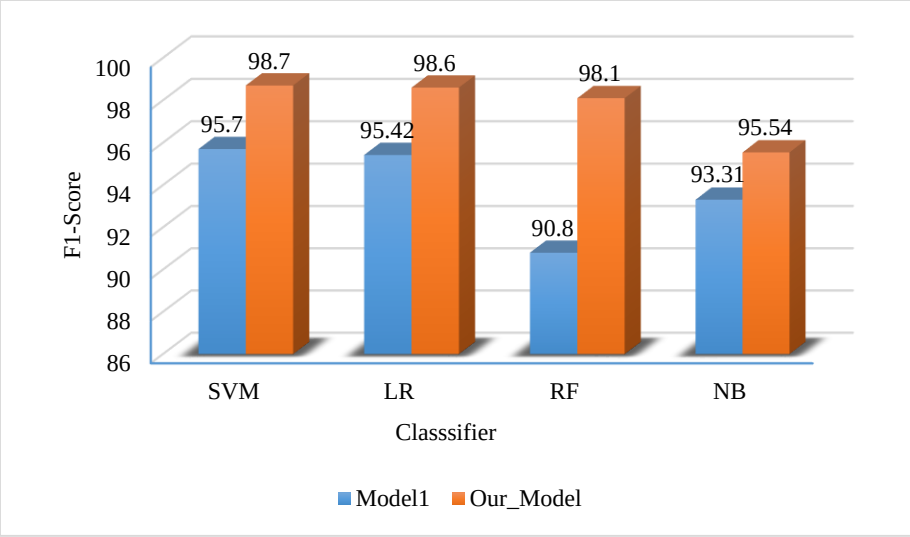


Fig.4. Comparative analysis with Model 1 on F1 score

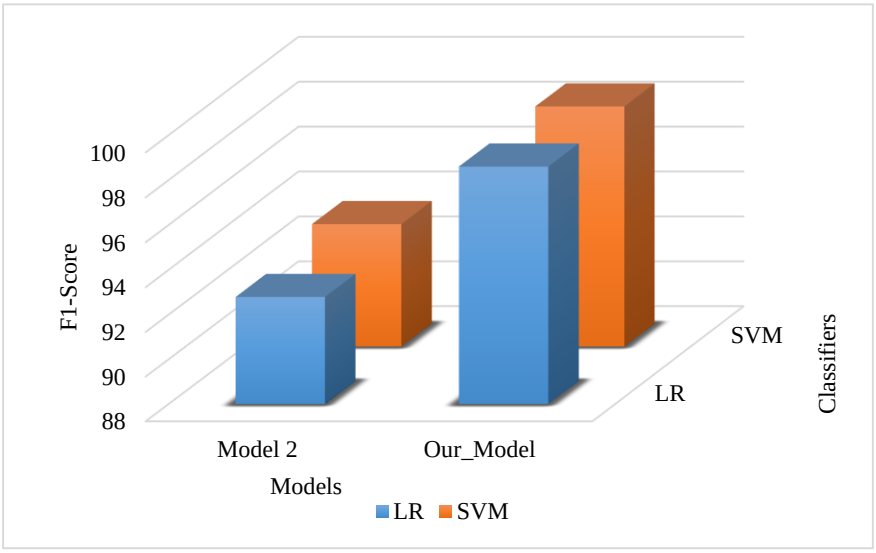


Fig.5. Comparative analysis with Model 2 on F1 score

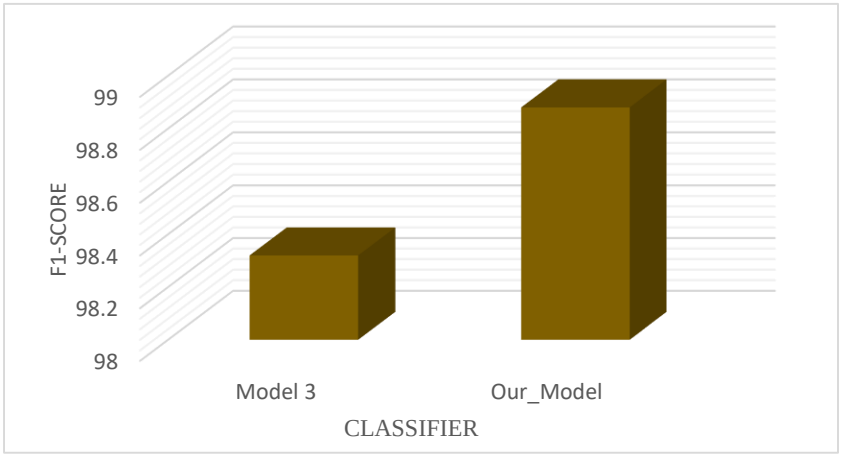


Fig.6. Comparative analysis with Model 3 on F1 score

6. Conclusion and Future Work

In this paper, we developed an intelligent generalized strategy for identifying possible clues to predict misleading information, where fake news proliferated during the COVID-19 outbreak is considered as a special case study and detailed analysis has been discussed. We proposed an automated Multi-Web Platform Voting Framework considering YouTube and Google as major sources for the retrieval of clues. The four set of novel features based on content, linguistics/semantic cues, similarity and sentiments has been gathered from these platforms that further fed into ensemble based machine learning model to classify the news as Misleading or real. Voting is applied to validate the news and to check the confidence/support given by different web platforms. It has been observed that Google web platform itself

performing good in retrieving crucial knowledge, giving best F1-Score of 98.89 by employing Ensemble based model incorporating LR, LSVM and CART and their voting gives the final decision. However, considering YouTube as a web platform alone for retrieving knowledge it only is able to give an F1-Score of 86.90 by employing LR which is quite low. Here, we can see YouTube alone is not able to retrieve effective clues to predict the news, however, incorporating multi web platform scheme we can improve the performance of the model by taking support from other platforms to validate the veracity of news when it is not available. Retrieving clues from multi-web platform improve the performance of the model and it outperforms other state-of the-art technique on the same dataset by employing ensemble based classification model. In the future we are planning to incorporates and explore other platforms (Instagram, WhatsApp etc.) to validate the news as well as also expand the work by including different modalities of data (images, videos, etc). Along with this, we are also planning to build a real time application for the users to predict misleading content.

Reference

[1] K. K. Sahu, A. K. Mishra, and A. Lal, "Comprehensive update on current outbreak of novel coronavirus infection (2019-nCoV)," *Ann. Transl. Med.*, vol. 8, no. 6, 2020.

[2] D. Varshney and D. K. Vishwakarma, "Analysing and Identifying Crucial Evidences for the prediction of False Information proliferated during COVID-19 Outbreak : A Case Study," 2021, pp. 47–51.

[3] C. Boididou, S. Papadopoulos, M. Zampoglou, L. Apostolidis, O. Papadopoulou, and Y. Kompatsiaris, "Detection and visualization of misleading content on Twitter," *Int. J. Multimed. Inf. Retr.*, vol. 7, no. 1, pp. 71–86, 2018.

[4] W. Zhou, A. Wang, F. Xia, Y. Xiao, and S. Tang, "Effects of media reporting on mitigating spread of COVID-19 in the early phase of the outbreak," 2020.

[5] D. Varshney and D. K. Vishwakarma, "A unified approach for detection of Clickbait videos on YouTube using cognitive evidences," *Appl. Intell.*, pp. 1–22.

[6] P. Meel and D. K. Vishwakarma, "Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities," *Expert Systems with Applications*, vol. 153. Elsevier Ltd, Sep. 01, 2020, doi: 10.1016/j.eswa.2019.112986.

[7] P. Meel and D. K. Vishwakarma, "HAN, image captioning, and forensics ensemble multimodal fake news detection," *Inf. Sci. (Ny).*, vol. 567, pp. 23–41, 2021, doi: 10.1016/j.ins.2021.03.037.

[8] C. Castillo, M. Mendoza, and B. Poblete, "Information Credibility on Twitter," in *20th international conference on World wide web. ACM, 2011.*, 2011, pp. 675–684.

[9] D. Varshney and D. K. Vishwakarma, "Hoax news-inspector: a real-time prediction of

fake news using content resemblance over web search results for authenticating the credibility of news articles," *J. Ambient Intell. Humaniz. Comput.*, pp. 1–14, 2020.

[10] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel Visual and Statistical Image Features for Microblogs News Verification," *IEEE Trans. Multimed.*, vol. 19, no. 3, pp. 598–608, 2017, doi: 10.1109/TMM.2016.2617078.

[11] D. K. Vishwakarma, D. Varshney, and A. Yadav, "Detection and veracity analysis of fake news via scrapping and authenticating the web search," *Cogn. Syst. Res.*, vol. 58, pp. 217–229, Dec. 2019, doi: 10.1016/j.cogsys.2019.07.004.

[12] D. Varshney and D. K. Vishwakarma, "A review on rumour prediction and veracity assessment in online social network," *Expert Syst. Appl.*, vol. 168, no. May 2020, p. 114208, 2020, doi: 10.1016/j.eswa.2020.114208.

[13] P. Meel and D. K. Vishwakarma, "A temporal ensembling based semi-supervised ConvNet for the detection of fake news articles," *Expert Syst. Appl.*, vol. 177, no. September 2020, p. 115002, 2021, doi: 10.1016/j.eswa.2021.115002.

[14] K. Wu, S. Yang, and K. Q. Zhu, "False Rumors Detection on Sina Weibo by Propagation Structures," in *IEEE 31st international conference on data engineering*, 2015, pp. 651–662, doi: 10.1109/ICDE.2015.7113322.

[15] B. D. Horne and S. Adali, "This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News," in *Eleventh International AAAI Conference on Web and Social Media*, 2017, pp. 759–766.

[16] A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly, "Stop clickbait: Detecting and preventing clickbaits in online news media," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016, pp. 9–16.

[17] A. Aker, L. Derczynski, and K. Bontcheva, "Simple open stance classification for rumour analysis," *arXiv Prepr. arXiv1708.05286*, 2017.

[18] E. J. Briscoe, D. S. Appling, and H. Hayes, "Cues to deception in social media communications," in *2014 47th Hawaii international conference on system sciences*, 2014, pp. 1435–1443.

[19] G. Giasemidis *et al.*, "Determining the Veracity of Rumours on Twitter," *CoRR*, vol. abs/1611.0, 2016.

[20] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *IEEE International Conference on Data Mining, ICDM*, 2013, pp. 1103–1108, doi: 10.1109/ICDM.2013.61.

[21] L. Zeng, K. Starbird, and E. S. Spiro, "# unconfirmed: Classifying rumor stance in crisis-related social media messages," 2016.

[22] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. W. S. Hoi, and A. Zubiaga, "SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours," *CoRR*, vol. abs/1704.0, 2017.

[23] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. De Alfaro, "Some Like it Hoax : Automated Fake News Detection in Social Networks," *arXiv Prepr. arXiv1704.07506*, pp. 1–12, 2017.

[24] L. Zhou, D. P. Twitchell, T. Qin, J. K. Burgoon, and J. F. Nunamaker, “An exploratory study into deception detection in text-based computer-mediated communication,” in *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the*, 2003, pp. 10--pp.

[25] W. Ferreira and A. Vlachos, “Emergent: a novel data-set for stance classification,” in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 2016, pp. 1163–1168.

[26] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[27] W.-Y. Loh, “Classification and regression trees,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 1, no. 1, pp. 14–23, 2011.

[28] T. Bodnar, C. Tucker, K. Hopkinson, and S. G. Bilén, “Increasing the veracity of event detection on social media networks through user trust modeling,” in *2014 IEEE International Conference on Big Data (Big Data)*, 2014, pp. 636–643.

[29] F. Yu *et al.*, “Detecting Rumors from Microblogs with Recurrent Neural Networks,” *CoRR*, vol. 8, no. 1, pp. 1435–1443, Aug. 2016, doi: 10.1016/j.cogsys.2019.07.004.

[30] E. Kochkina, M. Liakata, and A. Zubiaga, “All-in-one: Multi-task Learning for Rumour Verification,” in *Proceedings of the 27th International Conference on Computational Linguistics*, Aug. 2018, pp. 3402–3413.

[31] A. Jacovi, O. S. Shalom, and Y. Goldberg, “Understanding convolutional neural networks for text classification,” *arXiv Prepr. arXiv1809.08037*, 2018.

[32] F. Yu, Q. Liu, S. Wu, L. Wang, and T. Tan, “Attention-based convolutional approach for misinformation identification from massive and noisy microblog posts,” *Comput. Secur.*, vol. 83, pp. 106–121, 2019.

[33] C. Song, C. Yang, H. Chen, C. Tu, Z. Liu, and M. Sun, “CED: Credible early detection of social media rumors,” *IEEE Trans. Knowl. Data Eng.*, 2019.

[34] O. Ajao, D. Bhowmik, and S. Zargari, “Fake News Identification on Twitter with Hybrid CNN and RNN Models,” in *Proceedings of the 9th International Conference on Social Media and Society. ACM*, 2018, no. July, pp. 226–230, doi: 10.1145/3217804.3217917.

[35] W. Wang, “‘Liar, Liar Pants on Fire’: A New Benchmark Dataset for Fake News Detection,” 2017.

[36] P. Patwa *et al.*, “Fighting an infodemic: Covid-19 fake news dataset,” *arXiv Prepr. arXiv2011.03327*, 2020.

[37] T. Felber *et al.*, “Constraint 2021: Machine Learning Models for COVID-19 Fake News Detection Shared Task,” *arXiv Prepr. arXiv2101.03545*, 2021.

[38] S. D. Das, A. Basak, and S. Dutta, “A Heuristic-driven Ensemble Framework for COVID-19 Fake News Detection,” *arXiv Prepr. arXiv2101.03545*, 2021.

2021 Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 28-29, 2021

An Evolution of Style Transfer from Artistic to Photorealistic : A Review

Gaurav Sohaliya
Department of Information Technology
Delhi Technological University
Delhi, India
sohaliyagaurav@gmail.com

Kapil Sharma
Department of Information Technology
Delhi Technological University
Delhi, India
kapil@ieee.org

***Abstract*— Redrawing an image in a given style requires a professional artist, but Neural Style Transfer (NST) opens a way to generate art in the desired style using computers. NST uses a Convolution Neural Network (CNN) to draw a content image in a given artistic style. It showed the potential of CNN in producing painterly images by extracting and blending image style and content. Since then, various methods are introduced to enhance and extend the fundamental NST approach, becoming an interesting topic to research. Generative Adversarial Networks (GANs) were recently introduced, which have shown excellent potential in image-to-image translation. The purpose of image-to-image translation is similar to the style transfer task, and this new GAN-based approach extended the boundaries of style transfer from artistic applications to real-life use cases. This paper will first deliver a comprehensive summary of CNN-based and GAN-based state-of-the-art style transfer algorithms and then do a qualitatively comparative study of these algorithms. Finally, this review concludes with several applications of style transfer and identifies some research gaps for future research.**

Keywords—Style transfer, Convolution neural network, Generative adversarial networks, Image generation, Deep learning

I. INTRODUCTION

Style transfer is redrawing Semantic Content in different artistic styles. It is simply generating a new image with the content of one provided image but in the style of another image. Neural style transfer (NST) [1] was the first example of style transfer. It is based on a CNN which is generally utilized to classify images. VGG-19 [2] is a pre-trained Deep CNN, is employed as the feature extractor in the original paper.

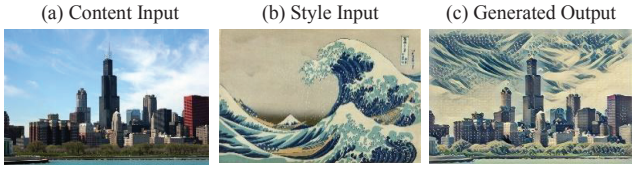


Fig. 1 Neural Style Transfer result. [1]

Fig. 1 (a) is a content image, and image Fig. 1 (b) is a style image by famous Japanese painter Katsushika Hokusai. Fig. 1 (c) is the output image of NST, which has content like buildings, trees, sky, etc., extracted from a given content input but in the style of a given style input. NST majorly focuses on artistic style transfer [3], and this is the main limitation of this algorithm. NST usually produces painterly images. Hence, NST can be used only for art generation and entertainment purposes. Deep Photo Style Transfer [3] is proposed to overcome the limitation of NST at a certain level. NST works amazingly well if we're looking for a painterly result, but if we apply this algorithm for real-life images, which contain complex objects and scenes, it introduces unnecessary distortions to the image, refer to Fig 2 (c). They won't look realistic anymore. Deep Photo Style Transfer introduced Photorealism (Photorealistic Style Transfer) and Semantic accuracy (Semantic Style Transfer) in the NST algorithm. It can generate a beautiful and realistic stylized image, refer to Fig 2 (d).

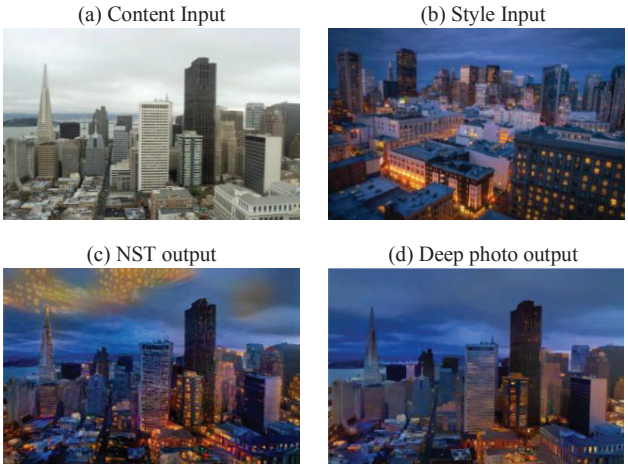


Fig. 2 Comparison between NST and Deep photo style transfer result. From [3]

Many advancements are made on NST to overcome its limitation and to add new features. NST algorithm takes a long time to stylize a single image [4], so it cannot use in real-time applications. Johnson et al. used image transform net to reduce time and proposed a real-time style transfer model [4], but this model is bound to only one style. Conditional Fast Style Transfer [5] can perform style transfer for multiple styles in real-time. Instance Style Transfer [6] aims to stylize only a particular object of a given image. Video Style transfer [7, 8] can stylize video with smooth frame transition. But all the proposed methods still have one major limitation: Incapability of learning the complete style [9, 10] of any specific domain such as day, night, winter, summer, child, old, etc. An example given in Fig. 2 is a day-to-night transition, but it only works if content and style image are semantically similar. NST model cannot transfer any given image of a day into night; it will fail, as shown in Fig. 3. It means the model is not learning the

2021 Asian Conference on Innovation in Technology (ASIANCON) | 978-1-7281-8402-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ASIANCON51346.2021.9544924

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:48:44 UTC from IEEE Xplore. Restrictions apply.

complete domain style. Failure happens because the model doesn't have a complete understanding of the night domain style. The model cannot predict that how all different object appears in the night, from the given style image (Fig. 3 (b)) model can only get the idea about building and sky night style, not about trees and grass (Fig. 3 (a)) night style.



Fig. 3 Day to night transition with NST

The reason behind the failure is that NST uses single style and single content images; it cannot extract style from a collection of style images. Hence it creates a lack of information regarding the complete style. To extract style features of domain like winter, summer, day, night, or mimic style of artist or genre's style [10] using image dataset, is a complicated task in NST. NST unable to do style transition from an input domain to a different domain [9]. In 2014, Ian Goodfellow proposed Generative adversarial nets (GANs) [11]. GANs are powerful generative models. It has achieved remarkable outcomes in image generation. As GAN is a generative model, it can also use in style transfer. Style transfer using GANs can be achieved by the image-to-image translation (I2I) [12] method. This method generally takes images from one class and tries to represent them so that images have style from another class. It does not precisely match the procedure of style transfer, which generates a new image by merging two different attributes of two different images. Still, both these concepts have the same goal and are also used as synonyms. Two methods exist for I2I: The Unpaired method and a Paired method.

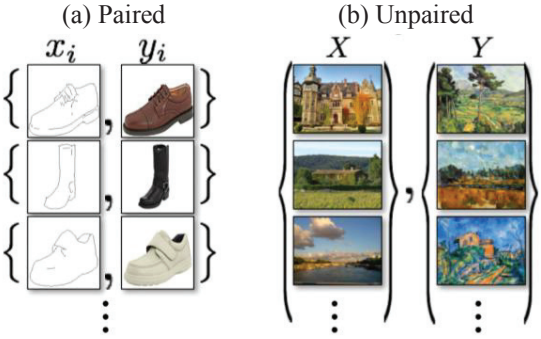


Fig. 4 Paired and Unpaired image dataset. From [9]

As shown in Fig. 4, the Paired approach requires paired image dataset. In an unpaired approach, there is no need that a particular image in the domain should be present in the codomain. For example, in Fig. 4, x_i and X are considered as content datasets whereas y_i and Y are considered as style datasets. The semantically similar images of x_i must be present in dataset y_i in a paired dataset, but this is not mandatory for an unpaired dataset. Pix2pix [13] architecture works on a paired approach, whereas Cycle GAN [9], and Disco GAN [14] work on an unpaired approach.

GAN can effectively solve the limitation of NST algorithms. It also opens some new research areas for style transfer. So, in subsequent sections, we do a comprehensive and comparative study on style transfer using both CNN and GANs and identify the result differences and limitations of these approaches. We also discuss possible ways to overcome those limitations and applications of style transfer in real life.

II. Style Transfer using CNN

A. Convolutional Neural Network (CNN)

CNNs [2] are robust classifiers that belong to the class of deep and feedforward neural networks. The Essential unit of CNN is Convolution. Convolution is an operation where the convolution matrix, also known as filter or kernel, move across the image and gives the sum of products of filter elements and the Corresponding image pixels. The downsampling block is one more vital unit of a CNN. The objective of downsampling is to reduce the dimensionality of image features. It helps to reduce computation by reducing the training parameters and helps to overcome the overfitting of the model. It gives the model more robustness to input changes.

A CNN is comprised of two essential components of feature extractor and classifier. Feature extraction contains a stack of convolution, downsampling and an activation layers. Fully connected layers (multilayer perceptron) are utilized to build a classification network. The input image is given to the feature extraction unit, which generates a feature map. This map then passes to a classifier unit, which allocates a label to the image, as per its features.

B. Neural Style Transfer

1) Content Representation: In CNN, feature extraction is done hierarchically. So that response of lower layer (layers that are nearest to input layer) of CNN is much similar to input image because feature extraction at lower layers is at low tendency, so it just reproduces much similar output with input. Higher layers give high-level content (e.g., objects in the scene, the relative positions of objects), so that we prefer to use a higher-level feature map as content representation.

So, let $\vec{o}$ and $\vec{g}$ be the original and generated image, respectively, whereas O^l and G^l are their corresponding feature maps in layer l. The feature reconstruction loss for a layer l can be formulated as:

$$\mathcal{L}_{content}(\vec{o}, \vec{g}, l) = \frac{1}{2} \sum_{ij} (G_{ij}^l - O_{ij}^l)^2 \qquad (1)$$

The above (1) is for one-layer content representation only; we can take multiple layers, with each layer has a different contribution factor.

2) Style Representation: The correlations among the filters represent the style of an image. These correlations are yielded using the Gram Matrix G_{ij}^l:

$$G_{ij}^l = \sum_k F_{ik}^l \cdot F_{jk}^l \qquad (2)$$

Here F^l is feature map at layer l. By incorporating feature correlations of layers, we can capture the style information of the image while ignoring the content information (e.g., objects

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:48:44 UTC from IEEE Xplore. Restrictions apply.

present in scenes, global arrangements of objects). Let $\vec{o}$ and $\vec{g}$ be the original and generated image, respectively, and O^l and G^l be their style feature map of layer l respectively with $N_l \times M_l$ dimensions. The style loss function for single layer l is constructed as follows:

$$\mathcal{L}_{style}(\vec{o}, \vec{g}, l) = \frac{1}{4N_l^2 M_l^2} \sum_{i,j} \left(G_{ij}^l - O_{ij}^l\right)^2 \quad (3)$$

3) Algorithm: Given style image $\vec{s}$ and content image $\vec{c}$, the algorithm tries to seek an aesthetic stylized image $\vec{g}$ that jointly minimizes content and style loss. The NST optimization function is:

$$\mathcal{L}_{total}(\vec{c}, \vec{s}, \vec{g}) = \alpha \mathcal{L}_{content}(\vec{c}, \vec{g}) + \beta \mathcal{L}_{style}(\vec{s}, \vec{g}) \quad (4)$$

α and β are regulators applied to regulate the stylized result's content component and style component. The working procedure of NST is as follow: $\vec{c}$ and $\vec{s}$ is passed through pre-trained VGG and content features C and style features S is extracted and stored. Then random noise image $\vec{g}$ is passed through VGG, and its content features C' and style features S' is calculated. Then content loss between C and C', and style loss between S and S' is calculated as $\mathcal{L}_{content}$ and $\mathcal{L}_{style}$. Finally, $\mathcal{L}_{total}$ is derived as per (1), and it is used to optimize image $\vec{g}$. This whole cycle is done iteratively till content features and style features of output image $\vec{g}$ match with content feature of $\vec{c}$ and style feature of $\vec{s}$, respectively.

C. Deep Photo Style Transfer

When the Content and Style images are photographs, the generated image looks similar to painting. The produced output is distorted. e.g., edges and textures wavy. This is because the NST algorithm concentrates on extracting the global artistic style of an image and transfers it into another. There is no constrain present in NST to generate photorealistic results. Hence, it is unsuitable for photorealistic style transition. Recent work "Deep photo style transfer" [3] introduces a phenomenon called photorealism. It is enforced on the resulting image to generate photorealistic output. This method effectively removes distortion and produces impressive photorealistic images in various situations, including daytime transition, season-weather transition, etc. Another significant contribution is semantic accuracy, which is a solution to undesirable style transfers between unrelated contents.

1) Photorealism: Describing photorealism in a scientific equation is an unsolved problem. But it is easy to obtain photorealism if the input picture is photorealistic. The strategy is to create a loss function that preserves the edges of an input image in small patches to prevent distortions. Then, locally affine transformations are adopted to design this function. Finally, this measure is applied throughout the transformation process. The affine loss is constructed as:

$$\mathcal{L}_m = \sum_{c=1}^{3} V_c[O]^T \mathcal{M}_I V_c[O] \quad (5)$$

Where $V_c[O]$ be the vectorized color channel c of the transfer image O. $\mathcal{M}_I$ is matting laplacian calculated to extract an object from an image.

2) Semantic accuracy and Faithful Transformation: In NST, the style loss is computed over the entire image, and the gram matrix represents the global style of an image. But in reality, each object has its unique style in real-life photographs, and they cannot be portraits using a single style. Moreover, NST style loss does not take semantic regions into account. The semantic segmentation method *[15]* is used to achieve the style transfer between the corresponding input image and style image regions. It generates an image segmentation mask for both the input and style images for common labels such as sky, building, water, etc., and then transfer style according to segmentations. Semantic segmentation of the inputs gives arbitrary direction to the transformation process to prevent the content-mismatch problem. This method generates excellent photorealistic results. The updated style loss is constructed as follows:

$$\mathcal{L}_{style+}(\vec{o}, \vec{g}, l) = \sum_{c=1}^{C} \frac{1}{2N_{l,c}^2} \sum_{i,j} (G_c^l - O_c^l)^2 \quad (6)$$

Where c is the segmentation class, C is the number of channels in the semantic segmentation mask, gram matrix of every class is calculated with its corresponding mask, and other things are same as style representation (3).

3) Algorithm: The Final optimization function is constructed by combining all three components:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{content} + \beta \mathcal{L}_{style+} + \lambda \mathcal{L}_m \quad (7)$$

λ is a parameter to regulate the photorealism. In semantically uniform regions, sometimes, deep photo algorithm generates inconsistent stylizations. The fast photorealistic image stylization method [16] is proposed to avoid inconsistent stylizations and generate more photorealistic results. This method is more effective and efficient than the deep photo algorithm. Fig. 5 shows the impressive results obtained by photorealistic style transfer.

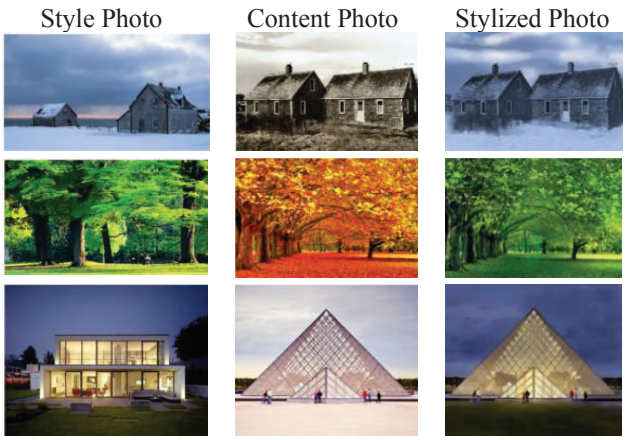


Fig. 5 Photo realistic style transfer results. From [16]

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:48:44 UTC from IEEE Xplore. Restrictions apply.

D. Real-Time Style Transfer

Previously proposed style transfer methods take 3-5 minutes per image generation; hence it is not a real-time application. To overcome this limitation, Johnson et al. introduce a feedforward convolution neural network [4]. This feedforward network stores the optimization-transformation process of style transfer. Hence when you pass any content image through the network, deconvolution layers of the network apply the learned transformation process to the image. It is a one-shot process after training the model; because of that, this method takes few seconds to generate a stylized image.

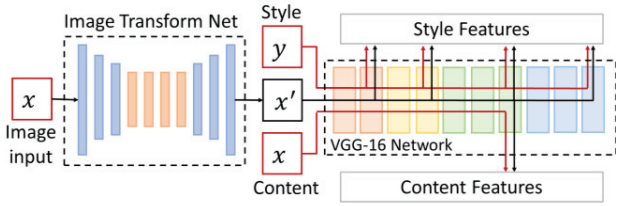


Fig. 6 Feedforward Network. From [4]

As shown in Fig. 6, the model takes content input x and passes through the Image Transform Net (feedforward network), which effectively learns the style transfer process and generates output image x'. Then after the remaining process is similar to NST, it takes VGG net to calculate content loss between content features of x' and content image x. It also calculates style loss between style features of x' and style image y. These losses are used to optimize the network parameters.

III. Style Transfer using GAN

A. Generative Adversarial Networks (GAN)

GAN [11] is constructed of a generator that creates an image from the random noise and a discriminator that determines if the freshly produced image is real or artificial. Assume that the discriminator considers the generator's result to be artificial, then the generator is penalized. The discriminator network loses when it fails to distinguish as an artificial image and categorize as a real image. So, it is a Game theory algorithm, both the networks attempting to deceive each other, and improve themselves by bearing loss. The objective of GAN is:

$$\min_G \max_D \mathbb{E}_{x\sim\mathbb{P}_r}[\log D(x)] + \mathbb{E}_{z\sim\mathbb{P}_g}\left[\log\left(1 - D\left(G(z)\right)\right)\right] \quad (8)$$

where $x\sim\mathbb{P}_r$ are images from the input dataset, and $z\sim\mathbb{P}_g$ are random noise units, $G(z)$ are the generated images and $D(x)$ is the probability of x being real.

B. Pix2Pix

Pix2Pix [13] uses conditional GAN model and adversarial training to learn $G: X \rightarrow Y$ mapping thus $\hat{y} = G(x), x \in X$ output is non-differentiable from $y \in Y$ by a discriminator. This is the first automatic I2I method. The architecture of Pix2Pix (as shown in Fig. 7) consists of one generator and one discriminator. It uses U-net as base generator architecture which uses to avoid bottleneck problem of basic encode-decoder problem and generate high resolution I2I results. Pix2Pix uses PatchGAN as discriminator because it captures local style statistics and model high-frequencies.

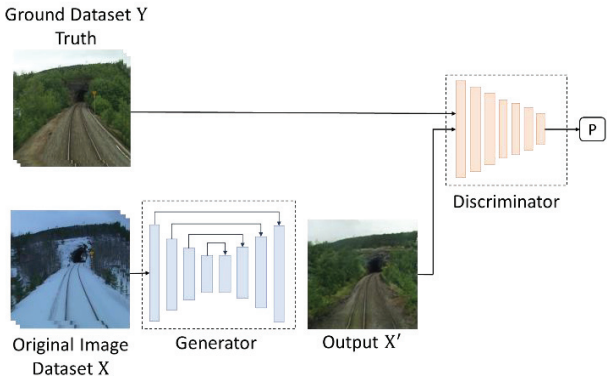


Fig. 7 Pix2Pix Architecture

The optimization function of Pix2pix consist one adversarial loss and one L1 distance loss. Following is conditioned adversarial loss adopted from cGAN:

$$\mathcal{L}_{cGAN}(G,D) = \min_G \max_D \mathbb{E}_{x\sim\mathbb{P}_r}[\log D(x,y)] + \mathbb{E}_{z\sim\mathbb{P}_g}\left[\log\left(1 - D\left(x, G(z)\right)\right)\right] \quad (9)$$

Here auxiliary condition or information x emphasized over discriminator. The L1 loss brings crispiness and reduce blurriness in resultant image in compare to L2 loss which is previously used in cGAN. Following equation is for L1 distance:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\| y - G(x,z) \|] \quad (10)$$

The final Pix2pix objective is:

$$G^* = arg \min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda\mathcal{L}_{L1}(G) \quad (11)$$

Fig. 9 shows the Pix2Pix aerial to map images transfer results. The Result obtained by Pix2Pix are absolutely stunning.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:48:44 UTC from IEEE Xplore. Restrictions apply.

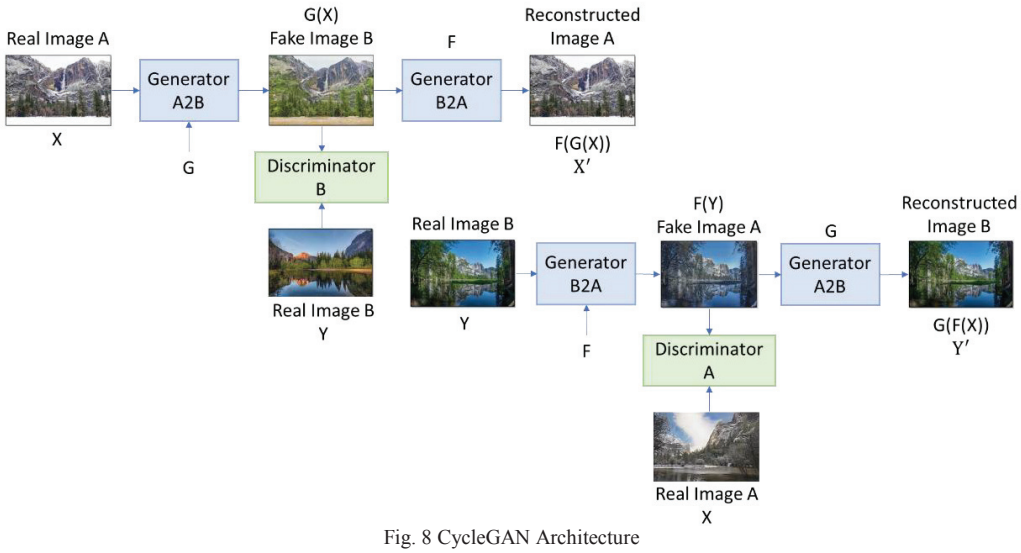


Fig. 8 CycleGAN Architecture

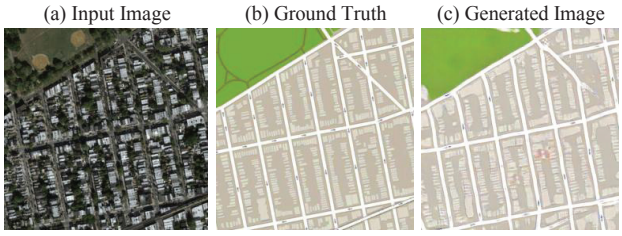


Fig. 9 Pix2pix Aerial to Map Translation

C. *Cycle GAN*

CycleGAN [9] follows an unpaired approach. To perform unpaired I2I; Cycle Consistency Loss (CCL) is introduced with the CycleGAN architecture. CycleGAN (Fig. 8) consists of two units. The First component tries to map $X \rightarrow Y$ domain. In contrast, the second component maps $Y \rightarrow X$ domain. Both these units are GANs, so they are composed of two Cycle-consistent generative models. CCL enforces invertible and bijection mapping between domains. Cycle Consistency Loss (Fig. 10) inspires $F(G(x)) \approx y$ and $G(F(y)) \approx x$ where $x \in X$ and $y \in Y$. It reduces the difference between input image x and reconstructed image $\hat{x}$ and same as for y and $\hat{y}$ by enforcing forward and backward consistency.

The cycleGAN consist two GANs model so that there is one CCL and two adversarial (GAN) losses. Following is GAN loss for the mapping $G: X \rightarrow Y$:

$$\mathcal{L}_{GAN}(G, D_y, X, Y) = [\log D_y(y)] + \mathbb{E}_{x \sim \mathbb{P}_x}\left[\log\left(1 - D_y(G(x))\right)\right] \quad (12)$$

The CCL is between $F(\hat{y})$ and $x \in$ X is:

$$\mathcal{L}_{cycle} = \|F(\hat{y}) - x\|_1 = \left\|F(G(x)) - x\right\|_1 \quad (13)$$

The CycleGAN objective is:

$$\mathcal{L}(G, F, D_x, D_y) = \mathcal{L}_{GAN}(G, D_y, X, Y) + \mathcal{L}_{GAN}(G, D_x, Y, X) + \lambda\mathcal{L}_{cycle} \quad (14)$$

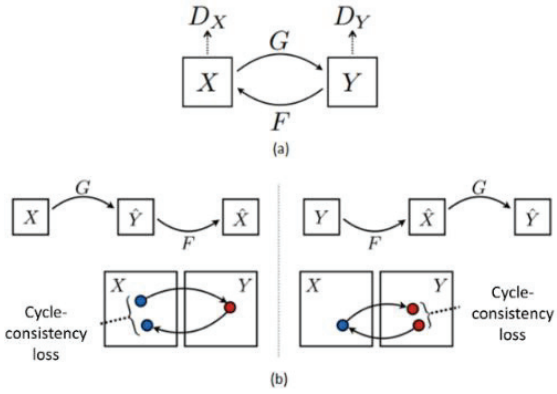


Fig. 10 Cycle consistency loss. From [9]

The final aim is to solve:

$$G^*, F^* = arg\ \min_{G,F} \max_{D_x, D_y} \mathcal{L}(G, F, D_x, D_y) \quad (15)$$

Fig. 11 shows the CycleGAN results. The results obtained by CycleGAN are excellent.



Fig. 11 CycleGAN style transfer results [9]

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:48:44 UTC from IEEE Xplore. Restrictions apply.

D. DiscoGAN

DiscoGAN [14] was proposed by Taeksoo Kim et al. in 2017. DiscoGAN is quite similar to CycleGAN; however, it introduces reconstruction loss, a similar concept of cycle consistency loss. Reconstruction loss computes how properly the input image is reconstructed after two iterations, and the goal is to minimize them. However, there's a difference between CycleGAN and DiscoGAN in measuring distance and the structure of reconstruction loss (or cycle consistency loss). First, CycleGAN measures the distance of cycle consistency with hinge loss, but DiscoGAN uses MSE loss in reconstruction loss. CycleGAN uses one cycle-consistency loss, while DiscoGAN requires two reconstruction losses for domains A and B. In the context of model architecture, DiscoGAN follows encoder-decoder structures like DCGAN with leaky relu activation as Generator and CNN with leaky relu as Discriminator, whereas CycleGAN uses Instance normalization as Generator. While transforming image color and pattern of a domain to codomain, it generates good results, but it fails while adjacent objects are present in the image, e.g., a person riding a horse.

IV. Applications

Style transfer has several practical applications in the entertainment and industrial field. NST algorithms produce good aesthetic and artistic results. So NST is useful in photo and video editing software. Prisma [17] is one of the first photo editing software which uses NST algorithms. Due to the high popularity of Prisma, many other social media sites such as Snapchat and Instagram also included the filter feature, which uses the NST algorithm for image processing. NST is widely used to generate license-free artwork, which is further used in fashion designing, art development, 3D model generation with a new style, etc.

The GAN approach for style transfer opens many new application fields because it can do domain style transfer. GAN-based style transfer algorithms are used in photo enhancement. Nowadays, the mobile camera gives DSLR effects in photos using the dual-camera system, but this can be done using a single camera and GAN style transfer algorithms [9]. Cycle GAN showed promising results in photo improvement applications such as defogging, deblurring, stabilization, super-resolution, etc [18]. Cycle GAN is very useful in domain transformation such as season transfer [9], face to emoji transfer [12], age and gender transfer [19], etc. FaceApp [20] is a mobile application that is one of the best examples of age and gender transfer. Since Pix2pix work on paired datasets still has many significant applications such as black and white images to colorized images [12], aerial images to map images [13], night to day transfer [21], etc. Night to day transfer can be helpful to improve night photography results and also while driving at night. As style transfer can also apply in the video, it makes cinematography, VFX, and CGI task very easy.

V. Comparative Analysis and Future Scope

We have done a comparative analysis of only state-of-the-art style transfer algorithms ([1], [3], [13], [9]) in the season transition domain (Winter to summer). The GAN-based algorithm requires a dataset for this process, where CycleGAN requires unpaired, and Pix2Pix requires paired image dataset. Hence, we used an unpaired image dataset of Yosemite National Park [9], taken during the summer and winter, for CycleGAN. Unfortunately, there is no paired dataset available for season transfer, so we extracting images from 10 hours of recorded footage of the train journey "Nordland Line" in Norway [22] to create dataset. Fig. 12 and Fig. 13 shows the results [23] [24].

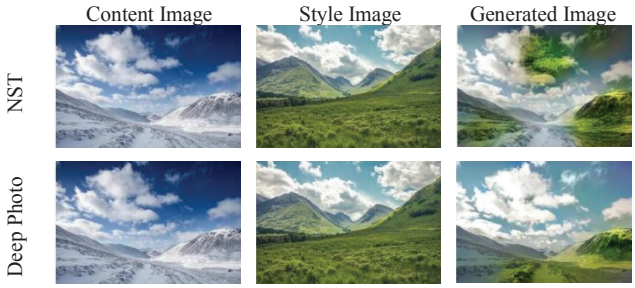


Fig. 12 CNN-based algorithms results

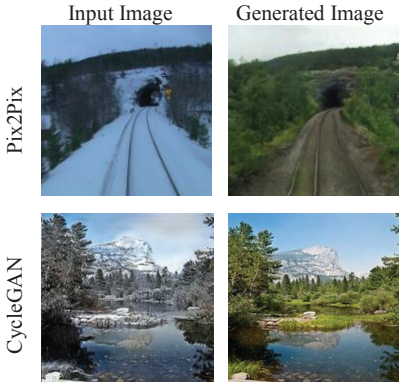


Fig. 13 GAN-based algorithms results

CNN-based algorithms are struggling, whereas GAN-based algorithms are producing excellent results in season transfer. There is no evaluation matric available to evaluate performance quantitatively [25]; hence we can only perform a qualitative evaluation. As a result of qualitative assessment, GAN-based algorithms perform well, but we found some limitations to these algorithms. Pix2pix requires paired dataset, and it is very costly and challenging to collect paired datasets in industrial applications. CycleGAN is an unpaired approach, but it requires high computation power and takes several days or weeks of training time to achieve good results because CycleGAN consists of six neural networks in its architecture (Fig. 8). This limitation bounds the advanced use cases of CycleGAN. To reduce training time, we can build a modified architecture of CycleGAN, which requires a smaller number of neural networks to train. Time-efficient CycleGAN

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:48:44 UTC from IEEE Xplore. Restrictions apply.

can also use in style transfer of Video, 3D modeling, VR, etc. smoothly.

VI. Conclusion

Style transfer potential is not only limited to generate aesthetic artwork but it is also able to solve many complex image-to-image translation and computer vision problems. It is useful in day-to-day life through social media, mobile camera, photo editing applications etc. Due to introduction of Generative adversarial networks into style transfer, it can mimic a painting style of any artist. It is a kind of visual intelligence. Most of the research has been done on images only because style transfer requires high computational power and graphics hardware but there is a scope of good research in video, AR-VR and 3D model style transfer as well.

References

[1] L. A. Gatys, A. S. Ecker and M. Bethge, "A Neural Algorithm of Artistic Style," CoRR, vol. abs/1508.06576, 2015.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[3] F. Luan, S. Paris, E. Shechtman and K. Bala, "Deep Photo Style Transfer," CoRR, vol. abs/1703.07511, 2017.

[4] J. Johnson, A. Alahi and F.-F. Li, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," CoRR, vol. abs/1603.08155, 2016.

[5] K. Yanai and R. Tanno, "Conditional Fast Style Transfer Network," in Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, New York, NY, USA, 2017.

[6] Z. Yu, Y. Wu and T. Wang, "A Method for Arbitrary Instance Style Transfer," CoRR, vol. abs/1912.06347, 2019.

[7] M. Ruder, A. Dosovitskiy and T. Brox, "Artistic style transfer for videos," CoRR, vol. abs/1604.08610, 2016.

[8] X. Xia, T. Xue, W.-s. Lai, Z. Sun, A. Chang, B. Kulis and J. Chen, "Real-time Localized Photorealistic Video Style Transfer," CoRR, vol. abs/2010.10056, 2020.

[9] J.-Y. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," CoRR, vol. abs/1703.10593, 2017.

[10] Z. Ma, J. Li, N. Wang and X. Gao, "Image style transfer with collection representation space and semantic-guided reconstruction," Neural Networks, vol. 129, pp. 123-137, 2020.

[11] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial networks," arXiv preprint arXiv:1406.2661, 2014.

[12] Y. Pang, J. Lin, T. Qin and Z. Chen, "Image-to-Image Translation: Methods and Applications," CoRR, vol. abs/2101.08629, 2021.

[13] P. Isola, J.-Y. Zhu, T. Zhou and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," CoRR, vol. abs/1611.07004, 2016.

[14] T. Kim, M. Cha, H. Kim, J. K. Lee and J. Kim, "Learning to Discover Cross-Domain Relations with Generative Adversarial Networks," CoRR, vol. abs/1703.05192, 2017.

[15] M. Thoma, "A Survey of Semantic Segmentation," CoRR, vol. abs/1602.06541, 2016.

[16] Y. Li, M.-Y. Liu, X. Li, M.-H. Yang and J. Kautz, "A Closed-form Solution to Photorealistic Image Stylization," CoRR, vol. abs/1802.06474, 2018.

[17] I. P. Labs, "Prisma: Turn memories into art using artificial," 2016. [Online]. Available: http://prisma-ai.com.

[18] S. Kaji and S. Kida, "Overview of image-to-image translation by use of deep neural networks: denoising, super-resolution, modality conversion, and reconstruction in medical imaging," Radiological physics and technology, vol. 12, p. 235–248, 2019.

[19] S. Palsson, E. Agustsson, R. Timofte and L. Van Gool, "Generative Adversarial Style Transfer Networks for Face Aging," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018.

[20] W. Lab, "FaceApp - AI Face Editor," 2016. [Online]. Available: https://www.faceapp.com/.

[21] I. Anokhin, P. Solovev, D. Korzhenkov, A. Kharlamov, T. Khakhulin, A. Silvestrov, S. I. Nikolenko, V. Lempitsky and G. Sterkin, "High-Resolution Daytime Translation Without Domain Labels," CoRR, vol. abs/2003.08791, 2020.

[22] T. N. B. C. (NRK), "Nordlandsbanen: minute by minute, season by season," 2013. [Online]. Available: https://nrkbeta.no/2013/01/15/nordlandsbanen-minute-by-minute-season-by-season/.

[23] S. Suri, A. Gupta and K. Sharma, "Comparative Analysis of Ranking Algorithms Used On Web," Annals of Emerging Technologies in Computing (AETiC), vol. 4, 2020.

[24] A. K. Tripathi, K. Sharma, M. Bala, A. Kumar, V. G. Menon and A. K. Bashir, "A Parallel Military-Dog-Based Algorithm for Clustering Big Data in Cognitive Industrial Internet of Things," IEEE Transactions on Industrial Informatics, vol. 17, p. 2134–2142, 2020.

[25] Y. Jing, Y. Yang, Z. Feng, J. Ye and M. Song, "Neural Style Transfer: A Review," CoRR, vol. abs/1705.04058, 2017.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:48:44 UTC from IEEE Xplore. Restrictions apply.

An Explicit Output Current-mode Quadrature Sinusoidal Oscillator and a Universal Filter Employing Only Grounded Passive Components - a Minimal Realisation

Tajinder Singh ARORA[1], *Manish GUPTA*[2], *Shiv Narain GUPTA*[3]

[1]Department of Electronics Engineering, National Institute of Technology, Uttarakhand, NH 58, 246174 Srinagar, India
[2]Department of Electronics and Communication Engineering, Inderprastha Engineering College, 63 Site IV, Sahibabad Industrial Area, Surya Nagar Flyover Road Sahibabad, 201010 Ghaziabad, India
[3]Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, A-10, Sector 62, 201309 Noida, India

tajarora@gmail.com, guptamanish2004@gmail.com, shivgnit@gmail.com

DOI: 10.15598/aeee.v19i3.4121

Article history: Received Feb 23, 2021; Revised Jul 06, 2021; Accepted Jul 07, 2021; Published Sep 30, 2021.
This is an open access article under the BY-CC license.

Abstract. *The use of voltage differencing current conveyor as an active device to design a current-mode oscillator along with a universal filter with only grounded passive elements is the main focus of this manuscript. This re-arranging circuit can work as a sinusoidal oscillator as well as a current-mode universal filter, by simple selection of passive switches. Both the circuits employ only two active devices and three grounded passive elements. The designed oscillator provides two distinctive current outputs with a quadrature-phase difference. It also maintains an independent condition of oscillation and frequency of oscillation. Moreover, the basic responses including low pass, high pass, and band pass are easily available from a current-mode universal filter. The low input impedance and high output impedance are amongst the noteworthy features of the current-mode derived filter. Non-ideal, parasitic, and sensitivity analysis of the designed circuits are also incorporated in the manuscript. Cadence PSPICE software simulation results are also included to justify the design idea. Experimental implementation of the described circuit has also been shown by employing special-purpose amplifier integrated circuit, i.e., OPA860.*

Keywords

Active filter, sinusoidal oscillator, Voltage Differencing Current Conveyor.

1. Introduction

The domain of analog signal processing is all about measurement, detection, and manipulation of the analog signals. Analog circuits find their applications in the area of control systems, communication engineering, instrumentation, and measurements [1]. Despite the dominance of digital signal processing, the analog signal processing finds its irreplaceable space in optical drives, analog to digital converters, etc. [2].

The rich culture of analog circuits is filled with multitudinous active devices. Few names in this long list are as follows: Current Conveyor (CC) [3] and [4], Operational Transconductance Amplifier (OTA) [5] and [6], Current Differencing Buffered Amplifier (CDBA) [7], Differential Voltage Current Conveyor (DVCC) [8], Current Differencing Transconductance Amplifier (CDTA) [9], Differential Voltage Current Conveyor Transconductance Amplifier (DVCCTA) [10], Unity Gain Cell (UGC) [11], Differential Difference Current Conveyor (DDCC) [12], Z-Copy Current Feedback Transconductance Amplifier (ZCCFTA) [13], etc. Each of the mentioned devices has its own uniqueness and features.

In [14], the authors proposed future looming devices, including Voltage Differencing Current Conveyor (VDCC). This is an electronic combination of OTA and

 © 2021 ADVANCES IN ELECTRICAL AND ELECTRONIC ENGINEERING

CC, and hence features/advantages of two different devices can be attained by this one device. The former input stage offers an electronically tunable transconductance (gm), whereas the latter stage conveys current from the input terminal to the output terminal. This active element is not explored and utilised up to its competence. The available applications in the open literature, realised with VDCC and its variants are mainly active filters and sinusoidal oscillators [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], and [27]. In a recent span of time, researchers have introduced different sinusoidal oscillators using different active building blocks. A comprehensive literature survey on the design of sinusoidal oscillators including VDCC has been performed; some prominent references which were taken into account are expressed in Tab. 1.

It can be summarised from Tab. 1 that:

- References [8], [9], [11], [28], [29], [30], [31], [32], [33], [34], [35], [36], and [37] do not utilise all grounded components.
- Higher number of passive elements are utilised in [8], [10], [11], [12], [15], [16], [17], [19], [28], [29], [30], [31], [32], [33], [34] [35], [37], [38], [39], [40], [41], and [42]; whereas higher count of active devices is encountered in [11], [13], [34], [41], [42], [43], [44], [45], and [46].
- Quadrature outputs are not available in [8], [10], [20], [28], [30], [33], [35], [40], and [43], whereas simultaneous availability of Current Mode (CM) and Voltage Mode (VM) is not provided by [9], [12], [13], [15], [16] [17], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [39], [40], [41], [43], [44], [45], and [46].
- Independent tunability of Condition of Oscillation (CO) along with the Frequency of Oscillation (FO) is not supported in references [8], [9], [16], [17], and [38].

There are only references [18], [19], [42], and [47] that fulfil all given constraints, but [19] and [42] use more passive elements, whereas [18] and [47] can work as an oscillator only (cannot perform the functions of an active filter). This manuscript is an attempt towards presenting a sinusoidal oscillator as well as a universal filter, both attained from the same configuration, employing VDCC and only grounded passive components. Designed circuits encounter the use of two passive switches, whose appropriate connection results in either an oscillator or a current mode active filter, delivering explicit current outputs in both cases. The fabrication of the circuit is also convenient as only grounded passive elements are used. The quadrature results are obtained in the oscillator circuit, in both CM as well as VM of operation. In the oscillator, CO and FO are independent and simple. The design of the universal filter provides all basic responses, like Low Pass (LP), High Pass (HP), Band Pass (BP), Band Reject (BR), and All Pass (AP).

The overall circuit offers low input and high output impedance, making it a preferable choice for cascading and for the realisation of higher-order filters.

The manuscript has been organized in various sections. The following section, i.e., Sec. 2. , will introduce the active device; VDCC. The designed configuration with its detailed ideal analysis is discussed in Sec. 3. Effect of non-idealities and sensitivity of design towards various elements is shown in Sec. 4. Section 5. details the effect of parasitic elements on the proposed circuit. The verification of the design through simulations has been represented in Sec. 6. and the justification of the same through experimentation is provided in Sec. 7. followed by concluding remarks.

2. Introduction to VDCC

The active device VDCC has in total six terminals namely P, N, Z, X, WP, and WN [22]. The input stage of the device has a current-controlled transconductance gain and a different current/voltage relationship between the different terminals, as given in Eq. (1).

$$\begin{bmatrix} I_P \\ I_N \\ I_Z \\ V_X \\ I_{WP} \\ I_{WN} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ g_m & -g_m & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} V_P \\ V_N \\ V_Z \\ I_X \end{bmatrix}. \quad (1)$$

The block diagram for the ideal VDCC is shown in Fig. 1. The realization of VDCC with Metal Oxide Semiconductor (MOS) transistors is shown in Fig. 2. The copy of the Z terminal named Zc− (for inverting output) and Zc+ (for non-inverting output) can also be found using MOS realization of the active device [19], as shown in Fig. 2.

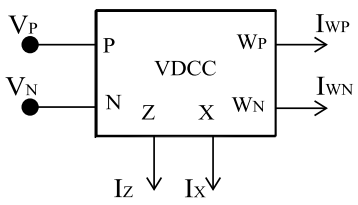


Fig. 1: Block Diagram of VDCC.

Applying the same concept [19], the copies of WP and WN terminals can also be realised, which have been named WPC and WNC, respectively.

Tab. 1: A comprehensive literature survey of sinusoidal oscillators including VDCC.

Ref. No.	Number of active devices employed	Number of used passive components	All grounded used passive elements	Availability of quadrature output	Attainment of independent CO and FO	Operating in both CM & VM
[8]	1	5	NO	NO	NO	YES
[9]	1	3	NO	YES	NO	NO
[10]	1	4	YES	NO	YES	YES
[11]	4	6	NO	YES	YES	YES
[12]	2	5	YES	YES	YES	NO
[13]	4	2	YES	YES	YES	NO
[15]	1	4	YES	YES	YES	NO
[16]	1	4	YES	YES	NO	NO
[17]	1	4	YES	YES	NO	NO
[18]	2	3	YES	YES	YES	YES
[19]	2	4	YES	YES	YES	YES
[20]	2	3	YES	NO	YES	YES
[23]	2	4	YES	YES	YES	NO
[24]	2	4	YES	YES	YES	NO
[25]	2	4	YES	YES	YES	YES
[26]	1	4	YES	YES	YES	NO
[27]	1	4	YES	NO	YES	NO
[28]	1	4	NO	NO	YES	NO
[29]	2	6	NO	YES	YES	NO
[30]	1	5	NO	NO	YES	NO
[31]	2	4	NO	YES	YES	NO
[32]	2	5	NO	YES	YES	NO
[33]	2	4	NO	NO	YES	NO
[34]	4	6	NO	YES	YES	NO
[35]	1	4	NO	NO	YES	NO
[36]	2	3	NO	YES	YES	NO
[37]	2	5	NO	YES	YES	NO
[38]	1	4	YES	YES	NO	YES
[39]	1	4	YES	YES	YES	NO
[40]	2	4	NO	NO	YES	NO
[41]	3	4	YES	YES	YES	NO
[42]	3	5	YES	YES	YES	YES
[43]	3	2	YES	NO	YES	NO
[44]	3	2	YES	YES	YES	NO
[45]	3	3	YES	YES	YES	NO
[46]	3	2	YES	YES	YES	NO
[47]	2	3	YES	YES	YES	YES
Proposed circuit	2	3	YES	YES	YES	YES

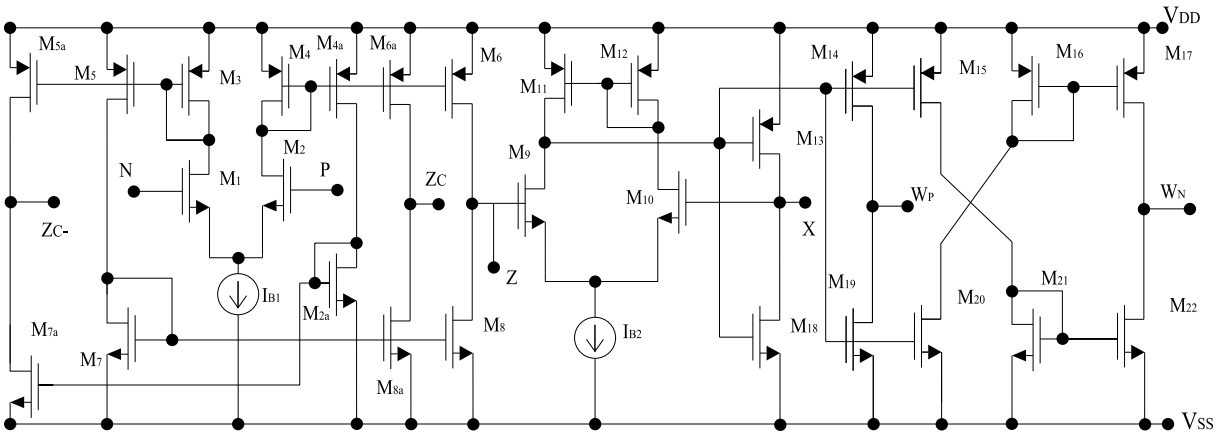


Fig. 2: CMOS realization of the Voltage Differencing Current Conveyor [19].

VDCC can also be implemented using commercially available Integrated Circuits (ICs), such as OPA 860 which is a special-function amplifier shown in Fig. 3. It is worth mentioning here that Fig. 3 is a modified version of figure of [21], which has been actively employed in this work.

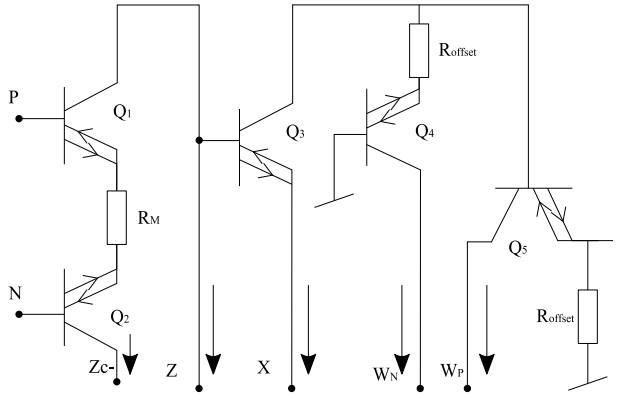


Fig. 3: Realization of VDCC utilizing discrete components.

3. Proposed Circuit

The proposed circuit diagram of a sinusoidal oscillator and a current mode active filter is shown in Fig. 4. This circuit can be converted from an oscillator to filter and vice-versa by appropriate selection of the passive switches i.e., S_1 and S_2, in Fig. 4. The patterns of engagement of these switches are given in Tab. 2.

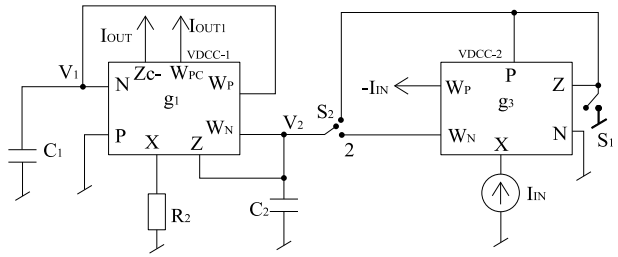


Fig. 4: Proposed circuit employing VDCC.

Tab. 2: Characteristic table for the proposed circuit.

S. No.	Switch position		Operation
	S_1	S_2	
1.	OFF	1	Dual-mode quadrature oscillator
2.	ON	2	Current-mode universal filter

Switch S_2 is the main switch that decides the functionality of the circuit, i.e., either filter or an oscillator. When switch S_2 is at position no. 1, the circuit works as an oscillator, otherwise, it acts as an active filter. The low input impedance for an active filter is achieved through switch S_1. When switch S_1 is in ON (closed) position, it makes Z terminal grounded offering low (zero) impedance at X terminal. Conversely, switch S_1 is kept in open (OFF) position for the oscillator application.

Appropriate connections of the switches (as shown in Tab. 2 at S. No. 1) result in a sinusoidal oscillator that can produce quadrature waveform in voltage mode as well as in the current mode. When simple fundamentals of circuit theory and techniques are applied to the designed circuit along with the characteristic equation of an ideal VDCC Eq. (1), we get the following equations given by Eq. (2), Eq. (3), Eq. (4), Eq. (5), and Eq. (6). The characteristic equation of the derived oscillator is represented in Eq. (2), whereas the CO and FO are represented in Eq. (3) and Eq. (4), respectively. Equation (3) shows that the CO can be independently tuned with the help of g_3 (transconductance factor of VDCC 2) without affecting the FO. It is worth noting here that FO can be electronically controlled from g_1 (the transconductance factor of VDCC 1) without altering the CO. The two output currents, i.e., I_{OUT} and I_{OUT_1}, are in quadrature with each other, as shown in Eq. (5). The relationship between the two voltages, i.e., V_1 and V_2, is shown in Eq. (6).

$$s^2 + \frac{s}{C_2}\left(\frac{1}{R_2} - g_3\right) + \frac{g_1}{R_2C_1C_2} = 0, \tag{2}$$

$$C \cdot O \cdot \left(\frac{1}{R_2} - g_3\right) = 0, \tag{3}$$

$$F \cdot O \cdot \omega_0 = \sqrt{\frac{g_1}{R_2C_1C_2}}, \tag{4}$$

$$\frac{I_{OUT}}{I_{OUT_1}} = \frac{g_1}{sC_1}, \tag{5}$$

$$\frac{V_1}{V_2} = \frac{g_2}{sC_1}. \tag{6}$$

When the switch combination of S. No. 2 from Tab. 2 is applied in Fig. 4, the proposed circuit works as a current-mode active filter. By simply applying the fundamentals of circuit simplification and the ideal equation of VDCC, i.e., Eq. (1), we get the following results for the current mode active filter. Eq. (7) and Eq. (8) show the responses of LP and BP filters. The rest of the responses, i.e., HP, BS, and AP, can be obtained as given in Eq. (9), Eq. (10) and Eq. (11). The common denominator polynomial along with angular frequency (ω_0) and quality factor (Q_0) are shown in Eq. (12), Eq. (13) and Eq. (14), respectively. Eq. (7), Eq. (8), Eq. (9), Eq. (10), and Eq. (11) justify the design idea of a universal filter.

$$\frac{I_{OUT}}{I_{IN}} = \frac{I_{LP}}{I_{IN}} = \frac{\frac{g_1}{R_2C_1C_2}}{D(s)}, \tag{7}$$

$$\frac{I_{OUT_1}}{I_{IN}} = \frac{I_{BP}}{I_{IN}} = \frac{\frac{s}{R_2C_2}}{D(s)}, \tag{8}$$

$$\frac{I_{HP}}{I_{IN}} = \frac{-I_{IN} + I_{OUT} + I_{OUT_1}}{I_{IN}} = -\frac{s^2}{D(s)}, \tag{9}$$

$$\frac{I_{BS}}{I_{IN}} = \frac{-I_{IN} + I_{OUT_1}}{I_{IN}} = -\frac{\left(s^2 + \frac{g_1}{R_2C_1C_2}\right)}{D(s)}, \tag{10}$$

$$\frac{I_{AP}}{I_{IN}} = \frac{I_{BS} + I_{BP}}{I_{IN}} = -\left(\frac{s^2 + \frac{s}{R_2C_2} + \frac{g_1}{R_2C_1C_2}}{D(s)}\right), \tag{11}$$

$$D(s) = s^2 + \frac{s}{R_2C_2} + \frac{g_1}{R_2C_1C_2}, \tag{12}$$

$$\omega_0 = \sqrt{\frac{g_1}{R_2C_1C_2}}, \tag{13}$$

$$Q_0 = \sqrt{\frac{R_2C_2g_1}{C_1}}. \tag{14}$$

4. Non-ideal and Sensitivity Analysis

In the preceding sections, we have considered only the ideal behavior model of the mentioned active device but this section of the manuscript deals with the non-ideal performance of VDCC. The characteristic equation while considering the non-idealities or commonly known tracking error of the device is given in Eq. (15) where α, β, γ_P, and γ_N are the tracking errors of the ports. Here, α is basically a transconductance error, β is the voltage transfer error between Z and X port and γ_P and γ_N are the tracking errors of W_P and W_N ports with respect to X terminal of the device.

When these non-ideal port transfer ratios/tracking errors are considered for interpreting the sinusoidal oscillator circuit (as given in Fig. 4), the following equations are derived. The characteristic equation is given by Eq. (16). The newly modified CO and FO are depicted in Eq. (17) and Eq. (18), respectively.

$$\left.\begin{array}{c} I_Z = \alpha \cdot g_m(V_P - V_N) \\ V_Z = \beta \cdot V_X \\ I_{WP} = \gamma_P \cdot I_X \\ I_{WN} = -\gamma_N \cdot I_X \end{array}\right\}, \tag{15}$$

$$s^2 + \frac{s}{C_2}\left(\frac{\beta_1\gamma_{N_1}}{R_2} - \alpha_2 \cdot g_3\right) + \frac{\alpha_1\beta_1\gamma_{P_1} \cdot g_1}{R_2C_1C_2} = 0, \tag{16}$$

$$C \cdot O \cdot \left(\frac{\beta_1\gamma_{N_1}}{R_2} - \alpha_2 \cdot g_3\right) = 0, \tag{17}$$

$$F \cdot O \cdot \omega_0' = \sqrt{\frac{\alpha_1\beta_1\gamma_{P_1} \cdot g_1}{R_2C_1C_2}}, \tag{18}$$

where α_1, β_1, γ_{P_1}, and γ_{N_1} are the non-ideal port transfer ratios of VDCC 1 and α_2, β_2, γ_{P_2}, and γ_{N_2} are the non-ideal port transfer ratios of VDCC 2. By considering the non-ideal behaviour of VDCC, the circuit of the active filter (as given in Fig. 4) is also interpreted. The obtained transfer functions for the LP and BP filters are shown in Eq. (19) and Eq. (20). The non-ideal angular frequency (ω_0') and non-ideal quality factor (Q_0') along with common denominator polynomial are represented by Eq. (21), Eq. (22), and Eq. (23), respectively.

$$\frac{I_{OUT}}{I_{IN}} = \frac{I_{LP}}{I_{IN}} = \frac{\alpha_1\beta_1\gamma_{P_1}\gamma_{N_2} \cdot \left(\frac{g_1}{R_2C_1C_2}\right)}{D'(s)}, \tag{19}$$

$$\frac{I_{OUT_1}}{I_{IN}} = \frac{I_{BP}}{I_{IN}} = \frac{\beta_1\gamma_{P_1}\gamma_{N_2} \cdot \left(\frac{s}{R_2C_2}\right)}{D(s)}, \tag{20}$$

$$\omega_0' = \sqrt{\frac{\alpha_1\beta_1\gamma_{P_1} \cdot g_1}{R_2C_1C_2}} = \omega_0\sqrt{\alpha_1\beta_1\gamma_{P_1}}, \tag{21}$$

$$Q_0' = \frac{1}{\gamma_{N_1}}\sqrt{\frac{\alpha_1\gamma_{P_1}R_2C_2 \cdot g_1}{\beta_1C_1}} = Q_0\left(\frac{1}{\gamma_{N_1}}\sqrt{\frac{\alpha_1\gamma_{P_1}}{\beta_1}}\right), \tag{22}$$

$$D'(s) = s^2 + s\left(\frac{1}{C_2}\right)\left(\frac{\beta_1\gamma_{N_1}}{R_2}\right) + \frac{\alpha_1\beta_1\gamma_{P_1} \cdot g_1}{R_2C_1C_2}. \tag{23}$$

The above-mentioned equations show that the ideal and non-ideal outputs (equations) are relatable, thereby justifying the designed idea. The sensitivity of both active and passive elements is considered for theoretical analysis. Here, the sensitivity analysis for ideal angular frequency (ω_0), non-ideal pole frequency (ω_0'), and quality factor (Q_0) and (Q_0' are scrutinised. Eq. (24) and Eq. (25) represent sensitivity calculation for ideal angular frequency (ω_0) and ideal quality factor (Q_0), whereas Eq. (26), Eq. (27), Eq. (28), and Eq. (29) represent non-ideal angular frequency (ω_0') and non-ideal quality factor (Q_0'). All these equations derived below show that the sensitivity in all the cases is under considerable limits.

$$S_{C_1}^{\omega_0} = S_{C_2}^{\omega_0} = S_{R_2}^{\omega_0} = -\frac{1}{2}; \quad S_{g_1}^{\omega_0} = \frac{1}{2}, \tag{24}$$

$$S_{R_2}^{Q_0} = S_{C_2}^{Q_0} = S_{g_1}^{Q_0} = \frac{1}{2}; \quad S_{C_1}^{Q_0} = -\frac{1}{2}, \tag{25}$$

$$S_{\alpha_1}^{\omega_0'} = S_{\beta_1}^{\omega_0'} = S_{\gamma_{P_1}}^{\omega_0'} = \frac{1}{2}, \tag{26}$$

$$S_{g_1}^{\omega_0'} = \frac{1}{2}; \quad S_{C_1}^{\omega_0'} = S_{C_2}^{\omega_0'} = S_{R_2}^{\omega_0'} = -\frac{1}{2}, \tag{27}$$

$$S_{C_2}^{Q_0'} = S_{R_2}^{Q_0'} = S_{g_1}^{Q_0'} = \frac{1}{2}; \quad S_{C_1}^{Q_0'} = -\frac{1}{2}, \tag{28}$$

$$S_{\alpha_1}^{Q_0'} = S_{\gamma_{P_1}}^{Q_0'} = \frac{1}{2}; \quad S_{\beta_1}^{Q_0'} = -\frac{1}{2}; \quad S_{\gamma_{N_1}}^{Q_0'} = -1. \tag{29}$$

The recognised parasitic model of VDCC [18] has been taken into consideration to contemplate the effects on the proposed circuits. When the VDCC model of Fig. 5 is reintegrated into the proposed circuit of Fig. 4 (only when Fig. 4 is acting as a sinusoidal oscillator), the resultant configuration resembles, as shown in Fig. 6. Some assumptions and simplifications have also been made before evaluating the circuit and are presented in Eq. (30). The characteristic polynomial is represented in Eq. (31). The CO and FO including

© 2021 ADVANCES IN ELECTRICAL AND ELECTRONIC ENGINEERING

device parasitics of the oscillator are shown in Eq. (32) and Eq. (33), respectively.

$$\left.\begin{array}{c} R_A = R_{P_1} \\ R_B = R_{Z_1}||R_{Z_2}||R_{N_1} \\ R_A >> R_{X_1} \\ C_A = C_1 + C_{P_1} \\ C_B = C_2 + C_{N_1} \end{array}\right\}, \tag{30}$$

$$\begin{aligned} &s^2 + s\Big(\frac{1}{C_A R_A} + \frac{1}{C_B R_B} - \frac{g_3}{C_B} + \frac{1}{C_B R_2}\Big) + \\ &+ \frac{1}{C_A C_B}\Big[\frac{1}{R_A}\Big(\frac{1}{R_B} + \frac{1}{R_2} - g_3\Big) + \frac{g_1}{R_2}\Big] = 0, \end{aligned} \tag{31}$$

$$CO: \Big(\frac{1}{C_A R_A} + \frac{1}{C_B R_B} - \frac{g_3}{C_B} + \frac{1}{C_B R_2}\Big) = 0, \tag{32}$$

$$FO: \tilde{\omega}_0 = \sqrt{\frac{1}{C_A C_B}\Big[\frac{1}{R_A}\Big(\frac{1}{R_B} + \frac{1}{R_2} - g_3\Big) + \frac{g_1}{R_2}\Big]}. \tag{33}$$

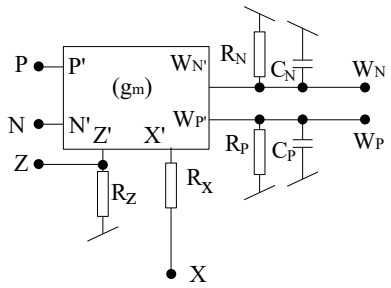


Fig. 5: Block diagram of VDCC including device parasitics [18].

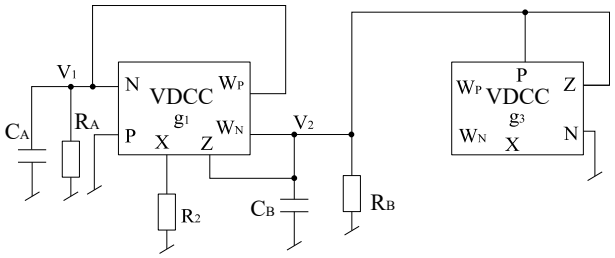


Fig. 6: Block diagram of electronically controlled oscillator including device.

When the derived circuit of a current-mode active filter (only when Fig. 4 is used as a universal filter) is reconfigured by replacing the ideal model of VDCC with the parasitic model of VDCC (from Fig. 5), the resultant structure looks as depicted in Fig. 7.

Before applying the regular circuit theory fundamentals to the derived circuit (Fig. 7), we have taken the following simplifications and assumptions, represented by Eq. (34), to conclude the outcome equations in a more suitable manner. The responses of LP and BP, including the effects of parasitic, are given in Eq. (35) and Eq. (36), respectively.

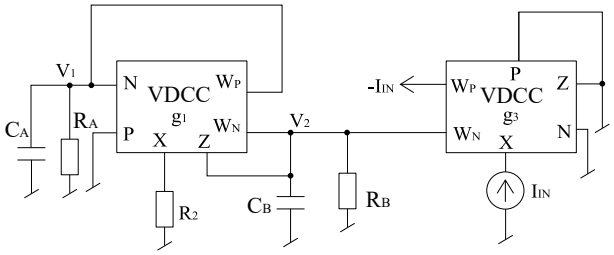


Fig. 7: Universal filter with device parasitics.

$$\left.\begin{array}{c} R_A = R_{P_1} \\ R_B = R_{Z_1}||R_{N_1}||R_{N_2} \\ R_2 >> R_{X_1} \\ C_A = C_1||C_{P_1} \\ C_B = C_2||C_{N_1}||C_{N_2} \end{array}\right\}, \tag{34}$$

$$\frac{I_{OUT}}{I_{IN}} = \frac{\frac{g_1}{R_2 C_A C_B}}{\tilde{\tilde{D}}(s)}, \tag{35}$$

$$\frac{I_{OUT_1}}{I_{IN}} = \frac{\Big[\frac{1}{R_A C_A} + s\Big] \cdot \frac{1}{R_2 C_B}}{\tilde{\tilde{D}}(s)}. \tag{36}$$

The denominator polynomial ($\tilde{\tilde{D}}(s)$), angular frequency ($\tilde{\tilde{\omega}}_0$), and quality factors ($\tilde{\tilde{Q}}_0$) are represented as Eq. (37), Eq. (38) and Eq. (39), respectively. It may be concluded from the above-given equations that both the derived circuits work well under the influence of the device parasitics.

$$\begin{aligned} \tilde{\tilde{D}}(s) = &s^2 + s\Big(\frac{1}{C_A R_A} + \frac{1}{C_B R_2} + \frac{1}{C_B R_B}\Big) + \\ &+ \frac{1}{C_A C_B}\Big[\frac{g_1}{R_2} + \frac{1}{R_A R_2} + \frac{1}{R_A R_B}\Big], \end{aligned} \tag{37}$$

$$\tilde{\tilde{\omega}}_0 = \sqrt{\frac{1}{C_A C_B}\Big[\frac{g_1}{R_2} + \frac{1}{R_A R_2} + \frac{1}{R_A R_B}\Big]}, \tag{38}$$

$$\begin{aligned} \tilde{\tilde{Q}}_0 = &\frac{R_2 R_A R_B}{C_B R_2 R_B + C_A R_A R_B + C_A R_A R_2} \cdot \\ &\cdot \sqrt{C_A C_B\Big[\frac{g_1}{R_2} + \frac{1}{R_A R_2} + \frac{1}{R_A R_B}\Big]}. \end{aligned} \tag{39}$$

It may be concluded from the above-given equations that both the derived circuits work well under the influence of the device parasitics.

5. Simulation Results

The designed circuits have been tested to verify the theoretical results using Cadence PSPICE simulation software. Figure 2 has been used for representing the VDCC block, made up of MOS transistors. In Fig. 2, the bias supply requirements are

±0.9 V voltage source for voltage biasing and the values of I_{B_1} and I_{B_2} are 50 µA and 100 µA respectively, to get the value of transconductance factor (g_m) as 277 µA·V^{-1}. The simulation results are bifurcated as mentioned. Firstly, the simulations of an electronically controllable sinusoidal oscillator are shown, where the results achieved through simulations for the active filter are represented later. Aspect ratios used for Fig. 2 are taken from [18] shown in Tab. 3. Here, 0.18 µm CMOS process parameters used for simulations are taken from [48].

Tab. 3: Aspect ratios of the MOS devices.

CMOS Transistors	W·L^{-1} (µm)
M_1–M_4	3.6/1.8
M_5–M_6, M_{4a}, M_{5a}, M_{6a}	7.2/1.8
M_7–M_8, M_{2a}, M_{7a}, M_{8a}	2.4/1.8
M_9–M_{10}	3.06/0.72
M_{11}–M_{12}	9/0.72
M_{13}–M_{16}	14.4/0.72
M_{17}	13.85/0.72
M_{18}–M_{22}	0.72/0.72

For the simulations of the designed quadrature oscillator, the passive components are selected as $C_1 = C_2 = 43$ pF, $R_2 = 3.60$ kΩ. The transconductance factors are $g_1 = g_3 = 277$ µA·V^{-1}. To show the working waveform of the designed oscillator, Fig. 8 has been presented. The Fast Fourier Transform (FFT) of the same waveform is shown in Fig. 9.

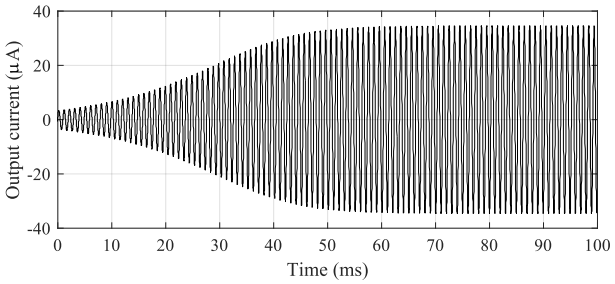


Fig. 8: Transient response for current-mode sinusoidal oscillator (I_{OUT}).

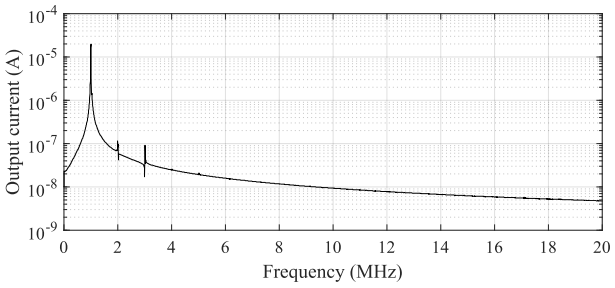


Fig. 9: Fast Fourier Transform of the I_{OUT}.

The theoretical value of the frequency of oscillation is chosen as 1.03 MHz. The simulated value of the frequency is achieved as 1.01 MHz, giving an error of 1.76 %. The deviation of oscillation frequency due to the effect of parasitics have also been calculated. The values of the parasitic elements have been taken from [19]. It is found that for the ideal frequency of 1.03 MHz, the computed oscillation frequency, under the impact of parasitics, is 1.0 MHz and gives an error of 2.72 %. The two distinct and explicit outputs, i.e., I_{OUT} and I_{OUT_1}, which must be theoretically phase-shifted by 90°, have been shown in Fig. 10. The calculated phase difference between the two signals in current mode is derived as 89.3°. The designed oscillator also provides voltage-mode quadrature outputs (the waveform of a few cycles only), as shown in Fig. 11.

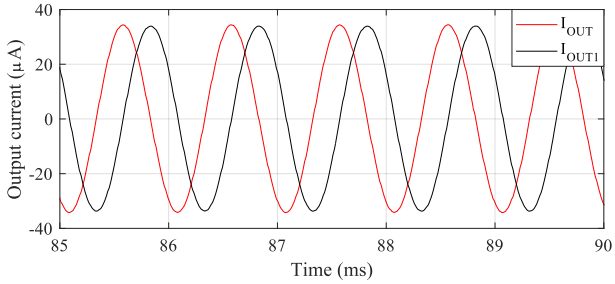


Fig. 10: Current-mode quadrature outputs.

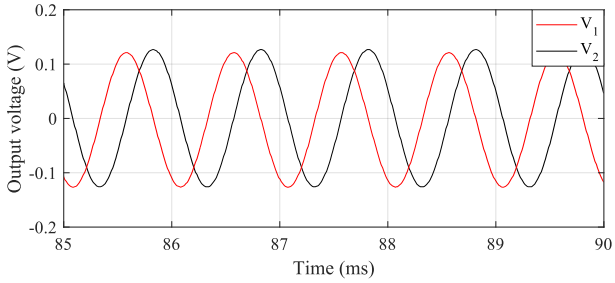


Fig. 11: Voltage-mode quadrature outputs.

The calculated angle difference between these two voltage signals is 89.2°. It can be seen from Eq. (4) that FO is directly dependent upon transconductance factor g_1. The same can be adjusted with the help of bias current, i.e., I_{B_1}, in Fig. 2 of the VDCC block. Thus, to check the wide span of the frequency of the oscillator, FO is varied with respect to the bias current of VDCC.

In Fig. 12, the value of I_{B_1} varies from 30 µA to 65 µA and the corresponding frequency ranges from 913 kHz to 1.08 MHz. If the bias current is further increased beyond this range, the active device gets saturated, increasing the percentage of error. The ideal and simulated values in Fig. 12 are closely related, thereby justifying the designed idea.

Theoretically, the designed circuit should deliver the constant output over the proposed range of frequencies. Unfortunately, while testing, the obtained responses did not match up with expectations. Variations in current outputs are shown in Fig. 13. Figure 14 shows the deviations in voltage mode quadrature outputs. It is evident from Fig. 13 and Fig. 14, that to get

stabilised output (s), either buffers or Automatic Gain Control (AGC) circuitry need to be deployed [49], [50], [51], and [52].

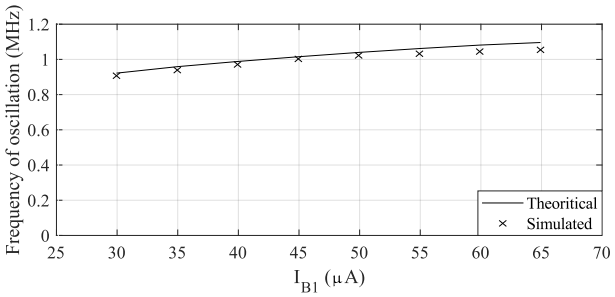


Fig. 12: Variation of frequency of operation with respect to bias current (I_{B_1}).

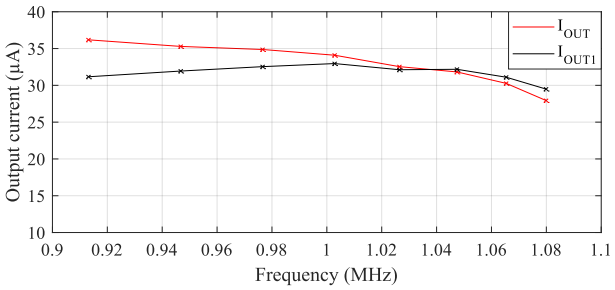


Fig. 13: Variation of current magnitude for sinusoidal oscillator.

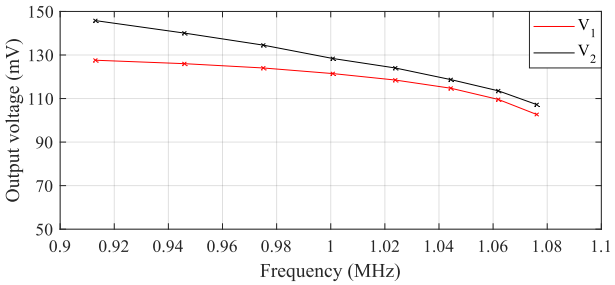


Fig. 14: Voltage variations with respect to oscillation frequency.

An already discussed approach for the utilization of AGC network [51] has been deployed in the designed oscillator. The output waveform at three different frequencies with almost the same magnitude is shown in Fig. 15, thus justifying the use of AGC network in the proposed oscillator.

Due to the brevity of the manuscript, the details of the AGC circuits are not discussed here and readers are encouraged to refer to the suggested references [49], [50], [51], and [52]. From Fig. 16, the phase dependency of the designed quadrature oscillator over a wide range of the frequency can be seen. To attain the constant phase difference between the two waveforms of quadrature oscillator, AGC circuit may be utilised. A graphical representation of oscillation frequency and corresponding Total Harmonic Distortion (THD) is represented in Fig. 17.

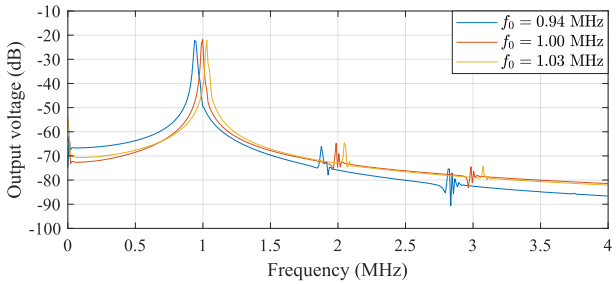


Fig. 15: Voltage variations with respect to oscillation frequency including AGC network.

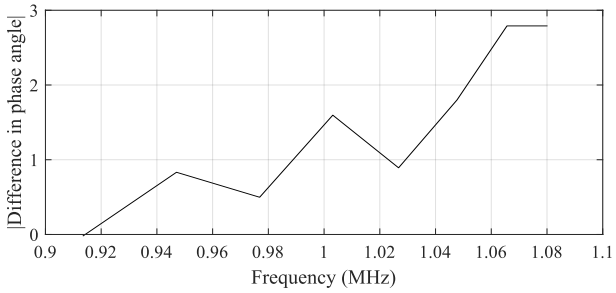


Fig. 16: Phase swing of a quadrature oscillator (current mode).

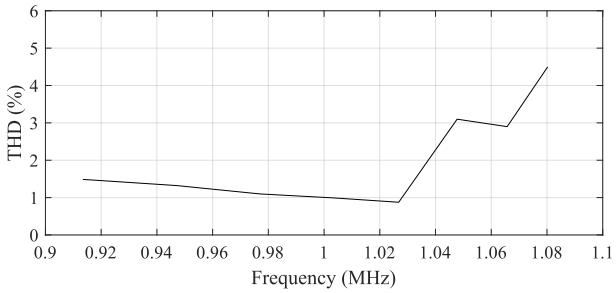


Fig. 17: THD values with respect to oscillation frequency.

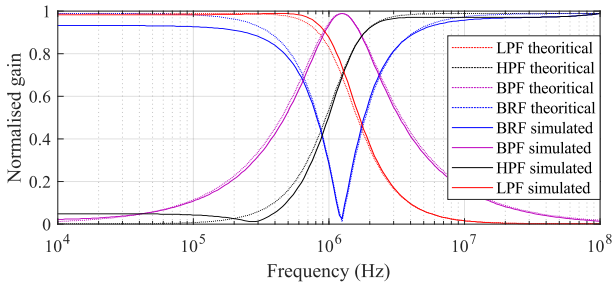


Fig. 18: Current-mode universal filter (LP, HP, BP and BR) responses.

To get the simulation results for a current-mode active filter, the values of the passive elements are $C_1 = C_2 = 50$ pF, $R_2 = 1.8$ kΩ, and $g_1 = g_3 = 277$ μA·V^{-1}. Figure 18 shows the ideal and practical frequency responses of LP, HP, BP, and BS filters. The calculated center frequency of the universal filter is 1.25 MHz,

whereas through simulation, 1.245 MHz was achieved giving an error of 0.4 % and therefore, supporting the design circuit. The following figure, i.e., Fig. 19, shows the gain as well as phase response of the AP filter. To see the fidelity of the filter circuit, transient analysis is performed on band pass filter. The response of the same is shown in Fig. 20. The frequency of the input sinusoidal signal is 1.25 MHz and the magnitude of the current is 50 µA. Figure 20 shows that the output is closely followed by the input.

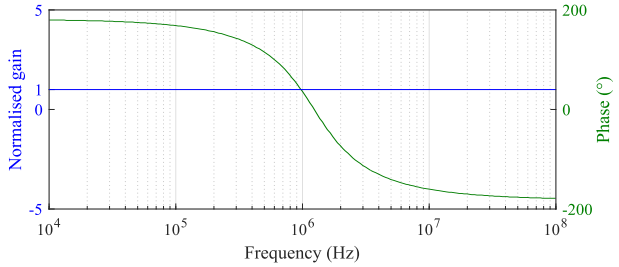


Fig. 19: All pass filter's gain and phase response.

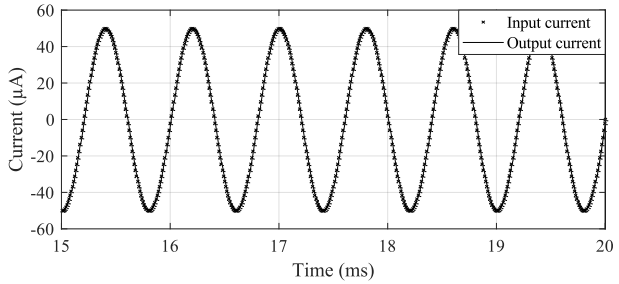


Fig. 20: Transient response of a BP filter.

The calculations for the THD of a BP filter have also been done, and its graphical representation is presented in Fig. 21. It is evident from Fig. 21 that the THD range of a BP filter is significantly low until the input magnitude of the current reaches a value of 120 µA.

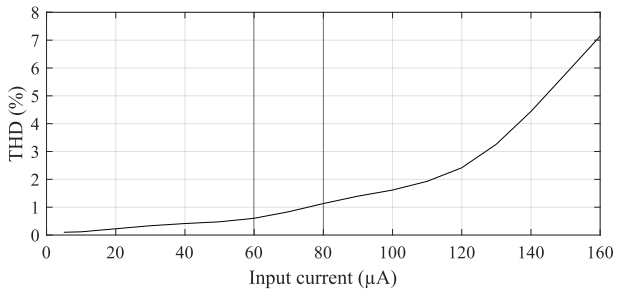


Fig. 21: A graphical representation of THD of a BP filter.

As discussed earlier, analogue circuits can find their applications in the area of control systems, communication engineering, instrumentation, and measurements [1]. In communication, Amplitude Modulation (AM) is one of the basic modulation techniques where the derived oscillator and designed filter circuit can occupy its space. The block diagram along with the waveforms' produced from the respective blocks have been shown in Fig. 22.

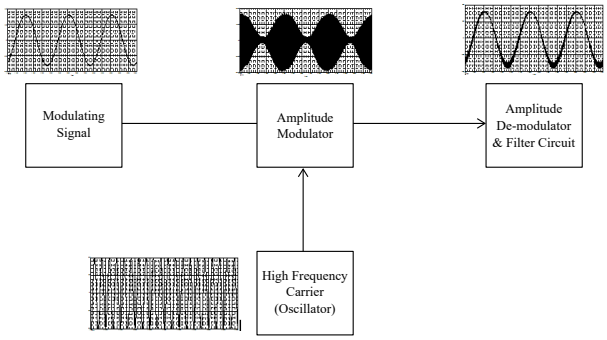


Fig. 22: Block diagram of Amplitude Modulation and Demodulation.

6. Experimental Results

There is no single commercially available Integrated Circuit (IC) that can successfully substitute the VDCC. Therefore, this active device is usually formulated from the combination of different off-the-shelf ICs. There are many approaches to realise this active building block such as (i) using LM13700 and AD844 [53], (ii) employing OPA860, (iii) using CA3080 and AD844. OPA860 is commonly known as the diamond transistor or special-purpose amplifier [54]. The approach utilised here is with the help of OPA860 due to the sterling features of this amplifier IC [54]:

- It is a special-function amplifier whose transconductance gain can be controlled with the help of single external resistor.
- The results obtained through OPA860 are close to the actual (derived) results.
- The bandwidth of this amplifier is very wide.

For the purpose of experiment, the transistor model of VDCC shown in Fig. 3 has been used along with some external circuitry. To explore the validity of the hardware arrangements, only the results of a sinusoidal oscillator are shown. It is worth noting that a resistor of 100 Ω, in series with the base of each diamond transistor and in series with input of the buffer of OPA860, is added [54]. For the realisation of hardware circuit, two Keysight programmable DC power supplies (E3632A) and a Caddo 30 MHz, 2-channel 4-trace Cathode Ray Oscilloscope (CRO) are used. The passive element values for the proper

© 2021 ADVANCES IN ELECTRICAL AND ELECTRONIC ENGINEERING

functionality of the sinusoidal oscillator were chosen as $C_1 = C_2 = 470$ pF, $R_{M1} = 1/g_{m1} = 330\ \Omega$, $R_{M3} = 1/g_{m3} = 0 - 1$ kΩ and $R_2 = 330\ \Omega$. The actual hardwired implementation for the sinusoidal oscillator is shown in Fig. 23.

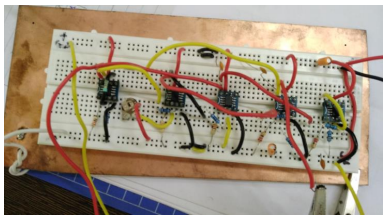


Fig. 23: Bread-board interconnection of the circuit.

The oscillator was ideally designed at 1 MHz and the corresponding simulations (using OPA860) were achieved at 940.2 kHz (giving an error of 5.8 %) and the hardware results were obtained at 910 kHz (giving an error of 9 % with respect to the ideal designed frequency). The complete range of the oscillator has also been tested for the designed hardware framework. Figure 24 shows the quadrature output, displayed on the CRO screen. Figure 25 shows the graph between the frequency of oscillation with respect to the variation in resistance value.

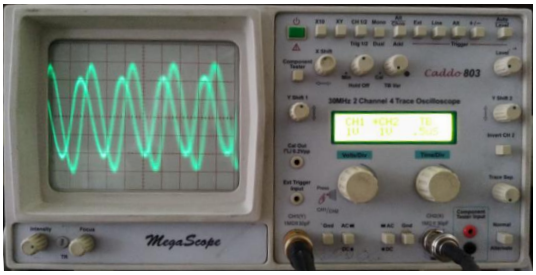


Fig. 24: Quadrature outputs.

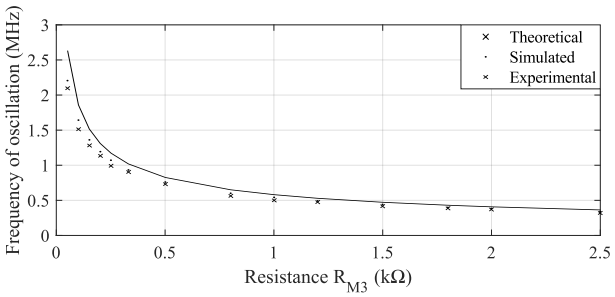


Fig. 25: FO with respect to RM3 (or g_3) using OPA 860.

7. Conclusion

The prime focus of this research article was to design a quadrature sinusoidal oscillator utilising only grounded passive elements and VDCC as an active element. The derived circuit offers quadrature outputs in current as well as voltage mode. The same circuit can be rearranged as an active filter by altering the position of the passive switches. The unique feature of the active filter is its impedance matching at the input and output port for making a higher-order current-mode filter. Regular mathematical calculations along with simulation results have also been included in the manuscript. At last, the experimental justification of the designed current-mode oscillator was also presented.

Acknowledgment

The experimental work was performed at Analog Electronics Lab of the department of Electronics Engineering, National Institute of Technology, Uttarakhand (Srinagar), Uttarakhand (India).

Author Contributions

M.G. conceived of the presented idea. T.S.A. developed the theory and S.N.G. performed the simulations and computations. T.S.A. and M.G. verified the analytical methods. T.S.A. carried out all the experimental work. T.S.A. and M.G. were in charge of overall direction and planning. T.S.A. wrote the manuscript with input from all authors. T.S.A., M.G. and S.N.G. discussed the results and contributed to the final manuscript.

References

[1] SENANI, R., D. R. BHASKAR, A. K. SINGH and V. K. SINGH. *Current Feedback Operational Amplifiers and Their Applications*. 1st ed. New York: Springer, 2013. ISBN 978-1-4614-5188-4.

[2] SENANI, R., D. R. BHASKAR, V. K. SINGH and R. K. SHARMA. *Sinusoidal Oscillators and Waveform Generators using Modern Electronic Circuit Building Blocks*. 1st ed. Cham: Springer, 2016. ISBN 978-3-319-23712-1.

[3] ARORA, T. S. and R. K. SHARMA. An All-Mode KHN Equivalent Biquad Using Third Generation Current Conveyor and All Grounded Passive Elements. *Proceedings of the National Academy of Sciences, India Section A: Physical Sciences*. 2017, vol. 87, iss. 1, pp. 97–108. ISSN 0369-8203, 2250-1762. DOI: 10.1007/s40010-016-0310-0.

[4] ARORA, T. S. and R. K. SHARMA. A Novel Cubic Generator Realised by CCIII-based Four

Quadrant Analog Multiplier and Divider. *Indian Journal of Science and Technology*. 2016, vol. 9, iss. 38, pp. 1–8. ISSN 0974-5645. DOI: 10.17485/ijst/2016/v9i38/100026.

[5] GUPTA, M. New Single Input Multiple Output Type Current Mode Biquad Filter Using OTAs. *Circuits and Systems*. 2016, vol. 7, iss. 4, pp. 231–238. ISSN 2153-1285. DOI: 10.4236/cs.2016.74022.

[6] SENANI, R., D. R. BHASKAR, M. GUPTA and A. K. SINGH. Canonic OTA-C Sinusoidal Oscillators: Generation of New Grounded-Capacitor Versions. *American Journal of Electrical and Electronic Engineering*. 2015, vol. 3, iss. 6, pp. 137–146. DOI: 10.12691/ajeee-3-6-2.

[7] ARORA, T. S. and S. GUPTA. A new voltage mode quadrature oscillator using grounded capacitors: An application of CDBA. *Engineering Science and Technology, an International Journal*. 2018, vol. 21, iss. 1, pp. 43–49. ISSN 2215-0986. DOI: 10.1016/j.jestch.2018.01.006.

[8] CHIEN, H.-C. Voltage- and Current-Modes Sinusoidal Oscillator Using a Single Differential Voltage Current Conveyor. *Journal of Applied science and Engineering*. 2013, vol. 16, iss. 4, pp. 395–404. ISSN 2708-9975. DOI: 10.6180/jase.2013.16.4.07.

[9] JIN, J. and C. WANG. Single CDTA-based current-mode quadrature oscillator. *AEU - International Journal of Electronics and Communications*. 2012, vol. 66, iss. 11, pp. 933–936. ISSN 1434-8411. DOI: 10.1016/j.aeue.2012.03.018.

[10] CHIEN, H.-C. and C.-Y. CHEN. CMOS realization of single-resistance-controlled and variable frequency dual-mode sinusoidal oscillators employing a single DVCCTA with all-grounded passive components. *Microelectronics Journal*. 2014, vol. 45, iss. 2, pp. 226–238. ISSN 0026-2692. DOI: 10.1016/j.mejo.2013.11.007.

[11] HORNG, J.-W., C.-L. HOU, C.-M. CHANG, W.-Y. CHUNG and H.-W. TANG. Current-Mode Multifunction Filters and Quadrature Oscillator Using Unit Gain Cells. *Journal of Active and Passive Electronic Devices*. 2006, vol. 1, iss. 1, pp. 247–258.

[12] KUMNGERN, M. and K. DEJHAN. DDCC-Based Quadrature Oscillator with Grounded Capacitors and Resistors. *Active and Passive Electronic Components*. 2009, vol. 2009, iss. 1, pp. 1–4. ISSN 1563-5031. DOI: 10.1155/2009/987304.

[13] TANGSRIRAT, W., P. MONGKOLWAI and T. PUKKALANUN. Current-mode high-Q bandpass filter and mixed-mode quadrature oscillator using ZC-CFTAs and grounded capacitors. *Indian Journal of Pure & Applied Physics*. 2012, vol. 50, iss. 1, pp. 600–607. ISSN 0975-1041.

[14] BIOLEK, D., R. SENANI, V. BIOLKOVA and Z. KOLKA. Active Elements for Analog Signal Processing: Classification, Review, and New Proposals. *Radioengineering*. 2008, vol. 17, iss. 4, pp. 15–32. ISSN 1210-2512.

[15] PRASAD, D., D. R. BHASKAR and M. SRIVASTAVA. New Single VDCC-based Explicit Current- Mode SRCO Employing All Grounded Passive Components. *Electronics*. 2014, vol. 18, iss. 2, pp. 81–88. ISSN 1450-5843. DOI: 10.7251/ELS1418081P.

[16] SOTNER, R., J. JERABEK, J. PETRZELA, N. HERENCSAR, R. PROKOP and K. VRBA. Second-Order Simple Multiphase Oscillator Using Z-Copy Controlled-Gain Voltage Differencing Current Conveyor. *Elektronika ir Elektrotechnika*. 2014, vol. 20, iss. 9, pp. 13–18. ISSN 2029-5731. DOI: 10.5755/j01.eee.20.9.8709.

[17] SOTNER, R., J. JERABEK, R. PROKOP and V. KLEDROWETZ. Simple CMOS voltage differencing current conveyor-based electronically tunable quadrature oscillator. *Electronics Letters*. 2016, vol. 52, iss. 12, pp. 1016–1018. ISSN 1350-911X. DOI: 10.1049/el.2016.0935.

[18] SRIVASTAVA, M. and D. PRASAD. VDCC Based Dual-Mode Quadrature Sinusoidal Oscillator with Outputs at Appropriate Impedance Levels. *Advances in Electrical and Electronic Engineering*. 2016, vol. 14, iss. 2, pp. 168–177. ISSN 1804-3119. DOI: 10.15598/aeee.v14i2.1611.

[19] GUPTA, M. and T. S. ARORA. Realization of Current Mode Universal Filter and a Dual-Mode Single Resistance Controlled Quadrature Oscillator Employing VDCC and Only Grounded Passive Elements. *Advances in Electrical and Electronic Engineering*. 2018, vol. 15, iss. 5, pp. 833–845. ISSN 1804-3119. DOI: 10.15598/aeee.v15i5.2397.

[20] MALIK, H., S. SRIVASTAVA, Y. R. SOOD and A. AHMAD. *Applications of Artificial Intelligence Techniques in Engineering*. 1st ed. Singapore: Springer, 2018. ISBN 978-981-13-1819-1.

[21] KACAR, F., A. YESIL and K. GURKAN. Design and experiment of VDCC-based voltage mode universal filter. *Indian Journal of Pure & Applied Physics*. 2015, vol. 53, iss. 5, pp. 341–349.

[22] KACAR, F., A. YESIL, S. MINAEI and H. KUNTMAN. Positive/negative lossy/lossless

grounded inductance simulators employing single VDCC and only two passive elements. *AEU - International Journal of Electronics and Communications*. 2014, vol. 68, iss. 1, pp. 73–78. ISSN 1434-8411. DOI: 10.1016/j.aeue.2013.08.020.

[23] ARORA, T. S., B. ROHIL and S. GUPTA. Fully Integrable/Cascadable CM Universal Filter and CM Quadrature Oscillator Using VDCC and Only Grounded Passive Elements. *Journal of Circuits, Systems and Computers*. 2019, vol. 28, iss. 11, pp. 1–36. ISSN 1793-6454. DOI: 10.1142/S0218126619501810.

[24] GUPTA, M., P. DOGRA and T. S. ARORA. Novel current mode universal filter and dual-mode quadrature oscillator using VDCC and all grounded passive elements. *Australian Journal of Electrical and Electronics Engineering*. 2019, vol. 16, iss. 4, pp. 220–236. ISSN 2205-362X. DOI: 10.1080/1448837X.2019.1648134.

[25] GUPTA, M. and T. S. ARORA. Various applications of analog signal processing employing voltage differencing current conveyor and only grounded passive elements: a re-convertible approach. *SN Applied Sciences*. 2020, vol. 2, iss. 9, pp. 1–18. ISSN 2523-3971. DOI: 10.1007/s42452-020-03379-6.

[26] ARORA, T. S. VDCC Based Sinusoidal Oscillators Using All Grounded Capacitors: A Series of Realization. *Wireless Personal Communications*. 2021, vol. 116, iss. 1, pp. 383–409. ISSN 1572-834X. DOI: 10.1007/s11277-020-07720-5.

[27] ARORA, T. S. A Current-Mode Single-Resistance-Controlled Oscillator Employing VDCC and All Grounded Passive Elements. *Iranian Journal of Electrical and Electronic Engineering*. 2020, vol. 16, iss. 2, pp. 184–191. DOI: 10.22068/IJEEE.16.2.184.

[28] BIOLEK, D., A. U. KESKIN and V. BIOLKOVA. Grounded capacitor current mode single resistance-controlled oscillator using single modified current differencing transconductance amplifier. *IET Circuits, Devices & Systems*. 2010, vol. 4, iss. 6, pp. 496. ISSN 1751-858X. DOI: 10.1049/iet-cds.2009.0330.

[29] LAHIRI, A. New realizations of voltage-mode quadrature oscillators using current differencing buffered amplifiers. *Journal of Circuits, Systems and Computers*. 2010, vol. 19, iss. 5, pp. 1069–1076. ISSN 1793-6454. DOI: 10.1142/S0218126610006608.

[30] HORNG, J.-W., Z.-R. WANG and T.-Y. YANG. Single ICCII Sinusoidal Oscillators Employing Grounded Capacitors. *Radioengineering*. 2011, vol. 20, iss. 3, pp. 608–613. ISSN 1805-9600.

[31] LAHIRI, A. Current-mode variable frequency quadrature sinusoidal oscillators using two CCs and four passive components including grounded capacitors. *Analog Integrated Circuits and Signal Processing*. 2012, vol. 71, iss. 2, pp. 303–311. ISSN 1573-1979. DOI: 10.1007/s10470-010-9571-8.

[32] SOTNER, R., J. JERABEK, N. HERENCSAR, Z. HRUBOS, T. DOSTAL and K. VRBA. Study of Adjustable Gains for Control of Oscillation Frequency and Oscillation Condition in 3R-2C Oscillator. *Radioengineering*. 2012, vol. 21, iss. 1, pp. 392–402. ISSN 1805-9600.

[33] PUSHKAR, K. L., D. R. BHASKAR and D. PRASAD. Single-Resistance-Controlled Sinusoidal Oscillator Using Single VD-DIBA. *Active and Passive Electronic Components*. 2013, vol. 2013, iss. 1, pp. 1–5. ISSN 1563-5031. DOI: 10.1155/2013/971936.

[34] MAHESHWARI, S. Voltage-Mode Four-Phase Sinusoidal Generator and Its Useful Extensions. *Active and Passive Electronic Components*. 2013, vol. 2013, iss. 1, pp. 1–8. ISSN 1563-5031. DOI: 10.1155/2013/685939.

[35] BHASKAR, D. R., D. PRASAD and K. L. PUSHKAR. Fully Uncoupled Electronically Controllable Sinusoidal Oscillator Employing VD-DIBAs. *Circuits and Systems*. 2013, vol. 4, iss. 3, pp. 264–268. ISSN 2153-1293. DOI: 10.4236/cs.2013.43035.

[36] JIN, J. and C. WANG. CDTA-based electronically tunable current-mode quadrature oscillator. *International Journal of Electronics*. 2014, vol. 101, iss. 8, pp. 1086–1095. ISSN 1362-3060. DOI: 10.1080/00207217.2013.805394.

[37] YUCEL, F. and E. YUCE. CCII based more tunable voltage-mode all-pass filters and their quadrature oscillator applications. *AEU - International Journal of Electronics and Communications*. 2014, vol. 68, iss. 1, pp. 1–9. ISSN 1434-8411. DOI: 10.1016/j.aeue.2013.06.012.

[38] HORNG, J.-W., C.-L. HOU, C.-M. CHANG, S.-T. CHENG and H.-Y. SU. Current Or/And Voltage-Mode Quadrature Oscillators with Grounded Capacitors and Resistors using FD-CCIIs. *WSEAS Transactions on Circuits and Systems*. 2008, vol. 7, iss. 3, pp. 129–138. ISSN 1109-2734.

[39] LAHIRI, A., W. JAIKLA and M. SIRIPRUCHYANUN. Voltage-mode quadrature sinusoidal oscillator with current tunable properties. *Analog Integrated Circuits and Signal Processing*. 2010, vol. 65, iss. 2, pp. 321–325. ISSN 1573-1979. DOI: 10.1007/s10470-010-9488-2.

[40] SOTNER, R., Z. HRUBOS, J. SLEZAK and T. DOSTAL. Simply Adjustable Sinusoidal Oscillator Based on Negative Three-Port Current Conveyors. *Radioengineering*. 2010, vol. 19, iss. 3, pp. 446–454. ISSN 1805-9600.

[41] MAHESHWARI, S. and B. CHATURVEDI. High output impedance CMQOs using DVCCs and grounded components. *International Journal of Circuit Theory and Applications*. 2011, vol. 39, iss. 4, pp. 427–435. ISSN 0098-9886. DOI: 10.1002/cta.636.

[42] ABDALLA, K. K., D. R. BHASKAR and R. SENANI. Configuration for realising a current-mode universal filter and dual-mode quadrature single resistor controlled oscillator. *IET Circuits, Devices & Systems*. 2012, vol. 6, iss. 3, pp. 1–9. ISSN 1751-8598. DOI: 10.1049/iet-cds.2011.0160.

[43] LAHIRI, A. Deriving (MO)(I)CCCII Based Second-order Sinusoidal Oscillators with Non-interactive Tuning Laws using State Variable Method. *Radioengineering*. 2011, vol. 20, iss. 1, pp. 349–353. ISSN 1805-9600.

[44] JIN, J. Current-Mode Resistorless Simo Universal Filter And Four-Phase Quadrature Oscillator. 2014, vol. 7, iss. 2, pp. 1–6. DOI: 10.5281/ZENODO.1335982.

[45] LI, Y. A. Electronically tunable current-mode biquadratic filter and four-phase quadrature oscillator. *Microelectronics Journal*. 2014, vol. 45, iss. 3, pp. 330–335. ISSN 0026-2692. DOI: 10.1016/j.mejo.2013.12.005.

[46] SUMMART, S., C. THONGSOPA and W. JAIKLA. New current-controlled current-mode sinusoidal quadrature oscillators using CDTAs. *AEU - International Journal of Electronics and Communications*. 2015, vol. 69, iss. 1, pp. 62–68. ISSN 1434-8411. DOI: 10.1016/j.aeue.2014.07.011.

[47] KUMNGERN, M., P. LAMUN and K. DEJHAN. Current-mode quadrature oscillator using current differencing transconductance amplifiers. *International Journal of Electronics*. 2012, vol. 99, iss. 7, pp. 971–986. ISSN 1362-3060. DOI: 10.1080/00207217.2011.651693.

[48] MINAEI, S. and E. YUCE. Novel Voltage-Mode All-Pass Filter Based on Using DVCCs. *Circuits, Systems and Signal Processing*. 2010, vol. 29, iss. 3, pp. 391–402. ISSN 1531-5878. DOI: 10.1007/s00034-010-9150-3.

[49] SOTNER, R., J. JERABEK, R. PROKOP and K. VRBA. Current Gain Controlled CCTA and its Application in Quadrature Oscillator and Direct Frequency Modulator. *Radioengineering*. 2011, vol. 20, iss. 1, pp. 317–326. ISSN 1805-9600.

[50] SOTNER, R., Z. HRUBOS, N. HERENCSAR, J. JERABEK, T. DOSTAL and K. VRBA. Precise Electronically Adjustable Oscillator Suitable for Quadrature Signal Generation Employing Active Elements with Current and Voltage Gain Control. *Circuits, Systems, and Signal Processing*. 2014, vol. 33, iss. 1, pp. 1–35. ISSN 1531-5878. DOI: 10.1007/s00034-013-9623-2.

[51] SOTNER, R., J. JERABEK, N. HERENCSAR, J. PETRZELA, K. VRBA and Z. KINCL. Linearly tunable quadrature oscillator derived from LC Colpitts structure using voltage differencing transconductance amplifier and adjustable current amplifier. *Analog Integrated Circuits and Signal Processing*. 2014, vol. 81, iss. 1, pp. 121–136. ISSN 1573-1979. DOI: 10.1007/s10470-014-0353-6.

[52] SOTNER, R., J. JERABEK, L. LANGHAMMER, J. POLAK, N. HERENCSAR, R. PROKOP, J. PETRZELA and W. JAIKLA. Comparison of Two Solutions of Quadrature Oscillators With Linear Control of Frequency of Oscillation Employing Modern Commercially Available Devices. *Circuits, Systems, and Signal Processing*. 2015, vol. 34, iss. 11, pp. 3449–3469. ISSN 1531-5878. DOI: 10.1007/s00034-015-0015-7.

[53] SATIPAR, D., P. INTANI and W. JAIKLA. Electronically Tunable Quadrature Sinusoidal Oscillator with Equal Output Amplitudes during Frequency Tuning Process. *Journal of Electrical and Computer Engineering*. 2017, vol. 2017, iss. 1, pp. 1–10. ISSN 2090-0155. DOI: 10.1155/2017/8575743.

[54] Wide Bandwidth Operational Transconductance Amplifier (OTA) and Buffer. In: *Texas Instruments* [online]. 2008. Available at: `https://www.ti.com/lit/ds/symlink/opa860.pdf?ts=1628152082082&ref_url=https%253A%252F%252Fwww.google.com%252`.

About Authors

Tajinder Singh ARORA (Senior Member, IEEE) was born 1979 in Dehradun, Uttarakhand, India. He has completed his B.Tech. (Electronics and Communication Engineering) from C.C.S.University in 2002, M.Tech. (Signal Processing) from Netaji Subhas Institute of Technology, New Delhi in 2008, Ph.D. (Analog Signal Processing) from Uttarakhand Technical University, Dehradun in 2017. He is currently working as an Assistant Professor with Department of Electronics Engineering, National Institute of Technology, Uttarakhand, India. His research interests are in the area of current mode signal processing, electronic circuit design, electronic devices etc. He has authored or co-authored 25 research papers in International Journals and Conferences.

Manish GUPTA was born 1975 in Ghaziabad, Uttar Pradesh, India. He obtained B.Tech Degree in Electronics and Communication Engineering from Delhi College of Engineering, Delhi, (presently D.T.U. Delhi), M.Tech with specialization in Signal Processing from Netaji Subhas Institute of Technology, New Delhi and Ph.D. in area of Analog Integrated Circuits and Signal Processing from Jamia Millia Islamia, New Delhi, India. Currently he is working as an Associate Professor with Department of Electronics and Communication Engineering, Inderprastha Engineering College, Ghaziabad, Uttar Pradesh, India. His research interests are in the area of Bi-polar and CMOS, analog integrated circuits, current-mode signal processing.

Shiv Narain GUPTA was born 1982 in Jaunpur, Uttar Pradesh, India. He has completed his B.Tech degree in Electronics and Communication Engineering from Uttar Pradesh Technical University in 2006 and M.Tech degree in Digital Communication from Ambedkar Institute of Technology, Delhi in 2009. He is pursuing Ph.D. from Jaypee Institute of Information Technology (JIIT), Noida. His research areas are analog integrated circuits and analog signal processing applications.

2021 8th International Conference on Signal Processing and Integrated Networks (SPIN) | 978-1-6654-3564-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SPIN52536.2021.9566133

An Improved Fusion Technique for IR and VIS Image using LWT and PCA

Urvashi Rawat
Dept.of Electronics and Communication
Delhi Technological University
Delhi, India
urvashidpsgv99@gmail.com

Sudipta Majumdar
Dept.of Electronics and Communication
Delhi Technological University
Delhi, India
korsudipta@rediffmail.com

***Abstract*— Image fusion techniques are divided into two broad categories: spatial domain and transform domain. Principal component analysis (PCA) is a spatial domain technique which is computationally simpler and reduces redundant information but has the demerit of spectral degradation. Lifting wavelet transform (LWT) is a transform domain technique which has an adaptive design and demands less memory. In this paper, a new hybrid fusion algorithm has been introduced which combines the LWT and PCA in a parallel manner. These two fusion methods are applied on Infrared and Visible images. The hybrid method is then compared with conventional fusion techniques like PCA, LWT and DWT. It has been shown that the proposed method outperforms conventional methods. The results are analyzed using performance parameters standard deviation, average value, the average difference, and normalized cross-correlation.**

***Keywords—Image fusion, PCA, LWT, performance criteria,* hybrid method.**

I. INTRODUCTION

Image Fusion[1] represents a method for enhancing the data performance from a given image collection. With help of image fusion, the better/relevant data from individual given input image is fused into a resulting image of a greater quality than the original images. A number of operations are carried out for the given input images, which will enhance the prominent information in both of the images. The output image is then obtained by combining the data from the given input images. There are a variety of applications for image fusion such as fusing medical images like magnetic resonance imaging (MRI)[2], electro-encephalographic (EEG), computed tomography(CT)[2], positron emission tomography(PET)), electrocardiographic imaging (ECG), in surveillance[3], object tracking, face recognition[5] and in remote sensing[4] etc.

The image fusion can be widely divided into transform domain and spatial domain[1]. The spatial domain approaches include fusion methodologies such as the Brovey method, averaging, select minimum/maximum, principal component analysis (PCA) and IHS forms of methods. In these methods, the fusion algorithm is applied directly on the pixels of the input images. They are simple in terms of computation but the resulting image suffers distortion and lacks spatial information. The transform domain techniques on the other hand divide the input image to lower and higher frequency components thus providing us with details of the image and line/edge information. Some of the transform domain techniques are Discrete wavelet transform (DWT), stationary wavelet transform (SWT), lifting wavelet transform (LWT), complex contourlet transform(CCT), discrete cosine transform(DCT), non-subsampled shearlet transform (NSST) and nonsubsampled contourlet transform NSCT etc.

In this paper, Infrared (IR) and Visible (VIS) [3][12] images have been selected for the fusion purpose. Although there are many types of images available for fusion, IR and VIS images have complementary aspects which when fused together, yield a better image. The IR image captures the thermal radiation of the object and hence the target detection ability of this image is commendable. VIS image captures the light reflected from the target, i.e in normal visibility conditions, it would provide us with a detailed image. Due to these reasons, motivation for finding a new algorithm which would enhance the quality of the output fused image arises.

PCA is a spatial domain technique in which the fusion technique is directly applied on the pixels of the image. It is computationally simple and free from color distortions but suffers from spectral degradation. LWT technique on the other hand is transform domain technique in which the fusion method is applied on the high and low-frequency components obtained after decomposition. We have discussed a new fusion algorithm in this paper for IR and VIS images using LWT-PCA in order to combine the merits of both the techniques and yield better results in terms of performance parameters. Fusion of IR and VIS image using this technique would give us a robust image having the properties of both.

The rest part of the paper is consisting of the following sections. An introduction to image fusion and its types, applications is given in section I. Section II contains the literature review of some of the techniques which have been implemented recently. Then section III discusses about the theories related to LWT and PCA with their advantages and disadvantages. The proposed work is explained in section IV and the results of the experiments performed are specified in section V. The paper is concluded in section VI along with future scope of the method.

II. LITERATURE SURVEY

In 2016, K. Padmavathi, M. V. Karki and M. Bhat[6] developed a method for fusing the medical images with PCA and complex wavelet transform(CWT). It combined the advantages of shift-invariance of CWT and feature enhancement of PCA. The proposed algorithm proved to be better in representing the spatial details and soft tissues of the tumor. Analyzing the method in terms of Peak signal-to-noise ratio (PSNR), fusion factor and entropy, it was found that the proposed method outperformed other methods of fusion.

In 2016, the authors Saranya G and S. N. Devi [7] used several methods like DWT, PCA, SWT, NSCT and CCT for fusion of medical images. The result of the methods were analyzed using mutual information (MI) and it was found that CCT method had more MI than the other methods. It was suggested that the proposed algorithm could be used in future for diagnosis purposes in medical field.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:36:21 UTC from IEEE Xplore. Restrictions apply.

In 2016, a new method based on combination of framelet transform, PCA and DCT was introduced by P. Sharma, P. Kulkarni, H. Sanghvi and A. Verma[8]. The method inherited the merits of DCT, PCA and framelet domain. The proposed method was compared with some conventional techniques and it was concluded that, improved results were achieved with regard to mean square error (MSE), PSNR and visual look analysis.

In 2017, A. A. Mergin and M. S. G. Premi[9] introduced a hybrid method for medical signal fusion. The EEG and ECG signals were fused and then reconstructed using hybrid DCT-DWT method. This method proved to be better that the individual DWT and DCT in terms of PSNR which indicated a achievement of better signal reconstruction.

In 2017, authors B. L. Priya and K. Jayanthi[10] presented a technique for fusion which enhanced the edge information of CT image of the Abdomen. The method was based on NSST which gave better results in terms of segmentation accuracy of the liver. The proposed method showed improvement segmentation accuracy of 60.11% in terms of similarity index

In 2018, M. B. Abdulkareem[11] performed fusion of MRI and PET images using the haar wavelet approach. The proposed algorithm was tested on brain images of Alzheimer's disease and the input images were decomposed at four levels of high and low frequency regions. The fused image's quality was tested using the PSNR and MSE. This method gave 90-95% accuracy.

In 2018, the authors S. M. Nemalidinne, A. P. Sindhu and D. Gupta[12] presented a fusion method for IR and VIS fire images using NSCT which was based on pulse coupled neural network and Gabour energy. The proposed method was better in terms of high visual quality and high entropy than the conventional methods. The NSCT method has shift-invariance and provides with multi-scale decomposition.

In 2019, N. Taxak and S. Singhal[13] described a method for fusion using the Weighted Average and Brovey Transform. The fused image had good clarity, increased PSNR value and decreased MSE value than the other fusion techniques like PCA, DWT, DCT.

In 2019, A. Yehia, H. Elhifnawy and M. Safy[14] introduced a method for fusion of multi sensor images integrating IHS and SWT transform method. Initially SWT of the input image was taken which preserved the spectral information and then it was combined with IHS algorithm. This method gave favorable results concerning spectral and spatial information. Two different types of fusion rules were chosen which are: average method and window-based fusion.

In 2019, V. Brindha. and P. Jayashree [15] introduced a fusion method for the Glioblastoma Multiforme images. The fusion methodology included the weighted average method and maximum selection of the enhanced DWT. The method was demonstrated to be better than some of the existing techniques in terms of fusion quality.

In 2019, J. R. Benjamin and T. Jayasree[16] introduced a methodology for fusion of multimodal brain images. The authors used NSST method and used the Maximum rule as fusion method. Two data sets of MRI and PET images were used. The proposed method proved to be better in terms of edge details, correlation and information preservation.

In 2020, M. Sayadi, H. Ghassemian, R. Naimi and M. Imani [17] introduced a fusion method in which the NSST was used as an edge detector, In order to match the luminance channel of MRI image with PET image, IHS transform was applied on PET image A gaussian filter was also applied in order to get the low resolution intensities. The final fused image was obtained by integrating the two images and then taking inverse HIS transform. This method proved to be better in terms of preserving the spatial information and having less spectral distortions than the other methods.

In 2020, T. Tyagi, P. Gupta and P. Singh[18] introduced a novel method which combined the PCA and SWT techniques in a parallel manner to yield better results for multi focus images. The hybrid methodology gave satisfactory results than the conventional methods with respect to PSNR and correlation coefficient.

In 2021, Kaur, H., Koundal, D. & Kadyan [19] presented a survey on image fusion techniques. Various spatial and transform domain techniques were discussed with their merits and demerits. Apart from this, different applications of image fusion like remote sensing, medical image, surveillance etc. were also discussed. At last, evaluation metrics were also discussed for evaluation of the fusion techniques.

III. Related Theories

A. Principal Component Analysis

PCA is a feature extraction methodology also known as Karhunen-Loève transform. Primarily the task of this method is to transform a correlated set of data into an orthogonal set of variables called as principal components [20]. Linear combinations of the observed weighted variables are coined as the principal components. The number of extracted components in the PCA is equal to the number of variables observed. Fig.1 shows the flow diagram of PCA. After arranging the original images in two columns, they are subtracted from their empirical means. The vector resulting from above method will be a n x 2 matrix with n being the each vector's length of the image. Then the eigen vectors and eigen values are calculated for the above-mentioned vector. Eigen vectors are obtained which corresponds to the larger eigen value. L1 and L2 are the normalized components that are obtained from the resulting eigenvector. The equation of the fused image can be written as follows:

$$If(p,q)= L1I1(p,q)+L2\ I2(p,q) \quad (1)$$

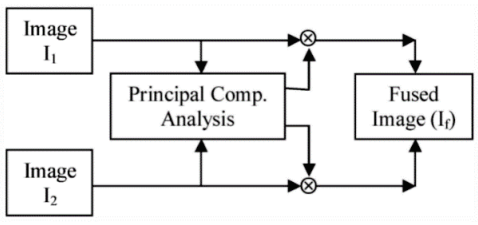


Fig.1: Flow diagram of Principal Component Analysis

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:36:21 UTC from IEEE Xplore. Restrictions apply.

The PCA algorithm can be understood with the following steps:

Step 1: First we arrange the details into the column vectors.
Step 2: Next, the empirical mean 'Me' is calculated along each of the columns.
Step 3: Subtract Me (Me vector) from column of data matrix.
Step 4: Calculate the matrix of covariance 'C'.
Step 5: Calculate the eigenvectors and values 'V', 'D' respectively of matrix 'C' and sort them from higher to lower eigenvalues.
Step 6: Lastly, For computation of L1 and L2, we consider first column of eigen vector V corresponding to the larger eigenvalue

$$L1 = \frac{V(1)}{\sum V} \text{ and } L2 = \frac{V(2)}{\sum V} \quad (2)$$

The fused image contains even more information compared to the original input images. PCA works in the spatial domain and produces high-quality images that are free of color distortion. Also for the analysis, it reduces multidimensional data to a lower dimension. However, it suffers from the problem of spectral degradation.

B. *Lifting Wavelet Transform*

The LWT-based fusion is a way to decompose the input image into four components of high and low frequency. The lower frequency components are called approximation coefficients and the higher frequency coefficients are horizontal, vertical, and diagonal coefficients[21]. This decomposition method can be applied as many times. After the decomposition is done, the coefficients of original input images are merged by using an appropriate rule for fusion. The new fused coefficients also contain four coefficients. After that, Inverse LWT (ILWT) is applied to get the final fused image. Fig.2 and Fig.3 shows the block diagram of LWT. It can be divided into 3 steps which are Split, Predict and Update[22].

(1) *Split:* The original image is split into two mutually disjoint sets S1, D1. S1 represents low frequency components and D1 represents high frequency components.

(2) *Predict:* To remove repetitive data, the subset S1 is used to predict subset D1. This is done to reduce computation and storage.

(3) *Update:* After prediction, it might be possible that coefficient produced by S1 might not align with the original source image. Therefore updation is required to maintain the original characteristics of the data.

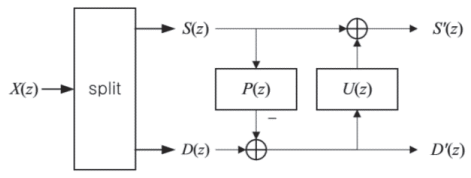


Fig.2: Lifting operation

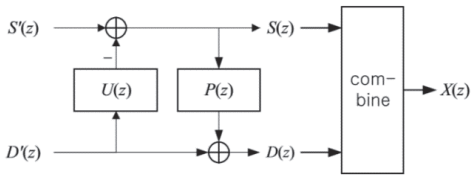


Fig.3: Inverse lifting operation

LWT is a different way of calculating wavelet coefficients. This scheme demands less computations and memory and also easily generates wavelet transform for the lossless compression.

IV. Proposed Work

In this proposed work, the two input images are processed in a parallel manner using two different type of fusion methods as shown in the block diagram Fig.4 . The input images are IR and VIS image as shown in Fig. 5 (a),(b) respectively. The steps of the proposed methodology are discussed below:

Step 1: LWT scheme is first applied on the input images which results in level-1 of decomposition. After this step, approximate coefficients LL1 and detailed coefficients LH1,HL1, HH1 are obtained.
Step 2: Level-2 decomposition is done again on the approximate coefficients LL1 to obtain LL2, LH2, HL2, HH2.
Step 3: The detailed part of the LWT contains maximum information, so it is important to get best of the two images. Therefore maximum rule is applied .
Step 4: The approximate part also holds some information but not much, therefore average rule is applied to them.
Step 5: The original IR and VIS images are also fused using the PCA based fusion method in parallel.
Step 6: Finally, both the fused images of LWT and PCA are combined to get the final output fused image.

V. Experimental Analysis and Performance Parameters

The proposed work was implemented in MATLAB 2017a and the dataset used was TNO data set for IR and VIS images. For quantitative analysis of the proposed method, various performance criteria are selected which are Standard Deviation (SD), Average Value (AV), Average Difference (AD) and Normalized cross correlation (NCC). The description of these performance parameters are provided in the Table I.

TABLE I. Description of Performance Parameters

Performance Parameters	Description
Standard Deviation[23]	Degree of dispersion between average pixel and single pixel of the image is given by SD.
Average Value[23]	Average brightness of the fused output image is described by AV.
Average Difference[24]	Difference between the output fused image and the original image is termed as AD.
Normalized cross correlation[24]	NCC provides similarities of output fused image with input image.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:36:21 UTC from IEEE Xplore. Restrictions apply.

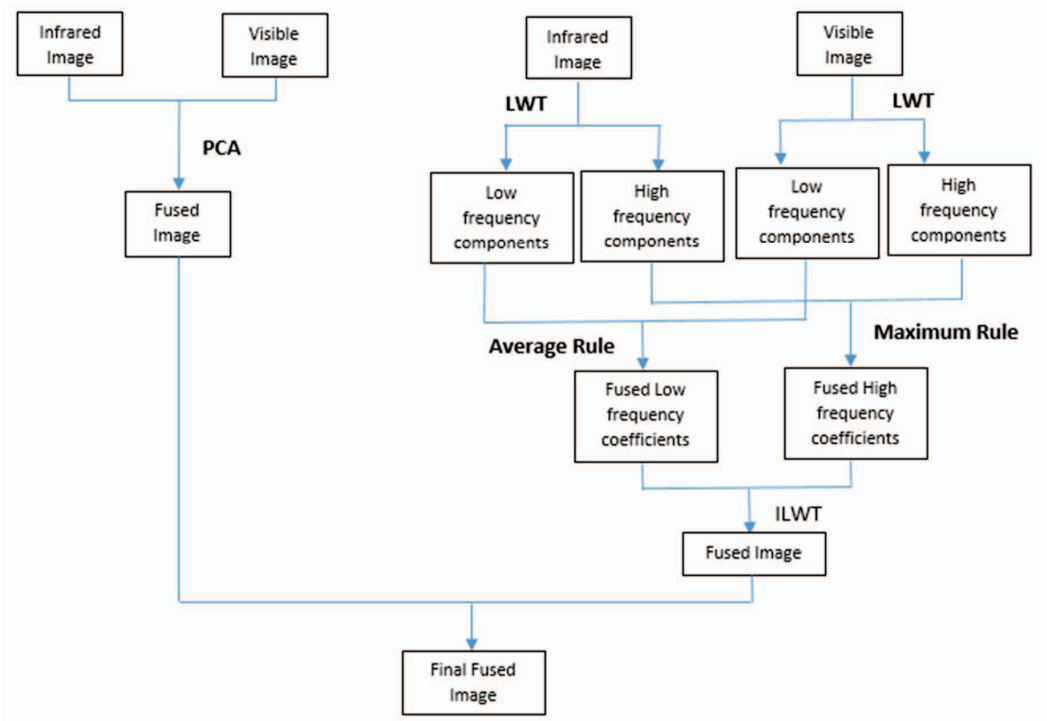


Fig.4: Block diagram of proposed work

(i) Standard Deviation (SD)

SD gives the degree of dispersion between average pixel and a single pixel of the image. Larger value of SD means high contrast, more information and wide gray value distribution. The formula for the SD is given below.

$$SD = \sqrt{\frac{1}{M*N}\sum_{I=1}^{M} \quad \sum_{i=1}^{N}(p(i,j)-\mu)^2} \qquad (3)$$

where P(i, j) is gray value and μ is mean value of pixel at (i, j).

(ii) Average Value (AV)

The average brightness of the output fused image is defined by the parameter AV. A larger value of this parameter means brighter image is obtained. The formula for AV is given below.

$$\mu = \frac{1}{M*N}\sum_{i=1}^{M}\sum_{j=1}^{N} p(i,j) \qquad (4)$$

where P(i, j) represents value of the pixel of fused image where N is the height and M is the width of the image.

(iii) Average Difference (AD)

AD represents the difference in the fused image and the original input image . This parameter must give lower values. The formula of AD is given below:

$$AD = \frac{1}{pq}\sum_{i=1}^{p}\sum_{i=1}^{q} |Aij - Bij| \qquad (5)$$

(iv) Normalized Cross-correlation (NCC)

NCC provides similarities of output fused image with the original image. NCC should have a high value. The formula is given below:

$$\frac{\sum_{i=1}^{p}\sum_{i=1}^{q}(Aij*Bij)}{\sum_{i=1}^{p}\sum_{i=1}^{q}(Aij)^2} A = \pi r^2 \qquad (6)$$

Fig.5 shows the output fused images using the new modified method. It has also been compared to other conventional methods such as DWT, PCA and LWT. The result of these methods are listed in Table II.

TABLE II. QUANTITATIVE ANALYSIS

Fusion Methods	SD	AV	AD	NCC
PCA	48.96	81.33	3.05	1.01
DWT	22.77	91.09	6.70	0.97
LWT	23.50	91.13	-6.74	0.97
Proposed method	67.22	163.14	-88.07	1.98

From the experimental analysis, it was found that the proposed method had a higher value of SD, NCC and AV i.e 67.22, 1.98 and 163.14 respectively which is greater than the conventional fusion methods used for comparison. Also, the performance of the novel method was analyzed

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:36:21 UTC from IEEE Xplore. Restrictions apply.

using AD performance criteria and it was found that this value came out to be less in case of the proposed method, which proves to be better since the difference between fused image and input image should be as less as possible.

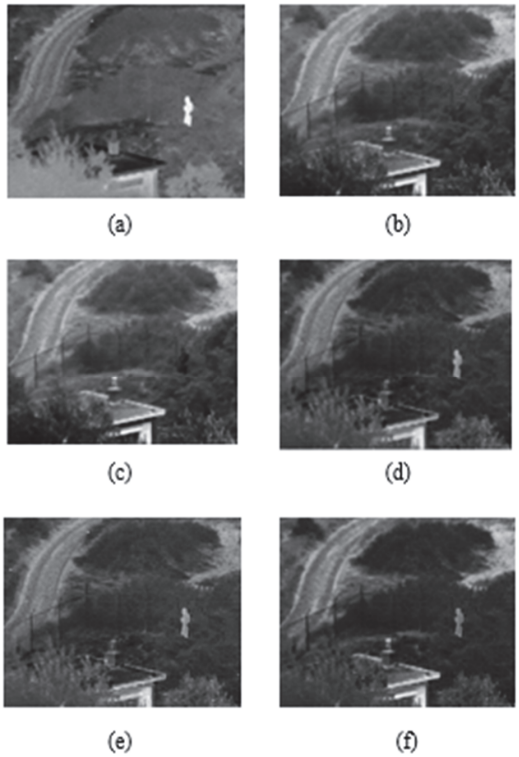


Fig. 5: Input images (a) IR, (b) VIS and output fused images using different techniques. (c) PCA, (d) DWT, (e) LWT, (f) Proposed method

VI. Conclusion

In this paper, a parallel fusion method for IR and VIS image dataset was introduced using LWT and PCA. At first, the LWT decomposition was applied to the input images. The approximate coefficients were fused using the average method and the detailed coefficients were fused using the maximum selection rule, ILWT was applied after that. The fusion of input images was also done using the PCA method in parallel. The method included fusion of the input images in a parallel manner. The final fused images obtained from both the methods were fused together to get the final output image. The image obtained from the proposed method was compared with PCA, DWT, and LWT using AD, SD, AV, and NCC and it was found that the new method gave better results in these parameters. The future work focuses on extending this fusion algorithm to medical images.

References

[1] [1] Li, Shutao, et al. “Pixel-level image fusion: A survey of the state of the art.” Information Fusion 33 (2017)

[2] [2] Alex Pappachen James, and Belur V. Dasarathy. “Medical image fusion: A survey of the state of the art.” Information Fusion 19 (2014): 4-19

[3] [3] P. Kumar, A. Mittal, P. Kumar, “Fusion of thermal infrared and visible spectrum video for robust surveillance,” in: Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing, 2006, pp. 528–539

[4] [4] G. Simone, A. Farina, F. C. Morabito, S. B. Serpico, L. Bruzzone, “Image fusion techniques for remote sensing applications,” Information Fusion 3 (1) (2002) 3–15

[5] [5] J. Han, B. Bhanu, “Fusion of color and infrared video for moving human detection,” Pattern Recognition 40 (6) (2007) 1771–1784.

[6] [6] K. Padmavathi, M. V. Karki and M. Bhat, “Medical image fusion of different modalities using dual tree complex wavelet transform with PCA,” 2016 International Conference on Circuits, Controls, Communications and Computing (I4C), 2016, pp. 1-5, doi: 10.1109/CIMCA.2016.8053288.

[7] [7] Saranya G and S. N. Devi, “Performance evaluation for image fusion technique in medical images using spatial and transform method,” 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2016, pp. 446-450, doi: 10.1109/WiSPNET.2016.7566173.

[8] [8] P. Sharma, P. Kulkarni, H. Sanghvi and A. Verma, “Image fusion based enhancement in framelet domain,” 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), 2016, pp. 316-319, doi: 10.1109/PDGC.2016.7913167.

[9] [9] A. A. Mergin and M. S. G. Premi, “Pixel level fusion of medical signals using DCT, DWT and hybrid(DWT-DCT) transform based on maximum selection rule-a comparison,” 2017 International Conference on Computing Methodologies and Communication (ICCMC), 2017, pp. 898-903, doi: 10.1109/ICCMC.2017.8282596.

[10] [10] B. L. Priya and K. Jayanthi, “Edge enhancement of liver CT images using non subsampled shearlet transform based multislice fusion,” 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2017, pp. 191-195, doi: 10.1109/WiSPNET.2017.8299746.

[11] [11] M. B. Abdulkareem, “Design and Development of Multimodal Medical Image Fusion using Discrete Wavelet Transform,” 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 1629-1633, doi: 10.1109/ICICCT.2018.8472997.

[12] [12] S. M. Nemalidinne, A. P. Sindhu and D. Gupta, “Nonsubsampled Contourlet Domain Fusion Approach for Infrared and Visible Fire Images,” TENCON 2018 - 2018 IEEE Region 10 Conference, 2018, pp. 2516-2521, doi: 10.1109/TENCON.2018.8650310.

[13] [13] N. Taxak and S. Singhal, “High PSNR based Image Fusion by Weighted Average Brovery Transform Method,” 2019 Devices for Integrated Circuit (DevIC), 2019, pp. 451-455, doi: 10.1109/DEVIC.2019.8783400.

[14] [14] A. Yehia, H. Elhifnawy and M. Safy, “An Improved Integrated Intensity-Hue-Saturation with Stationary Wavelet Transform Multi-sensor Image Fusion Approach,” 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), 2019, pp. 153-159, doi: 10.1109/ICICIS46948.2019.9014669.

[15] [15] V. Brindha. and P. Jayashree., “Fusion of radiological images of Glioblastoma Multiforme using weighted average and maximum selection method,” 2019 11th International Conference on Advanced Computing (ICoAC), 2019, pp. 328-332, doi: 10.1109/ICoAC48765.2019.246861.

[16] [16] J. R. Benjamin and T. Jayasree, “An Efficient MRI-PET Medical Image Fusion using Non-Subsampled Shearlet Transform,” 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 2019, pp. 1-5, doi: 10.1109/INCOS45849.2019.8951329.

[17] [17] M. Sayadi, H. Ghassemian, R. Naimi and M. Imani, “A New Composite Multimodality Image Fusion Method Based on Shearlet Transform and Retina Inspired Model,” 2020 International Conference on Machine Vision and Image Processing (MVIP), 2020, pp. 1-5, doi: 10.1109/MVIP49855.2020.9116919.

[18] [18] T. Tyagi, P. Gupta and P. Singh, "A Hybrid Multi-focus Image Fusion Technique using SWT and PCA," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2020, pp. 491-497, doi: 10.1109/Confluence47617.2020.9057960.

[19] [19] Kaur, H., Koundal, D. & Kadyan, V., “Image Fusion Techniques: A Survey,”. Arch Computat Methods Eng (2021). https://doi.org/10.1007/s11831-021-09540-7

[20] [20] Wenkao Yang, Jing Wang, Jing Guo, “A Novel Algorithm for Satellite Images Fusion Based on Compressed Sensing and

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:36:21 UTC from IEEE Xplore. Restrictions apply.

PCA", Mathematical Problems in Engineering, vol. 2013, ArticleID 708985, 10 pages, 2013. https://doi.org/10.1155/2013/708985

[21] [21] Kannan K, Perumal SA, Arulmozhi K (2010) "Performance comparison of various levels of fusion of multi-focused images using wavelet transform," Int J Comput Appl 1(6):71–78

[22] [22] X. Xiao and Z. Wu, "Image Fusion Based on Lifting Wavelet Transform," 2010 International Symposium on Intelligence Information Processing and Trusted Computing, 2010, pp. 659-662, doi: 10.1109/IPTC.2010.26

[23] [23] Varsha Patil1 , Deepali Sale2 and M.A.Joshi, "Image Fusion Methods and Quality Assessment Parameters," Asian Journal of Engineering and Applied Technology ISSN 2249-068X Vol. 2 No. 1, 2013, pp.40-46

[24] [24] Junwu, L., Li, B., & Jiang, Y. (2020). "An Infrared and Visible Image Fusion Algorithm Based on LSWT-NSST". IEEE Access, 8, 179857-179880

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:36:21 UTC from IEEE Xplore. Restrictions apply.

The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/1741-0398.htm

Analysis of barriers intensity for investment in big data analytics for sustainable manufacturing operations in post-COVID-19 pandemic era

Received 29 March 2021
Revised 19 July 2021
8 September 2021
Accepted 11 September 2021

Narender Kumar and Girish Kumar
Mechanical Engineering, Delhi Technological University, Delhi, India, and
Rajesh Kr Singh
Management Development Institute, Gurgaon, India

Abstract

Purpose – The study presents various barriers to adopt big data analytics (BDA) for sustainable manufacturing operations (SMOs) post-coronavirus disease (COVID-19) pandemics. In this study, 17 barriers are identified through extensive literature review and experts' opinions for investing in BDA implementation. A questionnaire-based survey is conducted to collect responses from experts. The identified barriers are grouped into three categories with the help of factor analysis. These are organizational barriers, data management barriers and human barriers. For the quantification of barriers, the graph theory matrix approach (GTMA) is applied.

Design/methodology/approach – The study presents various barriers to adopt BDA for the SMOs post-COVID-19 pandemic. In this study, 17 barriers are identified through extensive literature review and experts' opinions for investing in BDA implementation. A questionnaire-based survey is conducted to collect responses from experts. The identified barriers are grouped into three categories with the help of factor analysis. These are organizational barriers, data management barriers and human barriers. For the quantification of barriers, the GTMA is applied.

Findings – The study identifies barriers to investment in BDA implementation. It categorizes the barriers based on factor analysis and computes the intensity for each category of a barrier for BDA investment for SMOs. It is observed that the organizational barriers have the highest intensity whereas the human barriers have the smallest intensity.

Practical implications – This study may help organizations to take strategic decisions for investing in BDA applications for achieving one of the sustainable development goals. Organizations should prioritize their efforts first to counter the barriers under the category of organizational barriers followed by barriers in data management and human barriers.

Originality/value – The novelty of this paper is that barriers to BDA investment for SMOs in the context of Indian manufacturing organizations have been analyzed. The findings of the study will assist the professionals and practitioners in formulating policies based on the actual nature and intensity of the barriers.

Keywords Big data analytics, Investment, Sustainable operations, BDA barriers, Factor analysis, Graph theory matrix analysis, Pandemic, Sustainable development goals

Paper type Research paper

1. Introduction

In the present business environment of a circular economy, all organizations are trying to ensure sustainability in all operations across the whole supply chain for being globally competitive. Manufacturing organizations are shifting toward sustainability by reducing costs and waste as it is helping to improve productivity (Jabbour *et al.*, 2020; Roy and Roy, 2019). Sustainability has been applied to various stages of manufacturing operations such as planning, design, prototype, design for manufacturing, manufacturing, post-manufacturing, etc. (Saxena *et al.*, 2020; Machado *et al.*, 2020). Recently, awareness of sustainability has



Journal of Enterprise Information Management
© Emerald Publishing Limited
1741-0398
DOI 10.1108/JEIM-03-2021-0154

increased among manufacturers. There exists a relationship between manufacturing operations and the natural environment, and this fact is utilized in real-time decision-making (Kamble *et al.*, 2020a, b; Raut *et al.*, 2019). Sustainable manufacturing and consumption are considered as a strategy made for many manufacturing companies. Through this strategy, the organization can meet the overall development plans which include a reduction in resource utilization and pollution. This further promotes the organizations to save for future generations and elevates the energy usage while minimizing the cost at the same time (Roy and Singh, 2017; Kumar *et al.*, 2021a, b).

The current situation is increasingly uncertain and precarious in the business. Therefore, realizing how to adjust and change has become a critical challenge for the industry to achieve sustainability goals. The adoption of advanced technologies and digitization in the manufacturing sector is generating a high volume of data referred as big data. Generally, big data follows the 3V's High Volume (a large amount of data), high velocity (the high speed or the frequency with which the data are generated) and high variability (different types of data). But many authors have also added another two V named "Value" and "Veracity" (Manyika *et al.*, 2011). Value explains the economic importance of these data and how they are helping organizations to grow. Veracity or verification relates to the unreliable and invaluable data verification and to ensure quality data is being considered for the study.

A normal supply chain generates more than 100 gigabytes of data (Gupta *et al.*, 2019). There is a rapid increase in the number of sensors and radio-frequency identification (RFID) tags. The number of sensors used for the components of the supply chain in transportation, retail or automobile has increased at the rate of 30% each year (Manyika *et al.*, 2011). The RFID tags have also increased to 209 billion units by 2021 (Tachizawa *et al.*, 2015). This increase in sensor-based technology has helped organizations to reduce their operational cost by 10–25% (Hahn and Packowski, 2015).

Big data analytics (BDA) is considered a novel technology, which requires advanced technical skills in the era of data evolution (Guo *et al.*, 2017). BDA helps to create value from the data for the organizations. But only data collection and storing do not add value, the value is added only when the data is processed and analyzed to provide results that help in decision making.

The use of emerging technologies such as BDA, machine learning, blockchain, etc. may help in the containment of coronavirus disease (COVID-19) transmission by using data and information on a real-time basis. The COVID-19 has taught the importance of building resilience to many organizations as many companies have faced huge losses during this pandemic situation (PWC, 2020). Due to this many organizations are working with resilience models and strategies and BDA is one of the technologies, which helps organizations in building resilience (Bag *et al.*, 2020; Gunasekaran *et al.*, 2017; Saengchai and Jermsittiparsert, 2019). Maroufkhani *et al.* (2020) recognized BDA as the turning point for organizations to improve their performance. BDA provides researchers across developed and emerging economies with substantial and significant opportunities to research potential job patterns, global crises, and their effect on innovation and information sharing, by digital and analytical capabilities for the sustainability of supply chains (Sheng *et al.*, 2020). Innovations help the organizations to increase productivity, which further helps the organization to focus more on innovative sustainable practices such as eco-designing and eco-labeling of supply chains (Bhanot *et al.*, 2017). Therefore, there are many opportunities to make the use of big data more impactful in situations like the COVID-19 outbreak.

The COVID-19 pandemic has raised awareness of the main social and economic roles of manufacturing, especially concerning the supply of medical equipment, medication and other goods needed in the health-care sector (Devansh, 2020). At the same time, it has created an unparalleled financial hit for the manufacturing sector, cost reduction, and the development of transparent and flexible supply chains. Manufacturing companies now need to work

together more than ever to apply BDA to their manufacturing operations to unlock value and achieve sustainable growth. To be more flexible in reacting to the crisis, or when markets return to normal, the analytical upgrade will radically boost marketing strategies and innovation capabilities (Wang *et al.*, 2020). The use of emerging technologies will theoretically improve consumer loyalty by providing support for e-retailing, sales, customer care and so on (Davenport *et al.*, 2020).

BDA is recognized as a unique data processing tool that enhances the organizations' performance (Chen *et al.*, 2015). The emergence and development of BDA have created much insightful and diverse information for many organizations through the analysis of data (Guo *et al.*, 2018). To enhance firm performance, BDA evokes two major perspectives. These are data collection and data usage. First, the collection of big data from internal or external sources be in the form of an unstructured, semi-structured and structured data set (Frank *et al.*, 2019). Second, the use of big data with appropriate analytics helps in sustainable manufacturing operations (SMOs) for better decision making and hence improving the firm performance (Chams and Garcia-Blandon, 2019; Gunasekaran *et al.*, 2017). Acioli *et al.* (2021) highlighted the importance of Industry 4.0 technologies such as BDA for a sustainable supply chain during the COVID-19 pandemic. Exploiting the capability of BDA is essential for firms to mitigate the impact of global crises. COVID-19 has motivated organizations to shift to real-time analytics tools to look for accurate and reliable information to better understand the impact on their operations (Kent, 2020).

Several studies suggest that companies with high BDA capabilities achieve improved sustainable performance (Kamble *et al.*, 2020a, b; Belhadi *et al.*, 2019). BDA helps organizations to reduce cost (Choi *et al.*, 2018; Aydiner *et al.*, 2019; Dubey *et al.*, 2019a, b), faster production (Giannakis and Louis, 2016; Dubey *et al.*, 2018), resource optimization (Banyai *et al.*, 2019; Pinto *et al.*, 2020; Rachinger *et al.*, 2019), process improvement (Ciulli *et al.*, 2020; Julianelli *et al.*, 2020), optimize the production process for the first mile/last mile delivery along with innovative technological product development (Banyai *et al.*, 2019; Belinski *et al.*, 2020), cost reduction and improving the employee work effectiveness (Rachinger *et al.*, 2019), creation of new products and services which further helps the organizations to understand customers changing need and to timely meet them (Choi *et al.*, 2018; Ghasemaghaei and Calic, 2019).

The firms which exploit emerging technologies like BDA and artificial intelligence to improve their decision-making ability achieve better operational performance (Dubey *et al.*, 2020). Moreover, investments in BDA lead to advantages in terms of the financial performance of a firm.

In view of the future growth of manufacturing and the limited availability of natural resources, there is a need for the manufacturing industry to invest in new technologies such as BDA, to improve sustainability in operations. BDA, at the plant and process level, has shown positive outcomes in terms of performance improvement and energy-saving methods (Nascimento *et al.*, 2019; Yacob *et al.*, 2019). This technology helps in the reduction of losses associated with the manufacturing processes, provides accurate forecasting, reduces information distortion and delivers better utilization of equipment (Gandhi *et al.*, 2018).

In spite of the significant application of BDA for sustainable manufacturing operations, organizations hesitate in making investments in emerging technologies such as BDA (Kumar *et al.*, 2019). Limited research has been done on identifying major challenges in making investments for BDA applications in the context of sustainable manufacturing (Kumar *et al.*, 2021a). There have been many studies that have investigated various issues of sustainable operations. Adrian *et al.* (2017) investigated the factors influencing the success of BDA implementation for effective decision making toward organizational performance enhancement. Hilbert (2016) proposed a conceptual framework related to threats and opportunities for using BDA for international development. However, none of these studies have explicitly identified barriers to BDA investment for SMOs in the manufacturing sector.

Considering the important role of BDA in ensuring sustainable manufacturing operations, this study investigates various barriers to investment in BDA applications. The main objectives of this study are as follows:

(1) To identify different barriers to the investment on BDA for sustainable manufacturing operations.

(2) To categorize these barriers from a managerial perspective.

(3) To quantify different categories of barriers from a strategic point of view.

To achieve the research objectives, the work is conducted in three phases. The first phase deals with the identification of barriers to BDA adoption based on a literature review and addresses the first research objective. To find the number of categories of barriers, descriptive analysis is carried out in the second phase to achieve the second objective. Finally, the third phase is implemented for evaluation of the intensity of barriers using the graph theory matrix approach (GTMA). The findings of the study will contribute significantly to analyzing barriers to making an investment in BDA applications in manufacturing.

This study mainly addresses the following research questions:

(1) What are the key barriers to investment on BDA for SMOs?

(2) What is the impact of the barriers in adopting BDA in SMOs?

To find answers to these questions, different barriers to the investment on BDA are determined. Factor analysis and graph-theoretic matrix approaches are applied for the categorization and quantification of all barriers, respectively.

The rest of the paper is structured as follows. Section 2 presents a literature review for the identification of barriers to the investment of BDA in manufacturing. In Section 3, the authors focus on research methodology, while Section 4 covers results and discussion. Finally, Section 5 presents the conclusion, limitations and managerial implications.

2. Literature review

This section covers the literature on barriers to make an investment on BDA for sustainability in manufacturing operations.

2.1 BDA benefits for sustainable manufacturing

Sustainable manufacturing operations is a strategy for manufacturing industries that helps manufacturers in terms of reduction in the use of resources and low pollution levels for the entire lifecycle (Roy and Singh, 2017; Singh *et al.*, 2020). The Industry 4.0 technologies can be easily integrated with the business processes for sustainability. Integration results in cost reduction, sustainable process, timely information sharing, improving efficiency, flexibility, quality and collaboration (Ali and Gölgeci, 2019; Machado *et al.*, 2020; Pham *et al.*, 2019) and promotes the betterment of the operations pace (Lechler *et al.*, 2019). Acioli *et al.* (2021) investigated the effects of Industry 4.0 technologies in the COVID-19 sustainable supply chains. They emphasize the effect of Industry 4.0 technologies in the supply chains in the COVID-19 pandemic situation.

Ye *et al.* (2021) considered BDA as an advanced tool for sustainable manufacturing. To achieve sustainable development, BDA addresses the grand challenges which hinder the sustained efforts (Eisenhardt *et al.*, 2016; George *et al.*, 2016). Belaud *et al.* (2019) highlighted the importance of big data at different levels of sustainability, and Wu *et al.* (2017) encored that sustainability is an essential element for the business models of new technologies like BDA (Martínez-Olvera and Mora-Vargas, 2019). The BDA tools are being adopted for SMOs

to minimize production risks, market losses and process flaws. The utilization of BDA tools in SMOs also increases business effectiveness (Sharma *et al.*, 2020). BDA helps the organizations in profiling their customers (Ardito *et al.*, 2019) which further helps the organizations in satisfying their customers (Zhou *et al.*, 2019). BDA has helped organizations to increase their return on investment by 15–20%, enhance productivity and provide a competitive advantage (Ding, 2018; Manyika *et al.*, 2011; Zhou *et al.*, 2019). The innovations in the industries promoting sustainability practices have helped the organizations to consume less energy, reduce wastage and improve the organization's brand value (Ding, 2018). Through BDA, the data transparency in the entire supply chain has increased and has led to the prevention of goods being damaged (Birkel *et al.*, 2019; Buntak *et al.*, 2019; Junge, 2019). With the improvement in transparency, the errors are minimized and losses incurred are reduced (Birkel *et al.*, 2019; Gabriel and Pessel, 2016; Stock and Seliger, 2016).

Investment in BDA improves firm performance in terms of better economic, environmental and social indicators. Manufacturing industries in the developed world are implementing BDA for SMOs (Dubey *et al.*, 2016). Dubey *et al.* (2015) have identified management, governing pressures, supplier relationship management, employee's engagement, reconfigurable manufacturing systems and lean manufacturing as pillars for world-class sustainable manufacturing. Thakur and Mangla (2019) investigated the change for sustainability considering operational human-technological aspects in leading Indian industries in the home appliances sector. Akhtar *et al.* (2019) found a significant positive relationship between big-data savvy teams, big-data-driven actions and organization performance. Yasmin *et al.* (2020) identified that BDA capabilities have a positive impact on organizational performance. Scavarda *et al.* (2020) have also highlighted the importance of integrating Industry 4.0 technologies with circular economy practices.

Liu *et al.* (2020) explored the contribution of advanced technologies (such as Internet of Things, Cyber-Physical System, Artificial Intelligence, BDA, Cloud Computing, etc.) for the development of SMOs. Gawankar *et al.* (2020) found valuable insights for retail supply chain practitioners on planning BDA investments. A big data-driven supply chain has a strong influence on the supply chain performance measures in the retail supply chain with the contractual-based alliance.

Aho (2015) argued that big data can develop and transform organizational cultures. According to him "The transformative potential of Big Data lies in the treating data as an asset". Integration of big data from multiple sources has helped the organization to optimize its supply chain. For example, Walmart integrates millions of data from multiple sources, every hour into a single system (Sanders, 2016). Sanders (2016) observed that the application of BDA is not limited to price optimization, route optimization, inventory optimization, micro-segmentation of marketing or labor scheduling but has extended to many other areas of supply chain management. The application of BDA is further extended to boost the after-sales performance of service parts management.

Most organizations fail to extract value from the data. With the development of BDA, several new technologies are developed that helped organizations to extract meaningful information from the data (Marr, 2021). It includes Artificial Intelligence, Quantum Computing, Edge Computing, Natural Language Processing, and Hybrid cloud are some of the recent developments. Therefore, the development in the field of BDA has been a gamechanger for analytics (Gill, 2021).

BDA provides potential benefits both at the strategic level and the operational level for sustainable manufacturing. At the strategic end, they help in sustainable sourcing, supply chain networking and product designing. At the operational end, they help in improving visibility by providing real-time data, improve flexibility and help to manage the volatility and cost fluctuations during manufacturing (Mangla *et al.*, 2020). Table 1 summarizes the benefits of BDA for sustainable manufacturing.

Table 1. Benefits of BDA for sustainable manufacturing

Benefits of BDA applications	References
Decrease in operational cost and improvement in quality	Choi *et al.* (2018), Aydiner *et al.* (2019), Dubey *et al.* (2019a, b), Ren *et al.* (2019), Luo *et al.* (2017), Rachinger *et al.* (2019), Machado *et al.* (2020), Ali (2019), Pham *et al.* (2019), Ozkan-Ozen *et al.* (2020)
Real-time decision making	Machado *et al.* (2020), Ren *et al.* (2019), Inamdar *et al.* (2020), Kumar *et al.* (2021a, b)
Improved business transformation	Maroufkhani *et al.* (2020), Ren *et al.* (2019), Sharma *et al.* (2017), Arunachalam *et al.* (2018), Giannakis and Louis (2016), Dubey *et al.* (2018)
Better product/service quality	Luo *et al.* (2017), Shibin *et al.* (2017a), Choi *et al.* (2018), Ghasemaghaei and Calic (2019), Kazancoglu *et al.* (2021)
Minimization of resources/energy waste	Ren *et al.* (2019), Shibin *et al.* (2017a), Pinto *et al.* (2020), Rachinger *et al.* (2019), Banyai *et al.* (2019)
Improve safety and prevent risks	Luo *et al.* (2017), Raut *et al.* (2019)
Reduce or eliminate emissions from industrial processing	Ren *et al.* (2019), Shibin *et al.* (2017b)
Improvement in economic and environmental sustainability	Ren *et al.* (2019), Luo *et al.* (2017), Bag *et al.* (2020), Kamble *et al.* (2020a, b), Belaud *et al.* (2019)
Effective decision-making process	Machado *et al.* (2020), Dubey *et al.* (2019a, b), Arunachalam *et al.* (2018)

2.2 Barriers for investment in BDA

In the age of Industry 4.0, manufacturing organizations have begun to adopt BDA to optimize their decisions for SMOs (Manyika *et al.*, 2011). However, organizations are facing a number of challenges in making investments for BDA applications. These investment challenges are due to the limited knowledge of how to use BDA for sustainable manufacturing operations. Although BDA adoption requires high investments, but it is fast changing and offering new opportunities for data handling (Schull and Maslan, 2018). Alharthi *et al.* (2017) presented a qualitative analysis of barriers in using BDA. The barriers to the implementation of BDA for SMOs are categorized as, organizational barriers, data management barriers and human barriers. Organizational barriers have been one of the important barriers that influence BDA implementation for SMOs (Lamba and Singh, 2018; Sun *et al.*, 2018; Moktadir *et al.*, 2019). The important organizational barriers include security, privacy, digital infrastructure, organizational policies etc (Amerioun *et al.*, 2018; Mishra and Rane, 2019; Sivarajah *et al.*, 2017). A suitable organizational culture will help BDA implementation dramatically and reduce the risk associated with the BDA process (Diaz *et al.*, 2018; Sun *et al.*, 2018). Lack of management support and lack of digital vision hinder the success of BDA implementation from the organizational viewpoint. Management can promote and encourage employees to use the implementation of big data by designing incentive systems and linking them to the use of big data (Watson, 2019). Human barriers include people's awareness and skills of the employees in the context of BDA (Dubey *et al.*, 2019a, b).

Manufacturing industries face obstacles in managing BDA due to a lack of IT experts. Data management barriers are implicit complications of BDA because of the data development, broadening of data sources, different formats and the unstructured data that makes it difficult to deal with, store, and, regain the data (Alharthi *et al.*, 2017). The barriers need to be investigated for the successful implementation of BDA to minimize risks, improve productivity, enhance quality etc. Organizations that have the knowledge of their current state of BDA capabilities are in a better state to overcome the challenges associated with BDA. Maturity models can be used by the organizations to understand their present state of technology and to further benchmark themselves with the industry standards. Manufacturing industries are increasingly turning to BDA to get insights and make decisions based on big data. However, there are many barriers to the adoption of BDA in sustainable manufacturing operations (Moktadir *et al.*, 2019). These barriers should be

investigated for the adoption of BDA tools to minimize risks, improve productivity, quality control, decision-making capability etc. The reader could get a clearer idea of these barriers, which help them to make tactical and strategic policies regarding BDA implementation. A thorough investigation of barriers to BDA can make manufacturing organizations build more effective strategies. Important barriers identified from the literature are discussed as follows.

2.2.1 Lack of policies for data security and privacy. As organizations increasingly have access to confidential consumer data in the present Industry 4.0 era, information and cyber security issues arise. While access to confidential information is heavily regulated on an intra-organization level, data security may be a concern from a customer perspective (Kache and Seuring, 2017). Data security and privacy are among the most common issues for investment in BDA for SMOs (Maroufkhani *et al.*, 2020). Data must be secure if the organizations have to compete in the global market (Alharthi *et al.*, 2017; Delen and Ram, 2018). Jensen and Remmen (2017) have observed that while sharing the data, confidentiality must be maintained among all the stakeholders. There is a lack of policy framework which can give confidence to investors. Unethical use of big data and ineffective processing of data leads to the concern of privacy and security. Lack of policies, outdated regulations has been a big hurdle while data are shared especially when it comes to consumer data. Lack of policies becomes a major hurdle for multinational organizations as they are obliged to abide by the laws of different countries while sharing data across their global supply chains (Alfaro *et al.*, 2015). This barrier should be minimized by framing suitable policies for all the stakeholders.

2.2.2 Absence of data-driven decision-making culture. Data-driven culture and employees play a crucial role in absorbing and assimilating knowledge. Decisions are generally better if taken on the basis of information extracted from the data. However, the collection of data for proper analysis is a complicated task for manufacturers (Kamble and Gunasekaran, 2020). Manufacturers should develop data-driven decision-making culture to deal with this barrier (Gupta *et al.*, 2020). Data transparency and accountability should be nurtured to develop a data-driven culture (Malomo and Sena, 2017). In addition, the role of big data service providers can be extended to influence the attitude and decisions of top management regarding BDA (Lai *et al.*, 2018). The unavailability of suitable big data services discourages organizations to invest in BDA.

2.2.3 High cost of developing digital infrastructure. Developing a digital infrastructure is a prime requirement for implementing BDA for manufacturing organizations (Belhadi *et al.*, 2019). Data infrastructure plays a vital role in implementing BDA or BDA-enabled architecture (Kim *et al.*, 2014; Barbierato *et al.*, 2014). This infrastructure requires hardware, software systems, and various tools for the collection, processing and analysis of data. The architecture should be dynamic as well as smart which supports the scalability of a large amount of data. In addition, the architecture should support different sensors used under analytical tools (Raut *et al.*, 2019; Dubey *et al.*, 2019a, b).

However, to develop, digital infrastructure organizations require a high investment (Sivarajah *et al.*, 2017; Wang and Wiebe, 2016). Therefore, the high cost of digital infrastructure is a major problem for the organization. Moreover, there are problems in high-speed Internet access that have a severe impact on the implementation of emerging technologies (i.e. BDA/Industry 4.0) in manufacturing industries.

2.2.4 Ineffective performance framework for assessing the effectiveness of investments on new technologies. There are several challenges that organizations face in implementing new technology. Although the managers are well qualified and experienced for leading a team for the development of new technical innovation but face several challenges in its implementation (Belhadi *et al.*, 2019; Alharthi *et al.*, 2017). There is inadequate performance measure in practice to check the implementation of new technology. It is the responsibility of organizations to ensure that the effectiveness of new technology obtains organizational goals which are essential for their growth.

2.2.5 Rigid organization culture for making new investments on technologies. Organizational culture plays a crucial in the implementation of new technologies. Culture is the implicit norm that defines employee behavior. A major obstacle to investment in new technologies is the rigid and reluctant organizational culture. Conventional organizational culture lacks flexibility and is quite averse to change (Jahanshahi and Brem, 2017). A radical shift is required in the organization's culture to make investments on new technologies and to focus on all levels of the organization (Seth *et al.*, 2018). Therefore, organizations need to build a flexible culture to invest in new technologies that allow individuals of various specialties, expertise and skills to access the same information and to encourage each other in all aspects of work (Kalema *et al.*, 2016).

2.2.6 Lack of confidence of return on investment in BDA implementation. Return on investment refers to the ratio of net profit to the cost of investment in BDA implementation. In other words, it measures the amount of return on a specific investment, relative to the investment's cost. Justifying and measuring the return on investment is a challenge to BDA implementation (Frizzo-Barker *et al.*, 2016). The lag time between investment in big data and its return is important and should be reduced (Delen and Ram, 2018; Schüll and Maslan, 2018). The return on investment is a challenging task for BDA as this highly depends on the "downstream" employee, who is responsible for the execution of the task.

2.2.7 Lack of research on applications of BDA tools for sustainable operations. The research on applications of BDA tools for sustainable operations is the key to developing new technologies that may process the voluminous data for meaningful inferences. The manufacturing industries hesitate to invest in new technologies for BDA implementation as there is limited research on the application of BDA tools for sustainable operations (Belhadi *et al.*, 2019).

2.2.8 High cost associated in managing unstructured huge data. Organizations require huge investment costs in terms of implementation and managing an unstructured high volume of data to control over their data and provide more security (Schüll and Maslan, 2018). Data governance is also seen as a major challenge in most of the cases by organizations. Although there are advancements in cloud computing technology and hardware equipment, organizations are still facing problems with data storage, its management and mainly extracting valuable information from the data at a lower cost (Sivarajah *et al.*, 2017). Therefore, due to the high cost of managing unstructured data, organizations are unwilling to invest in BDA.

2.2.9 Unavailability of specific BDA tools as per industry requirements. Hadoop, MapReduce, Spark, Flink etc are commonly used BDA tools that have many similar functions. It is difficult for organizations to assess the appropriateness of the BDA tools. The processing of unstructured or semi-structured data is a bigger challenge than managing big data (Kaisler *et al.*, 2013; Maroufkhani *et al.*, 2020; Raut *et al.*, 2019). This requires specific BDA tools to deal with big data in the manufacturing sector. However, there is a scarcity of suitable tools and technology.

2.2.10 Absence of coordination among stakeholders for making investment on BDA-related activities. A stakeholder viewpoint is important to provide the required framework for a shift toward the new technologies (Moktadir *et al.*, 2019; Malomo and Sena, 2017). Flexible stakeholder management is the key starting point to develop a competitive strategy in terms of cost, quality and timings (Aboelmaged, 2018; Barzegar *et al.*, 2018). Additionally, it also includes an awareness of their current interests and coordination among stakeholders in the adoption of BDA. According to Zhou *et al.* (2019), the lack of participation of the stakeholders hinders the decision-making process for attaining sustainability. The coordination among stakeholders is important as many times, there is a shortage of time for checking the results of BDA as it might take up to 12–18 months. Because of this, BDA needs constant support from all the top management teams including all key stakeholders of the organization. Therefore,

there is a need for collective action on the part of all major stakeholders falling in the ambit of the BDA implementation framework. However, this aspect is lagging currently.

2.2.11 High cost associated in integrating data across the supply chain. BDA requires investment in new expensive IT infrastructure, employee skill training, business experts and analysis tools (Ahmed *et al.*, 2018; Schüll and Maslan, 2018; Sun *et al.*, 2018). Despite the declining high cost of information technology, BDA is still expensive (Sivarajah *et al.*, 2017). Due to the high cost associated with integrating data across the supply chain, the organizations are reluctant to invest in BDA (Lai *et al.*, 2018; Maduku *et al.*, 2016).

2.2.12 Inadequate data sharing policy among stakeholders. The ability to process and share a high volume of data remains a big task for existing data processing applications and database management systems (Jiang *et al.*, 2015). While forming a cross-functional team to implement BDA, organizations have an inadequate data sharing policy which gives rise to principle-agent conflicts and inappropriate incentive arrangements within the network. However, there is still no appropriate legal framework to regulate data sharing among stakeholders. Leaders play an important role at this stage. There should be an effective use of BDA to create business value for the firms.

2.2.13 Lack of competence for using BDA in resource optimization. Many organizations generally do not have sufficient technical skills to adopt new technologies related to BDA. There is a lack of adequate training programs to enhance the capabilities of employees for the adoption of BDA technologies. Gupta *et al.* (2019) emphasized that managerial and technical skills in big data predictive analytics have a positive impact on the adoption of BDA which is presently lacking. Organizations lack competence for using resource optimization as the supply chain partners have insufficient resources and lack the capability to share the data and to extract the information from them on a real-time basis. There must a proper collaboration and coordination among all the cross-functional teams of the organization to implement BDA.

2.2.14 Lack of support from employees for implementing new technologies. It is important for organizations to be able to update their tool on a regular basis in order to remain competitive and ensure that these changes are accepted by their employees (Moktadir *et al.*, 2019). Sometimes the ambiguity about what modern technology entails for employees create resistance to accept the new technologies. The organizations face problems while investing in new technology without proper management and training for their employees (Gawankar *et al.*, 2020).

2.2.15 High cost of hiring skilled BDA consultants. BDA requires highly skilled professionals (Kim *et al.*, 2014; Barbierato *et al.*, 2014). Some organizations hire consultants to act as advisors on various issues including the hardware and software for new technologies. However, hiring such experts involve a high cost that becomes an obstacle for organizations to invest in BDA (Alharthi *et al.*, 2017). Therefore, due to a lack of financial readiness to cover the cost of BDA, investment in BDA may lead to failure (Alalawneh and Alkhatib, 2020).

2.2.16 High cost of training programs on BDA. The employee must be well versed with new technologies. The organizations are facing the problem of a lack of big data skills in their employees. This is due to the lack of effective training programs and employees less involvement and interest in the new technologies (Oncioiu *et al.*, 2019). The employees should be ready to keep updating their knowledge in their areas through training and workshops (Raut *et al.*, 2019; Dubey *et al.*, 2019a, b). However, the cost incurred for these training and workshops is higher.

The skill training programs on BDA require high costs which are deterring organizations to invest in such programs due to fear of inadequate returns. To address the problem, all the IT leaders should come up together to work and develop new strategies that can address the issues of BDA. Without proper training, the potential of new technology cannot be fully tapped.

2.2.17 Lack of trust and commitment among employees. The employees trust and commitment play an important role in the successful implementation of any new technology in an organization (Zhang *et al.*, 2017). Employees have a pervasive fear of change that automation of their particular work process may lead to their retrenchment. They fear losing a competitive advantage and lacks trust due to the sensitivity of the data. Therefore, trust and commitment among employees is a crucial barrier to invest in BDA. The supportive leadership plays an important role to take out fear from the employees regarding the change (Schüll and Maslan, 2018).

The literature review reveals that many studies were conducted on the use of big data services by enterprises primarily on their perceptions about benefits, costs and data quality (Shin, 2016). In this context, no comprehensive review is available where analysis of barriers to BDA implementation in different manufacturing processes has been considered. Previous studies have mostly built conceptual frameworks for BDA in a given situation and have ignored examining the nature of the barriers to BDA. Due to the significance of BDA, this research aims to identify and analyze the barriers to BDA implementation in manufacturing. Manufacturing industries are not aware of the maturity level of BDA or whether the organization's current capabilities are sufficient for conducting the implementation of a BDA (Verma, 2017).

Table 2 summarizes barriers which were identified through extensive literature review and on the basis of the opinion of experts from industry and academia.

3. Research methodology

This section deals with the methodology adopted for identification, prioritization and categorization of barriers to the investment of BDA in SMOs. The study comprises three phases. The first phase is the identification of barriers to investment in BDA based on literature review and discussion with domain experts. In the second phase, the descriptive analysis for the selection of important barriers on the basis of descriptive measures and exploratory analysis for categorization of barriers is done using factor analysis. The third phase deals evaluation of the intensity of barriers using the GTMA.

Factor analysis is a technique to identify groups of factors that are related to the specified category (Dobriban and Owen, 2019). It is a statistical method that represents the relationships among a set of observed factors in terms of common factors. Factor analysis is an acceptable and more concise method that addresses the multicollinearity problem. Therefore, in the present study factor analysis is employed for the categorization of the barriers in the implementation of BDA.

Graph theory and matrix approach are used in modeling and solving a decision-making problem with multiple and interrelated attributes (Geetha and Sekar, 2018). The GTMA is employed for this work as it is a simple and powerful technique. Factor analysis is used to identify the hidden dimensions or constructs which may or may not be apparent from the direct analysis. It allows to investigate concepts that are difficult to quantify directly by reducing a huge number of variables to a few interpretable underlying factors. It is applied to extract the barriers in BDA by the statistical package for social sciences (SPSS Version 25.0). The GTMA is a multiple attribute decision-making (MADM) methodology used to find the intensity/quantification of barriers to adopting BDA.

Cluster analysis (CA), Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) and analytic network process (ANP) is the other decision-making approach (GCI). The factor analysis and graph theory hybrid approach are based on digraphs and permanent value computation, which does not require a hypothesis about interdependency (Tuljak-Suban and Bajec, 2020). Figure 1 depicts the research methodology framework adopted for this study and the same is discussed in detail in the following subsections.

SN	Barriers	References
1	Lack of policies for data security and privacy	Maroufkhani *et al.* (2020), Alharthi *et al.* (2017), Oncioiu *et al.* (2019), Malomo and Sena (2017), Jensen and Remmen (2017), Alfaro *et al.* (2015)
2	Absence of data-driven decision-making culture	Malomo and Sena (2017), Zhang *et al.* (2017), Gupta and George (2016), Kamble and Gunasekaran (2020), Gupta *et al.* (2020), Malomo and Sena (2017) Lai *et al.* (2018)
3	High cost of developing digital infrastructure	Maroufkhani *et al.* (2020), Sivarajah Zhang *et al.* (2017), Alharthi *et al.* (2017), Belhadi *et al.* (2019) Kim *et al.* (2014), Barbierato *et al.* (2014)
4	Ineffective performance framework for assessing effectiveness of investments on new technologies	Oncioiu *et al.* (2019), Belhadi *et al.* (2019), Schüll and Maslan (2018)
5	Rigid organization culture for making new investments on technologies	Schüll and Maslan (2018), Kamble *et al.* (2020a, b), Seth *et al.* (2018)
6	Lack of confidence of return on investment in BDA implementation	Oncioiu *et al.* (2019), Moktadir *et al.* (2019), Schüll and Maslan (2018)
7	Lack of research on applications of BDA tools for sustainable operations	Belhadi *et al.* (2019), Arunachalam *et al.* (2018)
8	High cost associated in managing unstructured huge data	Schüll and Maslan (2018), Belhadi *et al.* (2019)
9	Unavailability of specific BDA tools as per industry requirements	Maroufkhani *et al.* (2020), Raut *et al.* (2019), Dubey *et al.* (2019a, b)
10	Absence of coordination among stakeholders for making investment on BDA related activities	Zhou *et al.* (2019), Aboelmaged (2018), Barzegar *et al.* (2018), Moktadir *et al.* (2019), Malomo and Sena (2017)
11	High cost associated in integrating data across the supply chain	Maroufkhani *et al.* (2020), Raut *et al.* (2019), Arunachalam *et al.* (2018)
12	Inadequate data sharing policy among stakeholders	Janssen *et al.* (2017), Mishra and Rane (2019)
13	Lack of competence for using BDA in resource optimization	Mazzei and Noble (2017), Janssen *et al.* (2017), Raut *et al.* (2019)
14	Lack of support from employees for implementing new technologies	Maroufkhani *et al.* (2020), Moktadir *et al.* (2019)
15	High cost of hiring skilled BDA consultants	Oncioiu *et al.* (2019), Dubey *et al.* (2016), Malomo and Sena (2017), Alharthi *et al.* (2017), Raut *et al.* (2019), Dubey *et al.* (2019a, b), Kim *et al.* (2014), Barbierato *et al.* (2014)
16	High cost of training programs on BDA	Malomo and Sena (2017), Oncioiu *et al.* (2019), Kamble *et al.* (2020a, b), Raut *et al.* (2019), Dubey *et al.* (2019a, b)
17	Lack of trust and commitment among employees	Schüll and Maslan (2018), Zhang *et al.* (2017)

Table 2. Barriers for investment in of BDA for sustainable manufacturing operations

3.1 Statistical data analysis

The following sub-section discusses the statistical data analysis in terms of descriptive analysis for selecting the important barriers and exploratory analysis for categorizing the barriers.

3.1.1 Data collection. A questionnaire was developed in consultation with domain experts for data collection. The expert's responses on the impact of barriers on investment in BDA for sustainable operations in Indian manufacturing sector organizations were gathered on a five-point Likert scale (i.e. 1, Very low; 2, Low; 3, Medium; 4, High; 5, Very high). Overall, 201 responses were collected from manufacturing firms. The demographics details of respondents under the categories of respondent profile, type of industry, experience and number of employees are shown in Table 3.

General managers, managers, assistant managers and others fall in the respondent's profile category. The types of industries selected for data collection include automobile, metal and machinery, sheet metal and other industries. Under the experience category the respondents were grouped in the experience of 0–5, 6 to 10, 11–15, 16–20 and more than 20 years. The last category of respondents is related to number of employees in the industry

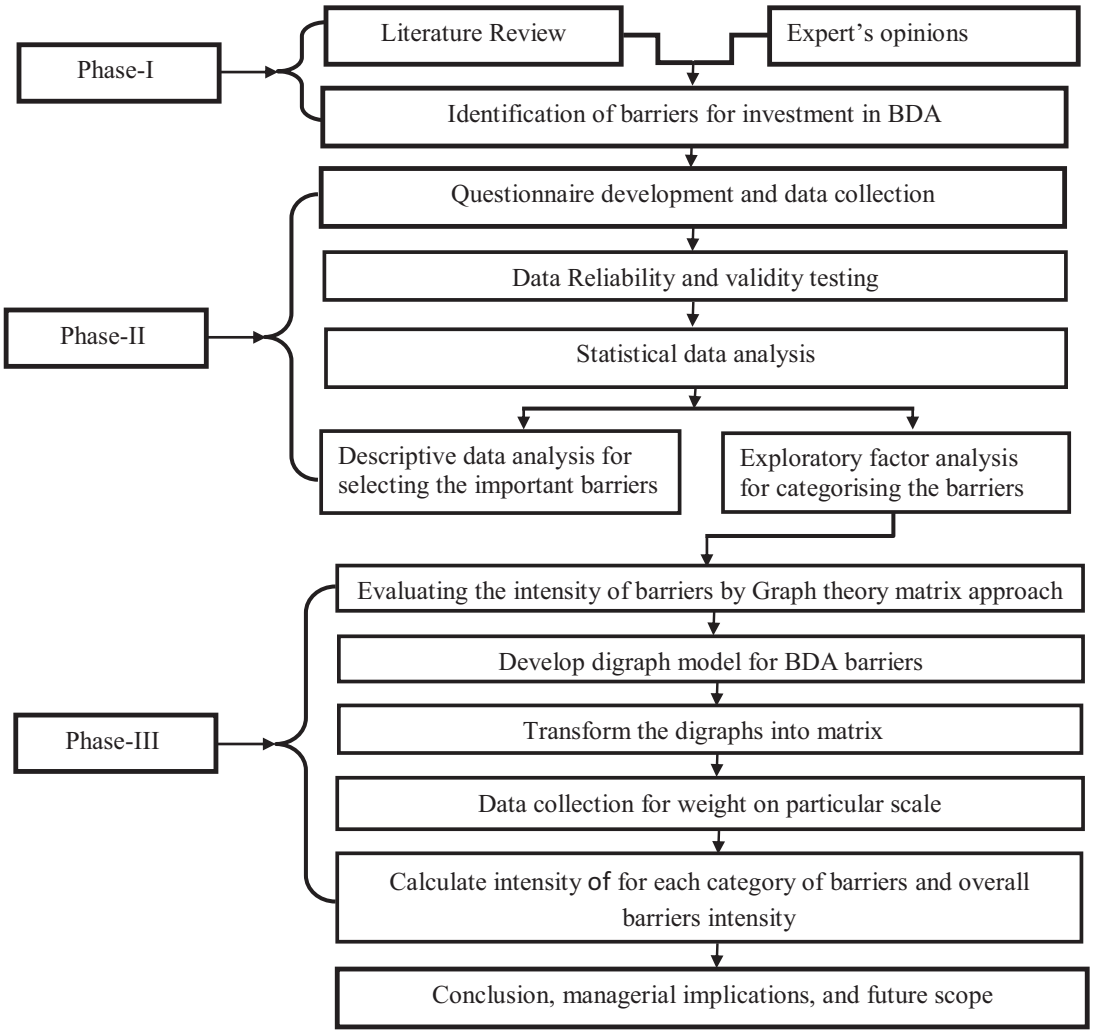


Figure 1.
Research methodology framework

(i.e. <100 to >500). The most responses were obtained from assistant managers, automobile industries and respondents having 16–20 years of experience and the industry with more than 500 employees. The least responses were from other respondent profiles, respondents having 0–5 years of experience and the industry with less than 100 employees.

3.1.2 Data validation/reliability testing. The data validation check is performed using a correlation test named as Bartlett's test of sphericity. It is performed to check the homogeneity of the data. A value of less than 0.05 of the significance levels indicates that factor analysis can be performed with the available data. The reliability test is conducted for the internal consistency of factor data. A reliability measure known as Cronbach's alpha is calculated for this purpose. The value of Cronbach's alpha less than 0.7 is considered a good indicator of reliability (Singh and Kumar, 2020). Further, the Kaiser-Meyer-Olkin (KMO) analysis is done for checking the sampling adequacy. KMO value should be greater than 0.6 indicate that factor analysis may be suitable with data. These tests can easily be done by statistical software such as SPSS and Minitab.

3.1.3 Descriptive analysis. This sub-section deals with the descriptive analysis of the barriers on the basis of descriptive measures (mean and standard deviation). Based on the mean value, the ranking of the BDA barriers is done, and the most important barriers are identified. Descriptive factor analysis can be performed using statistical software such as SPSS Version 25.0 (IBM, 2017).

Categories	Demographic detail	Number of respondent	Percentage of respondent
Respondent's profile	General managers	36	17.91
	Managers	47	23.38
	Assistant managers	75	37.32
	Others	43	21.39
	Total	201	100
Type of industry	Automobile	85	42.29
	Metal and machinery	25	12.43
	Steel industry	54	26.86
	Others	37	18.42
	Total	201	100
Experience (in years)	>20	58	28.85
	16–20	62	30.84
	11–15	48	23.89
	6–10	21	10.45
	0–5	12	5.97
	Total	201	100
Number of employees	>500	77	38.31
	251–500	39	19.41
	101–250	53	26.36
	<100	32	15.92
	Total	201	100

Table 3.
Demographics details of respondents

3.1.4 Factor analysis. In this study, factor analysis is employed to find the number of categories of barriers and merge the BDA barriers into a respective category. Factor analysis is a widely used technique for data reduction as well as for the construction of measurement scales. Rajput and Singh (2019) applied the factor analysis to understand the relationship between the circular economy and industry 4.0. Raut *et al.* (2019) have used this approach for linking BDA and operations for sustainable business management. Li *et al.* (2019) applied factor analysis technique to study the water-energy-food nexus conundrum. There are seven basic steps to perform factor analysis (DeCoster, 1998). These are briefly detailed below:

Step 1: Collect data

The data are collected through responses from experts on a questionnaire designed.

Step 2: Data validation and reliability check

Data validation and reliability check is performed as per discussion in Section 3.1.2.

Step 3: Select the number of categories for barriers

The eigenvalue is calculated which is used to determine the number of categories of barriers. The number of categories is equal to the number of extractions that have eigenvalue greater than one.

Step 4: Extraction of barriers into the respective category

For extraction of barriers into respective category principal component analysis (PCA) is used. PCA reduces the dimensionality of a dataset with a large number of interrelated variables.

Step 5: Rotate factors to a final solution

The aim of factor rotation is to rotate the factor matrix to achieve the simple structure in order to improve the interpretability of the factor solution. Factors are rotated for the purpose of explanation, to find a solution that is stingy in the sense that each variable receives a considerable contribution, i.e. factor loading, from only one factor.

Step 6: Interpret factor structure

After determining the number of factors, examine the loading pattern to determine the factor that has the most influence on each variable.

Step 7: Construct factor scores for further analysis

Based on the EFA results barrier are classified into three categories. Further, the GTMA is applied to find the intensity of extracted barriers category.

3.2 Graph theory matrix approach

This section discusses the methodology for the evaluation of barrier intensity. Quantification of barriers would help in strategy formulation. The GTMA is an important tool for decision-making that has been used widely. The GTMA has the power to solve some complex problems, and it has been applied in various applications.

The GTMA is a methodical and rational approach that is used in different areas. The digraph is dependent on the structure of the system so that changes can be analyzed. The traditional methods of representation such as flow charts, block diagrams, and cause and effect diagrams are not good in showing relationships among different variables, and therefore they are difficult to be analyzed, processed and written mathematically (Raj and Attri, 2010). The GTMA has an advantage above these traditional techniques. The GTMA makes the application more realistic and reliable as it is based on digraphs and permanent value computation, which does not require a hypothesis about interdependency. The important feature is that this approach provides a single mathematical index and converts qualitative values into quantitative values. Bhandari *et al.* (2019) used the GTMA for the evaluation of barriers intensity in the implementation of cleaner technologies. Singh and Kumar (2019) applied the GTMA approach for deriving a flexibility index for a supply chain. Singh *et al.* (2019) employed the GTMA for the evaluation of the supply chain coordination index. Agrawal (2015) and Kumar *et al.* (2017) applied the GTMA approach for the evaluation of reverse logistics strategies and maintainability, respectively. Gupta *et al.* (2017) and Agrawal *et al.* (2016) applied the GTMA for proposing a disassembly index in automotive systems and examining outsourcing decisions, respectively. This approach was also used to predict the acute ecotoxicity of chemical substances (Takata *et al.*, 2020).

The GTMA is a suitable tool for quantifying the impact of the barriers as it has been used for similar applications. Therefore, it is applied in this study. In the GTMA application initially, a relative importance matrix is developed. A 10 point scale for relative significance of attribute (r_{ij}'s) is employed (Muduli *et al.*, 2013). The data are collected in the form of responses to a questionnaire from experts.

Once the BDA barriers are identified, the relative priority of the ith barrier over jth (r_{ij}) is taken as per scale Here, the value 1 denotes very low and 5 denotes very high relative importance. The relative importance matrix is further used to calculate the index value of the individual category of barriers. Finally, the overall barrier intensity for investment in BDA is evaluated. A stepwise procedure of the GTMA is detailed in the following sub-section.

3.2.1 GTMA procedure. The main steps for applying the GTMA are as follows:

(1) Identify barriers affecting the investment in BDA for sustainable manufacturing operations and categories barriers into some groups.

(2) Construct the digraph on the basis of interdependencies among various categories.

(3) Construct a subsystem digraph and permanent matrix for the different categories of barriers. Compute permanent function value for each category of barriers.

(4) Construct inheritance and interdependency matrix for barriers by taking the expert's opinion.

(5) Calculate the index value of the different categories of barriers based on permanent function value of different categories of barriers and their interdependencies.

A generic description of the GTMA is provided in the following subsection.

3.2.2 Graph theory matrix analysis. Initially, a digraph is constructed in the GTMA. A digraph is a directed graph that contains vertices and edges. The digraph is converted into a matrix form. The permanent function of the matrix is computed in a similar manner as its determinant. To avoid any loss of information, change overall negative signs to positive signs while doing determinant calculation Grover *et al.* (2006). The permanent function value is the intensity index for this study.

In this study, three categories of barriers are identified using factor analysis. These are organization barriers (OB), data management barriers (DMB) and human barriers (HB). A digraph among these three categories of barriers is shown in Figure 2.

In the digraph, B_i's are the category of barriers represented by nodes while r_{ij}'s represent dependence through its edge. A particular value, r_{ij} represents the degree of dependence of jth barrier category on ith barrier category. A directed edge from node i to node j represents r_{ij} in the digraph. For a digraph with n number of nodes (category of barriers), an $n \times n$ matrix $[A]$ is obtained. The permanent function for the digraph is represented as follows:

$$|A| = \begin{vmatrix} B_1 & r_{12} & r_{13} & LL & r_{1n} \\ r_{21} & B_2 & r_{23} & LL & r_{2n} \\ r_{ij} & r_{ik} & B_i & LL & r_{in} \\ M & M & M & O & M \\ M & M & M & O & M \\ r_{n1} & r_{n2} & r_{n3} & LL & B_n \end{vmatrix}$$

In matrix A, elements B_i and r_{ij} are the absolute and relative values, respectively. The permanent function (per) is evaluated for index value (Gupta *et al.*, 2017).

In a similar way, digraphs for an individual category of barriers are constructed. For example, as there are seven barriers in the first category i.e. organizational barriers, the nodes $B_1^1, B_2^1, B_3^1, B_4^1, B_5^1, B_6^1, and\ B_7^1$ are used for the barriers in the first category, and r_{ij}'s symbolize the interrelationship between them in the construction of subsystem digraph. A digraph for the organization barrier representing the relationship of one node with all the other nodes is shown in Figure 3.

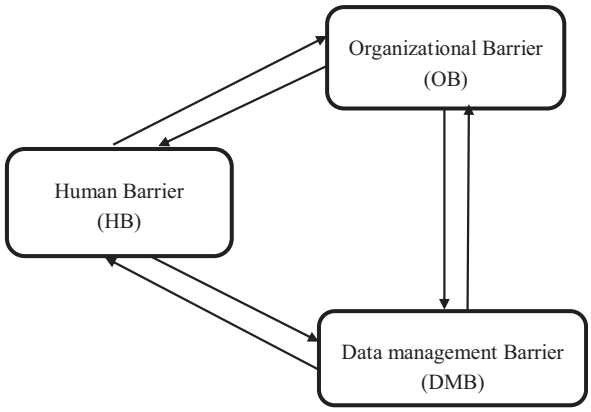


Figure 2.
Digraph of different categories of barriers

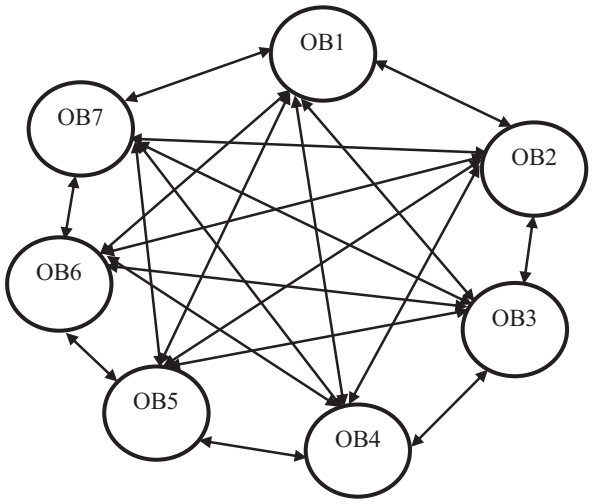


Figure 3.
Digraph for organization barrier category

As there are seven barriers in the organization barrier category, the digraph with seven nodes is converted into a 7 × 7 matrix. The permanent function, Per ($B1$) for the digraph (Figure 4) is represented as follows:

$$\begin{vmatrix} B_1^1 & r_{12}^1 & r_{13}^1 & r_{14}^1 & r_{15}^1 & r_{16}^1 & r_{17}^1 \\ r_{21}^1 & B_2^1 & r_{23}^1 & r_{24}^1 & r_{25}^1 & r_{26}^1 & r_{27}^1 \\ r_{31}^1 & r_{32}^1 & B_3^1 & r_{34}^1 & r_{35}^1 & r_{36}^1 & r_{37}^1 \\ r_{41}^1 & r_{42}^1 & r_{43}^1 & B_4^1 & r_{45}^1 & r_{46}^1 & r_{47}^1 \\ r_{51}^1 & r_{52}^1 & r_{53}^1 & r_{54}^1 & B_5^1 & r_{56}^1 & r_{57}^1 \\ r_{61}^1 & r_{62}^1 & r_{63}^1 & r_{64}^1 & r_{65}^1 & B_6^1 & r_{67}^1 \\ r_{71}^1 & r_{72}^1 & r_{73}^1 & r_{74}^1 & r_{75}^1 & r_{76}^1 & B_7^1 \end{vmatrix}$$

Where, $B_1^1, B_2^1, B_3^1, B_4^1, B_5^1, B_6^1, B_7^1$ represent OB1, OB2, OB3, OB4, OB5, OB6 and OB7 (sub barriers), Per ($B1$) = Per (OB).

Manufacturing industries face obstacles in handling data due to the complexity of data. Data management barriers refer to acquiring, storing, protecting and processing a high volume of data and transform them into useful information. A digraph for the data management barrier indicating the relationship of one node with all the other nodes is shown in Figure 4.

The data management category includes six barriers, the digraph with six nodes is converted into a 6 × 6 matrix. The permanent function, per ($B2$) for the digraph (Figure 5) is represented as follows:

$$\begin{vmatrix} B_1^1 & r_{12}^1 & r_{13}^1 & r_{14}^1 & r_{15}^1 & r_{16}^1 \\ r_{21}^1 & B_2^1 & r_{23}^1 & r_{24}^1 & r_{25}^1 & r_{26}^1 \\ r_{31}^1 & r_{32}^1 & B_3^1 & r_{34}^1 & r_{35}^1 & r_{36}^1 \\ r_{41}^1 & r_{42}^1 & r_{43}^1 & B_4^1 & r_{45}^1 & r_{46}^1 \\ r_{51}^1 & r_{52}^1 & r_{53}^1 & r_{54}^1 & B_5^1 & r_{56}^1 \\ r_{61}^1 & r_{62}^1 & r_{63}^1 & r_{64}^1 & r_{65}^1 & B_6^1 \end{vmatrix}$$

Where, $B_1^1, B_2^1, B_3^1, B_4^1, B_5^1, B_6^1$ represent DMB1, DMB2, DMB3, DMB4, DMB5 and DMB6 (sub barriers), Per ($B2$) = Per (DMB).

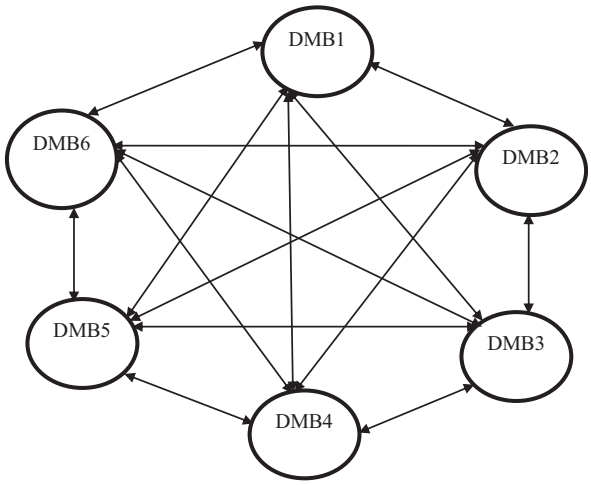


Figure 4.
Digraph for data management barrier category

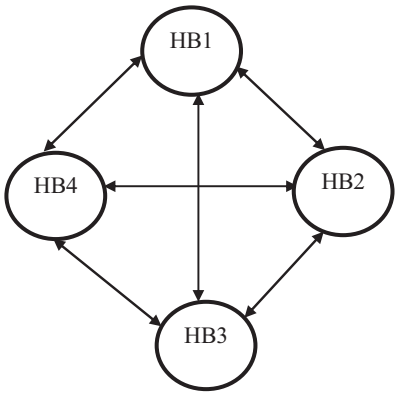


Figure 5.
Digraph for human barrier category

Human barriers refer to the lack of awareness and skill in the context of BDA that influence the implementation of BDA within the organization. This category of barrier includes the high cost of training programs on BDA, lack of support from employees for implementing new technologies, lack of trust and commitment among employees and high cost of hiring skilled BDA consultants.

A digraph for the organization barrier indicating the connection of one node with all the other nodes is shown in Figure 5. As there are four barriers in the human barriers category, the digraph with four nodes is converted into a 4×4 matrix. The permanent function, Per (B3) for the digraph (Figure 5) is represented as follows:

$$\begin{vmatrix} B_1^1 & r_{12}^1 & r_{13}^1 & r_{14}^1 \\ r_{21}^1 & B_2^1 & r_{23}^1 & r_{24}^1 \\ r_{31}^1 & r_{32}^1 & B_3^1 & r_{34}^1 \\ r_{41}^1 & r_{42}^1 & r_{43}^1 & B_4^1 \end{vmatrix}$$

Where, $B_1^1, B_2^1, B_3^1, B_4^1$ represent HB1, HB2, HB3 and HB4 (sub-barriers), Per ($B3$) = Per (HB).

This method will be used to find the permanent index value for each category of barriers. The higher the value of the permanent function more the intensity of barriers for investment in BDA for SMOs.

4. Results and discussion

On the basis of the literature review and experts' opinion, identified barriers were categorized into three categories namely organizational barriers, data management barriers and human barriers selected. Manufacturing sectors face unique barriers to implement BDA in manufacturing operations. The identified barriers, from the related literature, was included in the questionnaire. First, BD experts were asked to identify the expected BDA benefits across sectors. Then, they were asked to select the most relevant barriers to BDA adoption in sustainable manufacturing operations. These findings may help decision makers to understand the barriers to BDA implementation. These barriers may also be an issue in the adoption of BDA, and manufacturers are unwilling to adopt BDA due to the extra investment required. In this section, the results and the discussion are presented in two parts. The first part covers the descriptive and exploratory statistical analysis of BDA barriers, and the second part discusses the results of the GTMA.

4.1 Findings of statistical analysis

This sub-section presents the findings of statistical analysis of data obtained from respondents and factor analysis is applied to factorize the 17 identified BDA barriers into three groups. The data were collected from 201 respondents for this analysis. The demographics details of respondents are given in Table 3. The reliability test was conducted for checking the internal consistency of the data by using statistical software – SPSS Version 25.0. The value of Cronbach's alpha was found to be 0.869 which is a good indicator of reliability i.e. more than 0.7. Further, Bartlett's test was performed and found to be significant as the p-value is less than 0.05. The determinant of the correlation matrix is 0.000 greater than 0.00001, then there is no multi collinearity. The KMO value obtained is 0.882 which is higher than 0.6 which shows that the sampling is a good indicator of the internal consistency of the barriers. These statistical tests established that the data was suitable for factor analysis.

Further, the descriptive analysis of the barriers was carried out that includes mainly mean and standard deviation. The means for the seventeen barriers were used to rank order the barrier from the most important to least important (refer Table 4). The most important barrier to invest in BDA was the lack of support from employees for implementing new technologies (mean = 4.1642). High cost associated in integrating data across the supply chain, high cost of training programs on BDA, lack of trust and commitment among employee's inadequate data sharing policy among stake holders high cost of hiring skilled BDA consultants very closely as the next most important barriers (mean = 4.0746, 3.9403, 3.9204, 3.8905 and 3.8856). Less important barriers were lack of policies for data security and privacy and lack of research on applications of BDA tools for sustainable operations (mean = 3.5423 and 3.5920) and these are low obstacles to investment in BDA.

On the basis of eigenvalues greater than 1 obtained in factors analysis, three categories of barriers are formulated. These three categories are named as organizational barriers, data management barriers and human barriers in consultation with domain experts (refer Table 5). The organization barriers include seven barriers that explained 22.617% of the variance. The PCA was used as an extraction method for segregating the barriers. Varimax with Kaiser Normalization (cutoff = 0.50) was used as a rotation method. Finally, three components are taken after rotation of the component matrix with Varimax. These barriers are segregated into a particular category on the basis of their factor loading.

The seven barriers extracted for organization barriers category were lack of policies for data security and privacy (77.10%), high cost of developing digital infrastructure (68.70%), absence of data-driven organizational culture (66.90%), rigid organization culture for making new investments (65.50%), lack of confidence of return on investment in BDA implementation (63.90%), lack of research on applications of BDA tools for sustainable operations (63%) and ineffective performance framework for assessing effectiveness

Barriers to investment in BDA for sustainable operations	Mean	Std. deviation	Ranking
Lack of policies for data security and privacy	3.5423	1.28041	17
Absence data-driven decision-making culture	3.6020	1.17507	15
High cost of developing digital infrastructure	3.6517	1.22805	12
Ineffective performance framework for assessing effectiveness of investments on new technologies	3.6567	1.17753	11
Rigid organization culture for making new investments	3.8060	1.21539	8
Lack of confidence of return on investment in BDA implementation	3.7463	1.09558	9
Lack of research on applications of BDA tools for sustainable operations	3.5920	1.11926	16
High cost associated in managing unstructured huge data	3.7214	1.04498	10
Unavailability of specific BDA tools as per industry requirements	3.6368	1.17151	14
Lack of support from employees for implementing new technologies	4.1642	1.01386	1
Absence of coordination among stakeholders for BDA related activities	3.6468	1.18305	13
High cost associated in integrating data across the supply chain	4.0746	1.02927	2
High cost of hiring skilled BDA consultants	3.8856	1.23363	6
High cost of training programs on BDA	3.9403	1.11194	3
Inadequate data sharing policy among stakeholders	3.8905	1.10361	5
Lack of trust and commitment among employees	3.9204	1.08795	4
Lack of competence for using BDA in resource optimization	3.8061	1.12123	7

Table 4.
Descriptive analysis of the barriers for investment in BDA

Categories of barriers	Barriers for investment in BDA for sustainable operations	Factor loading	Eigen value
Organizational barriers	Lack of policies for data security and privacy	0.771	5.625
	High cost of developing digital infrastructure	0.687	
	Absence of data-driven organizational culture	0.669	
	Rigid organization culture for making new investments	0.655	
	Lack of confidence of return on investment in BDA implementation	0.639	
	Lack of research on applications of BDA tools for sustainable operations	0.630	
	Ineffective performance framework for assessing effectiveness of investments on new technologies	0.618	
Data management barriers	Lack of competence for using BDA in resource optimization	0.62	1.367
	Absence of coordination among stakeholders for BDA related activities	0.617	
	Unavailability of specific BDA tools as per industry requirements	0.615	
	Inadequate data sharing policy among stakeholders	0.58	
	High cost associated in integrating data across the supply chain	0.517	
	High cost associated in managing unstructured huge data	0.501	
Human barriers	High cost of training programs on BDA	0.700	1.255
	Lack of support from employees for implementing new technologies	0.641	
	Lack of trust and commitment among employees	0.588	
	High cost of hiring skilled BDA consultants	0.568	

Table 5.
Factor analysis results to barriers for investment in BDA

(61.80%). This category accounted for 33.088% of the variance (refer Table 6). The authors claim that the awareness level of BDA among the manufacturing industry is moderate for the majority of the survey respondents. Verma *et al.* (2017) found that proper technology infrastructures, organization culture and architecture standards should be prepared for a successful investment in BDA. Beath *et al.* (2012) observed that the lack of data-driven

Table 6.
Total variance explained

Component	Initial eigenvalues			Extraction sums of squared loadings			Rotation sums of squared loadings		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
1	5.625	33.088	33.088	5.625	33.088	33.088	3.845	22.617	22.617
2	1.367	9.044	42.132	1.367	9.044	42.132	2.580	16.178	38.795
3	1.255	8.385	50.517	1.255	8.385	50.517	1.823	11.722	50.517
4	0.990	5.826	54.343						
5	0.932	5.481	59.825						
6	0.848	4.990	64.815						
7	0.796	4.682	69.497						
8	0.771	4.537	74.034						
9	0.683	4.015	78.050						
10	0.610	3.589	81.638						
11	0.551	3.239	84.878						
12	0.532	3.132	88.010						
13	0.501	2.948	90.958						
14	0.429	2.526	93.484						
15	0.418	2.457	95.941						
16	0.375	2.206	98.147						
17	0.315	1.853	100.000						

Note(s): Extraction method: Principal component analysis

culture among organizations is the major reason for the high failure rate of BDA implementation.

The six barriers extracted for the data management barriers category were lack of competence for using BDA in resource optimization (62%), absence of coordination among stakeholders for BDA-related activities (61.70%), unavailability of specific BDA tools as per industry requirements (61.50%), inadequate data sharing policy among stake holders (58%), high cost associated in integrating data across the supply chain (51.70%) and high cost associated in managing unstructured huge data (51%). The numbers in parentheses are the factor loadings. This category accounted for 9.044% of the variance. Nwankpa and Roumani (2014) observed that data management is an important barrier category as the top management plays an important role by providing support and resources to implement new technologies. Baldwin (2015) argued that 65% of organizations have reported below-average returns on their data management. Therefore, proper consideration of data management is necessary for the manufacturing sector.

The four barriers extracted for the human barriers category were high cost of training programs on BDA (70%), lack of support from employees for implementing new technologies (64.10%), lack of trust and commitment among employees (58.80%) and high cost of hiring skilled BDA consultants (56.80%). This category accounted for 8.385% of the variance. Alalawneh and Alkhatib (2020) argued that organizations should provide good training/education programs, sufficient resources, supportive facilities and appropriate consulting services to facilitate investment in BDA.

4.2 GTMA application for evaluation of barriers intensity index

For an illustration of the suggested methodology for the analysis of barriers intensity in the investment of BDA for SMOs, a passenger car manufacturing company, ABC Limited has been considered. The company is located in the National Capital Region of Delhi. The company needs to upgrade at the technology front to meet customers' expectations. Stakeholders have become increasingly conscious of fuel efficiency, CO_2 emissions and other indicators of environmental sustainability. There is an urgent need to accept different changes in manufacturing processes and product design to sustain and expand the business simultaneously. The ABC limited is currently working to meet its vehicle CO_2 emission reduction target to reduce by 30%. To ensure environment-friendly operations, the industry is looking for investment in BDA and new technologies for SMOs. In this context, first barriers are identified and further grouped into three categories with the help of factor analysis. Subsequently, the GTMA is used to quantify the barriers to invest in BDA for SMOs.

In this section, the intensity index for each category of barriers is evaluated as per the GTMA procedure detailed in section 3. These categories are organizational, data management and human barriers. Permanent matrices for these categories of barriers are constructed on 1–10 scale (1 for very low and 10 for very high) using, $r_{ji} = 10 - r_{ij}$. To evaluate the permanent matrix index of organizational barriers, required inputs received from experts for absolute and relative values of barriers are:

$$B_1^1 = 3,\ B_2^1 = 3,\ B_3^1 = 3,\ B_4^1 = 4,\ B_5^1 = 4,\ B_6^1 = 5,\ B_7^1 = 4.\ r_{12}^1 = 3\ r_{13}^1 = 2\ r_{14}^1 = 4,\ r_{15}^1 = 5,$$
$$r_{16}^1 = 2\ r_{17}^1 = 4,\ r_{21}^1 = 7\ r_{23}^1 = 4\ r_{24}^1 = 3\ r_{25}^1 = 2\ r_{26}^1 = 2\ r_{27}^1 = 3\ r_{31}^1 = 8\ r_{32}^1 = 6\ r_{34}^1 = 5\ r_{35}^1 = 4$$
$$r_{36}^1 = 3\ r_{37}^1 = 2,\ r_{41}^1 = 6\ r_{42}^1 = 7\ r_{43}^1 = 4\ r_{45}^1 = 2\ r_{46}^1 = 4\ r_{47}^1 = 3\ r_{51}^1 = 5\ r_{52}^1 = 8\ r_{53}^1 = 6\ r_{54}^1 = 8$$
$$r_{56}^1 = 3\ r_{57}^1 = 2,\ r_{61}^1 = 8\ r_{62}^1 = 8\ r_{63}^1 = 7\ r_{64}^1 = 6\ r_{65}^1 = 7\ r_{67}^1 = 4,\ r_{71}^1 = 6\ r_{72}^1 = 7\ r_{73}^1 = 8\ r_{74}^1 = 7$$
$$r_{75}^1 = 8\ r_{76}^1 = 6.$$

The permanent value of a matrix of organizational barriers (*OB*) is calculated as follows.

$$Per(OB) = \begin{vmatrix} 3 & 3 & 2 & 4 & 5 & 2 & 4 \\ 7 & 3 & 4 & 3 & 2 & 2 & 3 \\ 8 & 6 & 3 & 5 & 4 & 3 & 2 \\ 6 & 7 & 5 & 4 & 2 & 4 & 3 \\ 5 & 8 & 6 & 8 & 4 & 3 & 2 \\ 8 & 8 & 7 & 6 & 7 & 5 & 4 \\ 6 & 7 & 8 & 7 & 8 & 6 & 4 \end{vmatrix} = 210684578$$

Similarly, values of the permanent matrix of data management barriers and human barriers are calculated as:

$$Per(DMB) = \begin{vmatrix} 5 & 3 & 4 & 3 & 3 & 2 \\ 7 & 4 & 2 & 3 & 3 & 2 \\ 6 & 8 & 5 & 2 & 3 & 4 \\ 7 & 7 & 8 & 3 & 4 & 3 \\ 7 & 7 & 7 & 6 & 3 & 4 \\ 8 & 8 & 6 & 7 & 6 & 4 \end{vmatrix}$$

$$Per(HB) = \begin{vmatrix} 9 & 7 & 8 & 7 \\ 3 & 7 & 9 & 7 \\ 2 & 1 & 8 & 8 \\ 3 & 3 & 2 & 9 \end{vmatrix}$$

The values obtained from the permanent matrix of data management barriers and human barriers are 6264473 and 21854, respectively.

On the similar lines, the overall barrier index value is evaluated. The overall barriers intensity (*OBI*) for investment in BDA index value is computed as:

$$Per(OBI) = \begin{vmatrix} 21854 & 6 & 5 \\ 4 & 6264473 & 7 \\ 5 & 3 & 210684578 \end{vmatrix} = 3.86 \times 10^{17}$$

The amount of negative influence of a specific class of barriers for investment in BDA depends on its index value. A higher index value means more the intensity of barriers for investment in BDA for SMOs. It is observed that the organizational barriers have the highest (210684578) intensity whereas the human barriers have the smallest (21854) intensity. The results are similar to other findings in the literature. Gunasekaran and Spalanzani (2012) concluded in their study that organizations play an important role in the implementation of BDA and new technologies for sustainable activities. Organizations should motivate employees to work for the goal of sustainable operations and inculcate business ethics (Govindarajulu and Daily, 2004). The organizational barriers must be given more importance followed by data management barriers and human barriers. To reduce the impact of these barriers, top management should have planned for confirming SMOs. These technologies (such as BDA, I4.0) improve organizational performance as well as reduce pollution and resource consumption (Saleem *et al.*, 2020). Organizations should develop capabilities for

implementing BDA to meet the requirements of sustainable operations (Vargas *et al.*, 2018). Many organizations focus on short-term goals thereby ignore the implementation of new technologies such as BDA, industry 4.0, etc. due to the fear of heavy investment and risks of failure. Organizational culture also plays a crucial role in motivating employees for adopting the changes within the organization. In terms of organizational barriers, there is a clear compromise among experts (65.22%) about the importance of this category for the manufacturing sector (Alalawneh and Alkhatib, 2020).

The data management barrier category has the second-highest index (6264473) value, and it also influences to investment in BDA. Further organizations adopt emerging technologies in their manufacturing operations and manage the increasing data flow in their value chain for effective management (Ghadge *et al.*, 2020). Organizations should be flexible to adopt new technologies. Therefore, organizations need to comprehend the significance of new technologies and overcome different barriers to investment in new technologies.

5. Conclusion

Organizations are confronting various problems after the outbreak of the COVID-19 pandemic such as industrial data access, data integration, data sharing, data storage/management and data analysis while implementing new technologies for sustainability in manufacturing operations. To ensure sustainable operations, BDA plays a crucial role by setting up efficient manufacturing processes, with data-driven decision making and through efficient usage of available resources for manufacturing organizations. However, there are many challenges to invest for implementing BDA in manufacturing organizations. In this study, barriers in the investment of BDA are identified through literature and by taking the expert's opinions. Further, factor analysis and GTMA tools are employed to factorize the various barriers and to evaluate the intensity of different categories of barriers.

From the literature and experts' opinion, a total of 17 barriers have been finalized and based on factor loading the barriers are grouped into three categories i.e. organization barriers, data management barriers and human barriers. These categories are prioritized on the basis of the value of the index for barriers intensity. High index value depicts the higher intensity of barriers, whereas the low value of index signifies the lesser intensity of barriers in BDA implementation. To overcome these barriers, top management should have comprehensive performance measures such as efficiency, effectiveness and industry rankings etc. Also, the organizations should run awareness programs for taking corrective actions to eliminate the root cause of the problem, preventing future issues regarding BDA implementation. Top management should inspire employees so that they could take part in the initiatives such as skill enhancement, employee engagement activities, training and development program etc. of the company for BDA implementation. Organizations may utilize an illustrated approach to quantify the barriers in different sectors.

This research work might contribute useful information to develop a relevant theory for implementing BDA. More broadly, this work provides methods to examine the impact of barriers to BDA implementation. The findings of the study will also contribute significantly to motivate organizations to invest in BDA application in manufacturing organizations for sustainable operations. This research may also help manufacturing organizations in developing strategies for making effective investments for BDA implementation.

The main contribution of this study is to encourage organizations toward BDA adoption by strategically overcoming the barriers encountered. The unique contribution of the research is understanding barriers and overcoming them will lead to the successful implementation of BDA. Also, the intensity of BDA barriers will help managers to put their efforts in the right direction and formulate appropriate strategies accordingly. Thus, results will be extremely useful for industry professionals and practitioners willing to transform

their manufacturing processes from conventional to digitalization. Ultimately, the adoption of BDA will result in increased productivity, resources optimization and efficient decision making. Implementing BDA will also result in reducing wastage, attaining sustainability and introducing a concept of the circular economy into the manufacturing system.

5.1 Limitations and future scope

Although this study contributes significantly to analyzing the intensity of barriers for making investments in BDA application for SMOs, findings cannot be generalized. It has got certain limitations. As the barriers taken into consideration are based on literature review and experts' opinions, there is a possibility that a few barriers might have been ignored or overlooked. Therefore, the list of barriers identified is not exhaustive and could change in the future depending upon the advancement in literature and variation in experts' opinions.

As a future scope, researchers could use the same approach to conduct research in context to different sectors and countries. The opinion of experts may be inconsistent while giving their inputs and it may have an impact on the overall intensity of barriers. In this study barriers to BDA adoption are examined in the context of Indian manufacturing industries. In the future, barriers to BDA using international data can be examined, other decision-making methods such as fuzzy ELECTRE, fuzzy GTMA and interval GTMA can be used to examine the barriers.

Therefore, as a future extension of this study, findings may be validated through empirical research and case studies to generalize the results. In addition, since this study is focused on the manufacturing sector, therefore, future research on service and other sectors is also recommended.

5.2 Managerial implications

With the development of digitalization, the manufacturing organizations are moving toward the BDA application but facing many challenges. The findings of this study will help managers in prioritizing strategies for making investments in BDA applications. Although BDA improves organization performance, Gobble (2013) observed that only 25% of firms that invested in BDA were able to implement it successfully. Business Software Alliance (BSA), based in the US, reported that 56% of the organizations implementing the BDA in their processes have shown more than 10% increase in their growth (Columbus, 2014). This work is expected to assist industrial managers to explore the barriers to BDA in the manufacturing sector. Upon identifying such barriers, managers will be able to formulate tactical and strategic policies regarding BDA adoption. In addition, this research may assist decision makers in preparing action plans to overcome the hurdles to using BDA. This study provides a thorough insight into the barriers of BDA for successful implementation in SMOs. This research may be used in analyzing and developing new sustainable manufacturing operations by prioritizing resources to overcome the important barriers. Successful implementation of BDA helps in efficient data collection and in ensuring the sustainability of manufacturing operations (Kamilaris *et al.*, 2017).

Thus, it is a prerequisite for managers and owners of organizations to understand the factors for the effective implementation of BDA for sustainable manufacturing. The findings of this study may assist industrial managers to explore the use of BDA for enhancing sustainability in manufacturing. Hence, the organizational culture may be an important moderating variable in context to the adoption of BDA. However, in the BDA context the data-driven decision-making skills, organizational learning, technical skills and management skills are important antecedents of BDA implementation.

The results of this study indicate that organizational barriers mostly influence the adoption of BDA. Findings also imply that organizations should motivate their employees to

adopt new technologies and manage the cost of data integration and training programs in a more effective manner. The barriers for investment in the BDA application identified from the study have the potential to make the manufacturing operations more sustainable if these are adequately addressed by organizations by prioritizing their resources.

References

Aboelmaged, M. (2018), "The drivers of sustainable manufacturing practices in Egyptian SMEs and their impact on competitive capabilities: a PLS-SEM model", *Journal of Cleaner Production*, Vol. 175, pp. 207-221.

Acioli, C., Scavarda, A. and Reis, A. (2021), "Applying Industry 4.0 technologies in the COVID–19 sustainable chains", *International Journal of Productivity and Performance Management*, Vol. 70 No. 5, doi: 10.1108/IJPPM-03-2020-0137.

Adrian, C., Abdullah, R., Atan, R. and Jusoh, Y.Y. (2017), "Factors influencing to the implementation success of big data analytics: a systematic literature review", *2017 International Conference on Research and Innovation in Information Systems (ICRIIS)*, Langkawi, Malaysia, IEEE, pp. 1-6.

Agrawal, K.P. (2015), "Investigating the determinants of Big Data Analytics (BDA) adoption in emerging economies", *Academy of Management Proceedings*, Vol. 2015 No. 1, p. 11290.

Agrawal, S., Singh, R.K. and Murtaza, Q. (2016), "Outsourcing decisions in reverse logistics: sustainable balanced scorecard and graph theoretic approach", *Resources, Conservation and Recycling*, Vol. 108, pp. 41-53.

Ahmed, V., Aziz, Z., Tezel, A. and Riaz, Z. (2018), "Challenges and drivers for data mining in the AEC sector", *Engineering, Construction and Architectural Management*, Emerald Publishing, Vol. 25 No. 11, pp. 1436-1453.

Aho, A.-M. (2015), "Product data analytics service model for manufacturing company", *International Conference on Knowledge Management in Organizations*, pp. 282-296.

Akhtar, P., Frynas, J.G., Mellahi, K. and Ullah, S. (2019), "Big data-savvy teams' skills, big data-driven actions, and business performance", *British Journal of Management*, Vol. 30 No. 2, pp. 252-271.

Alalawneh, A.A.F. and Alkhatib, S.F. (2020), "The barriers to big data adoption in developing economies", *The Electronic Journal of Information Systems in Developing Countries*, Wiley Online Library, Vol. 87 No. 1, e12151.

Alfaro, L.A., Le, T.M.H., Choi, H.R. and Cho, M.J. (2015), "Deployment model of Big Data for port logistics", *International Information Institute (Tokyo)*, Vol. 18 No. 1, p. 7.

Alharthi, A., Krotov, V. and Bowman, M. (2017), "Addressing barriers to big data", *Business Horizons*, Vol. 60 No. 3, pp. 285-292.

Ali, I. and Gölgeci, I. (2019), "Where is supply chain resilience research heading? A systematic and co-occurrence analysis", *International Journal of Physical Distribution and Logistics Management*, Vol. 49 No. 8, pp. 793-815.

Amerioun, A., Alidadi, A., Zaboli, R. and Sepandi, M. (2018), "The data on exploratory factor analysis of factors influencing employees effectiveness for responding to crisis in Iran military hospitals", *Data in Brief*, Vol. 19, pp. 1522-1529.

Ardito, L., Petruzzelli, A.M., Panniello, U. and Garavelli, A.C. (2019), "Towards industry 4.0", *Business Process Management Journal*, Vol. 25 No. 2, pp. 323-346.

Arunachalam, D., Kumar, N. and Kawalek, J.P. (2018), "Understanding big data analytics capabilities in supply chain management: unravelling the issues, challenges and implications for practice", *Transportation Research Part E: Logistics and Transportation Review*, Vol. 114, pp. 416-436.

Aydiner, A.S., Tatoglu, E., Bayraktar, E., Zaim, S. and Delen, D. (2019), "Business analytics and firm performance: the mediating role of business process performance", *Journal of Business Research*, Elsevier, Vol. 96, pp. 228-237.

Bag, S., Wood, L.C., Xu, L., Dhamija, P. and Kayikci, Y. (2020), "Big data analytics as an operational excellence approach to enhance sustainable supply chain performance", *Resources, Conservation and Recycling*, Vol. 153, p. 104559.

Baldwin, H. (2015), "When Big Data projects go wrong", *Forbes*, available at: http://www.forbes.com/sites/howardbaldwin/2015/01/22/when-big-data-projects-go-wrong/#78ae288a623.

Banyai, T., Tamas, P., Illes, B., Stankeviciutė, Z. and Banyai, A. (2019), "Optimization of municipal waste collection routing: impact of industry 4.0 technologies on environmental awareness and sustainability", *International Journal of Environmental Research and Public Health*, Vol. 16 No. 4, p. 634.

Barbierato, E., Gribaudo, M. and Iacono, M. (2014), "Performance evaluation of NoSQL big-data applications using multi-formalism models", *Future Generation Computer Systems*, Elsevier, Vol. 37, pp. 345-353.

Barzegar, M., Ehtesham Rasi, R.N and Niknamfar, A.H. (2018), "Analyzing the drivers of green manufacturing using an analytic network process method: a case study", *International Journal of Research in Industrial Engineering*, Vol. 7 No. 1, pp. 61-83.

Belaud, J.-P., Prioux, N., Vialle, C. and Sablayrolles, C. (2019), "Big data for agri-food 4.0: application to sustainability management for by-products supply chain", *Computers in Industry*, Vol. 111, pp. 41-50.

Beath, C., Becerra-Fernandez, I., Ross, J. and Short, J. (2012), "Finding value in the information explosion", *MIT Sloan Management Review*, Vol. 53 No. 4, p. 18.

Belhadi, A., Zkik, K., Cherrafi, A., Yusof, S.M. and El fezazi, S. (2019), "Understanding big data analytics for manufacturing processes: insights from literature review and multiple case studies", *Computers and Industrial Engineering*, Vol. 137, p. 106099.

Belinski, R., Peixe, A.M.M., Frederico, G.F. and Garza-Reyes, J.A. (2020), "Organizational learning and Industry 4.0: findings from a systematic literature review and research agenda", *Benchmarking: An International Journal*, Vol. 27 No. 8, pp. 2435-2457.

Bhandari, D., Singh, R.K. and Garg, S.K. (2019), "Prioritisation and evaluation of barriers intensity for implementation of cleaner technologies: framework for sustainable production", *Resources, Conservation and Recycling*, Vol. 146, pp. 156-167.

Bhanot, N., Rao, P.V. and Deshmukh, S.G. (2017), "An integrated approach for analysing the enablers and barriers of sustainable manufacturing", *Journal of Cleaner Production*, Vol. 142, pp. 4412-4439.

Birkel, H., Veile, J., Müller, J., Hartmann, E. and Voigt, K.-I. (2019), "Development of a risk framework for industry 4.0 in the context of sustainability for established manufacturers", *Sustainability*, Vol. 11 No. 2, p. 384.

Buntak, K., Kovačić, M. and Mutavdžija, M. (2019), "Internet of things and smart warehouses as the future of logistics", *Tehnički Glasnik*, Vol. 13 No. 3, pp. 248-253.

Chams, N. and Garcia-Blandon, J. (2019), "On the importance of sustainable human resource management for the adoption of sustainable development goals", *Resources, Conservation and Recycling*, Vol. 141, pp. 109-122.

Chen, D.Q., Preston, D.S. and Swink, M. (2015), "How the use of big data analytics affects value creation in supply chain management", *Journal of Management Information Systems*, Vol. 32 No. 4, pp. 4-39.

Choi, T.-M., Wallace, S.W. and Wang, Y. (2018), "Big data analytics in operations management", *Production and Operations Management*, Wiley Online Library, Vol. 27 No. 10, pp. 1868-1883.

Ciulli, F., Kolk, A. and Boe-Lillegraven, S. (2020), "Circularity brokers: digital platform organizations and waste recovery in food supply chains", *Journal of Business Ethics*, Vol. 167 No. 2, pp. 299-331.

Columbus, L. (2014), "Making analytics accountable: 56% of executives expect analytics to contribute to 10% or more growth in 2014", *Forbes*, available at: Http://Www. Forbes. Com/Sites/

Louiscolumbus/2014/12/10/Making-Analytics-Accountable-56-of-Executivesexpect-Analytics-Tocontribute-to-10-or-More-Growth-in-2014.

Davenport, T., Guha, A., Grewal, D. and Bressgott, T. (2020), "How artificial intelligence will change the future of marketing", *Journal of the Academy of Marketing Science*, Vol. 48 No. 1, pp. 24-42.

DeCoster, J. (1998), *Overview of Factor Analysis*, Tuscaloosa, AL.

Delen, D. and Ram, S. (2018), "Research challenges and opportunities in business analytics", *Journal of Business Analytics*, Taylor & Francis, Vol. 1 No. 1, pp. 2-12.

Devansh, S. (2020), "Demystifying the new normal with data and analytics", 7 September 2020, *Dataversity*, available at: https://www.dataversity.net/demystifying-the-new-normal-with-data-and-analytics/.

Diaz, A., Rowshankish, K. and Saleh, T. (2018), "Why data culture matters", *McKinsey Quarterly*, Vol. 3, pp. 36-53.

Ding, B. (2018), "Pharma industry 4.0: literature review and research opportunities in sustainable pharmaceutical supply chains", *Process Safety and Environmental Protection*, Vol. 119, pp. 115-130.

Dobriban, E. and Owen, A.B. (2019), "Deterministic parallel analysis: an improved method for selecting factors and principal components", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, Vol. 81 No. 1, pp. 163-183.

Dubey, R., Gunasekaran, A. and Chakrabarty, A. (2015), "World-class sustainable manufacturing: framework and a performance measurement system", *International Journal of Production Research*, Vol. 53 No. 17, pp. 5207-5223.

Dubey, R., Gunasekaran, A., Childe, S.J., Wamba, S.F. and Papadopoulos, T. (2016), "The impact of big data on world-class sustainable manufacturing", *The International Journal of Advanced Manufacturing Technology*, Vol. 84 Nos 1-4, pp. 631-645.

Dubey, R., Gunasekaran, A., Childe, S.J., Luo, Z., Wamba, S.F., Roubaud, D. and Foropon, C. (2018), "Examining the role of big data and predictive analytics on collaborative performance in context to sustainable consumption and production behaviour", *Journal of Cleaner Production*, Vol. 196, pp. 1508-1521.

Dubey, R., Gunasekaran, A., Childe, S.J., Blome, C. and Papadopoulos, T. (2019a), "Big data and predictive analytics and manufacturing performance: integrating institutional theory, resource-based view and big data culture", *British Journal of Management*, Wiley Online Library, Vol. 30 No. 2, pp. 341-361.

Dubey, R., Gunasekaran, A., Childe, S.J., Fosso Wamba, S., Roubaud, D. and Foropon, C. (2019b), "Empirical investigation of data analytics capability and organizational flexibility as complements to supply chain resilience", *International Journal of Production Research*, Vol. 59 No. 1, pp. 1-19.

Dubey, R., Gunasekaran, A., Childe, S.J., Bryde, D.J., Giannakis, M., Foropon, C., Roubaud, D. and Hazen, B.T. (2020), "Big data analytics and artificial intelligence pathway to operational performance under the effects of entrepreneurial orientation and environmental dynamism: a study of manufacturing organisations", *International Journal of Production Economics*, Vol. 226, 107599.

Eisenhardt, K.M., Graebner, M.E. and Sonenshein, S. (2016), "Grand challenges and inductive methods: rigor without rigor mortis", *Academy of Management Journal*, Vol. 59 No. 4, pp. 1113-1123.

Frank, A.G., Dalenogare, L.S. and Ayala, N.F. (2019), "Industry 4.0 technologies: implementation patterns in manufacturing companies", *International Journal of Production Economics*, Vol. 210, pp. 15-26.

Frizzo-Barker, J., Chow-White, P.A., Mozafari, M. and Ha, D. (2016), "An empirical study of the rise of big data in business scholarship", *International Journal of Information Management*, Elsevier, Vol. 36 No. 3, pp. 403-413.

Gabriel, M. and Pessel, E. (2016), "Industry 4.0 and sustainability impacts: critical discussion of sustainability aspects with a special focus on future of work and ecological consequences", *International Journal of Engineering*, Vol. 14 No. 2, pp. 131-136.

Gandhi, N.S., Thanki, S.J. and Thakkar, J.J. (2018), "Ranking of drivers for integrated lean-green manufacturing for Indian manufacturing SMEs", *Journal of Cleaner Production*, Vol. 171, pp. 675-689.

Gawankar, S.A., Gunasekaran, A. and Kamble, S. (2020), "A study on investments in the big data-driven supply chain, performance measures and organisational performance in Indian retail 4.0 context", *International Journal of Production Research*, Vol. 58 No. 5, pp. 1574-1593.

Geetha, N.K. and Sekar, P. (2018), "An unprecedented multi attribute decision making using graph theory matrix approach", *Engineering Science and Technology, an International Journal*, Vol. 21 No. 1, pp. 7-16.

George, G., Howard-Grenville, J., Joshi, A. and Tihanyi, L. (2016), "Understanding and tackling societal grand challenges through management research", *Academy of Management Journal*, Vol. 59 No. 6, pp. 1880-1895.

Ghadge, A., Er Kara, M., Moradlou, H. and Goswami, M. (2020), "The impact of Industry 4.0 implementation on supply chains", *Journal of Manufacturing Technology Management*, Emerald Publishing, Vol. 31 No. 4, pp. 669-686.

Ghasemaghaei, M. and Calic, G. (2019), "Does big data enhance firm innovation competency? The mediating role of data-driven insights", *Journal of Business Research*, Elsevier, Vol. 104, pp. 69-84.

Giannakis, M. and Louis, M. (2016), "A multi-agent based system with big data processing for enhanced supply chain agility", *Journal of Enterprise Information Management*, Emerald Group Publishing, Vol. 29 No. 5, pp. 706-727.

Gill, N.S. (2021), "10 latest trends in Big Data analytics that you should know in 2021", available at: https://www.xenonstack.com/blog/latest-trends-in-big-data-analytics.

Gobble, M.M. (2013), "Big data: the next big thing in innovation", *Research-Technology Management*, Taylor & Francis, Vol. 56 No. 1, pp. 64-67.

Govindarajulu, N. and Daily, B.F. (2004), "Motivating employees for environmental improvement", *Industrial Management and Data Systems*, Emerald Group Publishing, Vol. 104 No. 4, pp. 364-372.

Grover, S., Agrawal, V.P. and Khan, I.A. (2006), "Role of human factors in TQM: a graph theoretic approach", *Benchmarking: An International Journal*, Vol. 13 No. 4, pp. 447-468, doi: 10.1108/14635770610676290.

Gunasekaran, A. and Spalanzani, A. (2012), "Sustainability of manufacturing and services: investigations for research and applications", *International Journal of Production Economics*, Elsevier, Vol. 140 No. 1, pp. 35-47.

Gunasekaran, A., Papadopoulos, T., Dubey, R., Wamba, S.F., Childe, S.J., Hazen, B. and Akter, S. (2017), "Big data and predictive analytics for supply chain and organizational performance", *Journal of Business Research*, Vol. 70, pp. 308-317.

Guo, Y., Barnes, S.J. and Jia, Q. (2017), "Mining meaning from online ratings and reviews: tourist satisfaction analysis using latent Dirichlet allocation", *Tourism Management*, Elsevier, Vol. 59, pp. 467-483.

Guo, Y., Zhu, Y., Barnes, S.J., Bao, Y., Li, X. and Le-Nguyen, K. (2018), "Understanding cross-product purchase intention in an IT brand extension context", *Psychology and Marketing*, Wiley Online Library, Vol. 35 No. 6, pp. 392-411.

Gupta, M. and George, J.F. (2016), "Toward the development of a big data analytics capability", *Information and Management*, Vol. 53 No. 8, pp. 1049-1064.

Gupta, K., Dangayach, T., Kumar, G. and Gupta, P. (2017), "Disassembly index evaluation of automotive systems using graph theory and AHP", *2017 6th International Conference on*

Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, IEEE, pp. 261-266.

Gupta, S., Altay, N. and Luo, Z. (2019), "Big data in humanitarian supply chain management: a review and further research directions", *Annals of Operations Research*, Springer, Vol. 283 No. 1, pp. 1153-1173.

Gupta, S., Modgil, S. and Gunasekaran, A. (2020), "Big data in lean six sigma: a review and further research directions", *International Journal of Production Research*, Vol. 58 No. 3, pp. 947-969.

Hahn, G.J. and Packowski, J. (2015), "A perspective on applications of in-memory analytics in supply chain management", *Decision Support Systems*, Elsevier, Vol. 76, pp. 45-52.

Hilbert, M. (2016), "Big data for development: a review of promises and challenges", *Development Policy Review*, Wiley Online Library, Vol. 34 No. 1, pp. 135-174.

IBM (2017), *IBM SPSS Statistics for Windows*, IBM Corp, Armonk, NY.

Inamdar, Z., Raut, R., Narwane, V.S., Gardas, B., Narkhede, B. and Sagnak, M. (2020), "A systematic literature review with bibliometric analysis of big data analytics adoption from period 2014 to 2018", *Journal of Enterprise Information Management*, Emerald Publishing. doi: 10.1108/JEIM-09-2019-0267.

Jabbour, C.J.C., Fiorini, P.D.C., Ndubisi, N.O., Queiroz, M.M. and Piato, É.L. (2020), "Digitally-enabled sustainable supply chains in the 21st century: a review and a research agenda", *Science of The Total Environment*, Elsevier, Vol. 725, p. 138177.

Jahanshahi, A.A. and Brem, A. (2017), "Sustainability in SMEs: top management teams behavioral integration as source of innovativeness", *Sustainability*, Multidisciplinary Digital Publishing Institute, Vol. 9 No. 10, p. 1899.

Janssen, M., van der Voort, H. and Wahyudi, A. (2017), "Factors influencing big data decision-making quality", *Journal of Business Research*, Vol. 70, pp. 338-345.

Jensen, J.P. and Remmen, A. (2017), "Enabling circular economy through product stewardship", *Procedia Manufacturing*, Vol. 8, pp. 377-384.

Jiang, H., Chen, Y., Qiao, Z., Weng, T.-H. and Li, K.-C. (2015), "Scaling up MapReduce-based big data processing on multi-GPU systems", *Cluster Computing*, Springer, Vol. 18 No. 1, pp. 369-383.

Julianelli, V., Caiado, R.G.G., Scavarda, L.F. and Cruz, S.P.D.M.F. (2020), "Interplay between reverse logistics and circular economy: critical success factors-based taxonomy and framework", *Resources, Conservation and Recycling*, Vol. 158, p. 104784.

Junge, A.L. (2019), "Digital transformation technologies as an enabler for sustainable logistics and supply chain processes – an exploratory framework", *Brazilian Journal of Operations and Production Management*, Vol. 16 No. 3, pp. 462-472.

Kache, F. and Seuring, S. (2017), "Challenges and opportunities of digital information at the intersection of Big Data analytics and supply chain management", *International Journal of Operations and Production Management*, Vol. 37 No. 1, pp. 10-36.

Kaisler, S., Armour, F., Espinosa, J.A. and Money, W. (2013), "Big data: issues and challenges moving forward", *2013 46th Hawaii International Conference on System Sciences*, pp. 995-1004.

Kalema, B.M.M., Motsi, L. and Motjolopane, I.M. (2016), "Utilizing IT to enhance knowledge sharing for school educators in developing countries", *The Electronic Journal of Information Systems in Developing Countries*, Wiley Online Library, Vol. 73 No. 1, pp. 1-22.

Kamble, S.S. and Gunasekaran, A. (2020), "Big data-driven supply chain performance measurement system: a review and framework for implementation", *International Journal of Production Research*, Vol. 58 No. 1, pp. 65-86.

Kamble, S., Gunasekaran, A. and Dhone, N.C. (2020a), "Industry 4.0 and lean manufacturing practices for sustainable organisational performance in Indian manufacturing companies", *International Journal of Production Research*, Vol. 58 No. 5, pp. 1319-1337.

Kamble, S.S., Gunasekaran, A., Ghadge, A. and Raut, R. (2020b), "A performance measurement system for industry 4.0 enabled smart manufacturing system in SMMEs-A review and empirical investigation", *International Journal of Production Economics*, Elsevier, Vol. 229, p. 107853.

Kamilaris, A., Kartakoullis, A. and Prenafeta-Boldú, F.X. (2017), "A review on the practice of big data analysis in agriculture", *Computers and Electronics in Agriculture*, Vol. 143, pp. 23-37.

Kazancoglu, Y., Ozbiltekin Pala, M., Sezer, M.D., Luthra, S. and Kumar, A. (2021), "Drivers of implementing Big Data Analytics in food supply chains for transition to a circular economy and sustainable operations management", *Journal of Enterprise Information Management*, Vol. ahead-of-print No. ahead-of-print, doi: 10.1108/JEIM-12-2020-0521.

Kent, J. (2020), "Predictive analytics models help plan for COVID-19 demands", available at: https://healthitanalytics.com/news/predictive-analytics-models-help-plan-for-covid-19-demands.

Kim, G.-H., Trimi, S. and Chung, J.-H. (2014), "Big-data applications in the government sector", *Communications of the ACM*, ACM New York, NY, Vol. 57 No. 3, pp. 78-85.

Kumar, P., Singh, R.K. and Kumar, R. (2017), "An integrated framework of interpretive structural modeling and graph theory matrix approach to fix the agility index of an automobile manufacturing organization", *International Journal of System Assurance Engineering and Management*, Vol. 8 No. S1, pp. 342-352.

Kumar, K., Zindani, D. and Davim, J.P. (2019), *Industry 4.0*, Springer Singapore, Singapore, doi: 10.1007/978-981-13-8165-2.

Kumar, N., Kumar, G. and Singh, R.K. (2021a), "Big data analytics application for sustainable manufacturing operations: analysis of strategic factors", *Clean Technologies and Environmental Policy*, Vol. 23, pp. 965-989.

Kumar, P., Singh, R.K. and Kumar, V. (2021b), "Managing supply chains for sustainable operations in the era of industry 4.0 and circular economy: analysis of barriers", *Resources, Conservation and Recycling*, Vol. 164, p. 105215.

Lai, Y., Sun, H. and Ren, J. (2018), "Understanding the determinants of big data analytics (BDA) adoption in logistics and supply chain management", *The International Journal of Logistics Management*, Emerald Publishing, Vol. 29 No. 2, pp. 676-703.

Lamba, K. and Singh, S.P. (2018), "Modeling big data enablers for operations and supply chain management", *The International Journal of Logistics Management*, Emerald Publishing, Vol. 29 No. 2, pp. 629-658.

Lechler, S., Canzaniello, A., Robmann, B., von der Gracht, H.A. and Hartmann, E. (2019), "Real-time data processing in supply chain management: revealing the uncertainty dilemma", *International Journal of Physical Distribution and Logistics Management*, Vol. 49 No. 10, pp. 1003-1019, doi: 10.1108/IJPDLM-12-2017-0398.

Li, G., Huang, D., Sun, C. and Li, Y. (2019), "Developing interpretive structural modeling based on factor analysis for the water-energy-food nexus conundrum", *Science of the Total Environment*, Vol. 651, pp. 309-322.

Liu, Y., Zhang, Y., Ren, S., Yang, M., Wang, Y. and Huisingh, D. (2020), "How can smart technologies contribute to sustainable product lifecycle management?", *Journal of Cleaner Production*, Vol. 249, p. 119423.

Luo, Z., Dubey, R., Gunasekaran, A., Childe, S.J., Papadopoulos, T., Hazen, B. and Roubaud, D. (2017), "Sustainable production framework for cement manufacturing firms: a behavioural perspective", *Renewable and Sustainable Energy Reviews*, Vol. 78, pp. 495-502.

Machado, C.G., Winroth, M.P. and da Silva, E.H.D. (2020), "Sustainable manufacturing in Industry 4.0: an emerging research agenda", *International Journal of Production Research*, Taylor & Francis, Vol. 58 No. 5, pp. 1462-1484.

Maduku, D.K., Mpinganjira, M. and Duh, H. (2016), "Understanding mobile marketing adoption intention by South African SMEs: a multi-perspective framework", *International Journal of Information Management*, Elsevier, Vol. 36 No. 5, pp. 711-723.

Malomo, F. and Sena, V. (2017), "Data intelligence for local government? Assessing the benefits and barriers to use of big data in the public sector", *Policy and Internet*, Vol. 9 No. 1, pp. 7-27.

Mangla, S.K., Raut, R., Narwane, V.S., Zhang, Z.J. and Priyadarshinee, P. (2020), "Mediating effect of big data analytics on project performance of small and medium enterprises", *Journal of Enterprise Information Management*, Emerald Publishing, doi: 10.1108/JEIM-12-2019-0394.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and A.H.B. (2011), *Big Data: the Next Frontier for Innovation, Competition, and Productivity*, McKinsey Global Institute, San Francisco, CA.

Maroufkhani, P., Tseng, M.-L., Iranmanesh, M., Ismail, W.K.W. and Khalid, H. (2020), "Big data analytics adoption: determinants and performances among small to medium-sized enterprises", *International Journal of Information Management*, Elsevier, Vol. 54, p. 102190.

Marr, B. (2021), "The 4 biggest trends in Big Data and analytics right for 2021", available at: https://www.forbes.com/sites/bernardmarr/2021/02/22/the-4-biggest-trends-in-big-data-and-analytics-right-for-2021/?sh=49c499b07df8.

Martínez-Olvera, C. and Mora-Vargas, J. (2019), "A comprehensive framework for the analysis of industry 4.0 value domains", *Sustainability*, Vol. 11 No. 10, p. 2960.

Mazzei, M.J. and Noble, D. (2017), "Big data dreams: a framework for corporate strategy", *Business Horizons*, Vol. 60 No. 3, pp. 405-414.

Mishra, N. and Rane, S.B. (2019), "Prediction and improvement of iron casting quality through analytics and Six Sigma approach", *International Journal of Lean Six Sigma*, Vol. 10 No. 1, pp. 189-210.

Moktadir, M.A., Ali, S.M., Paul, S.K. and Shukla, N. (2019), "Barriers to big data analytics in manufacturing supply chains: a case study from Bangladesh", *Computers and Industrial Engineering*, Vol. 128, pp. 1063-1075.

Muduli, K., Govindan, K., Barve, A. and Geng, Y. (2013), "Barriers to green supply chain management in Indian mining industries: a graph theoretic approach", *Journal of Cleaner Production*, Vol. 47, pp. 335-344.

Nascimento, D.L.M., Alencastro, V., Quelhas, O.L.G., Caiado, R.G.G., Garza-Reyes, J.A., Rocha-Lona, L. and Tortorella, G. (2019), "Exploring Industry 4.0 technologies to enable circular economy practices in a manufacturing context", *Journal of Manufacturing Technology Management*, Vol. 30 No. 3, pp. 607-627.

Nwankpa, J. and Roumani, Y. (2014), "Understanding the link between organizational learning capability and ERP system usage: an empirical examination", *Computers in Human Behavior*, Elsevier, Vol. 33, pp. 224-234.

Oncioiu, I., Bunget, O.C., Türkeş, M.C., Capuşneanu, S., Topor, D.I., Tamaş, A.S., Rakoş, I.-S. and Hint, M.Ş. (2019), "The impact of big data analytics on company performance in supply chain management", *Sustainability*, Multidisciplinary Digital Publishing Institute, Vol. 11 No. 18, p. 4864.

Ozkan-Ozen, Y.D., Kazancoglu, Y. and Kumar Mangla, S. (2020), "Synchronized barriers for circular supply chains in industry 3.5/industry 4.0 transition for sustainable resource management", *Resources, Conservation and Recycling*, Vol. 161, p. 104986.

Pham, T.T., Kuo, T.-C., Tseng, M.-L., Tan, R.R., Tan, K., Ika, D.S. and Lin, C.J. (2019), "Industry 4.0 to accelerate the circular economy: a case study of electric scooter sharing", *Sustainability*, Vol. 11 No. 23, p. 6661.

Pinto, C.A.S., Reis, A.D.C. and Braga, M. (2020), "The supply chain as part of knowledge management in organisational environments", *International Journal of Logistics Systems and Management*, Vol. 36 No. 3, p. 385.

PWC (2020), "PWC", in *Global Annual Review 2020*, available at: https://www.pwc.com/gx/en/about/global-annual-review-2020.html.

Rachinger, M., Rauter, R., Müller, C., Vorraber, W. and Schirgi, E. (2019), "Digitalization and its influence on business model innovation", *Journal of Manufacturing Technology Management*, Vol. 30 No. 8, pp. 1143-1160.

Raj, T. and Attri, R. (2010), "Quantifying barriers to implementing total quality management (TQM)", *European Journal of Industrial Engineering*, Vol. 4 No. 3, p. 308.

Rajput, S. and Singh, S.P. (2019), "Connecting circular economy and industry 4.0", *International Journal of Information Management*, Vol. 49, pp. 98-113.

Raut, R.D., Mangla, S.K., Narwane, V.S., Gardas, B.B., Priyadarshinee, P. and Narkhede, B.E. (2019), "Linking big data analytics and operational sustainability practices for sustainable business management", *Journal of Cleaner Production*, Elsevier, Vol. 224, pp. 10-24.

Ren, S., Zhang, Y., Liu, Y., Sakao, T., Huisingh, D. and Almeida, C.M.V.B. (2019), "A comprehensive review of big data analytics throughout product lifecycle to support sustainable smart manufacturing: a framework, challenges and future research directions", *Journal of Cleaner Production*, Vol. 210, pp. 1343-1365.

Roy, M. and Roy, A. (2019), "Nexus of internet of things (IoT) and big data: roadmap for smart management systems (SMgS)", *IEEE Engineering Management Review*, IEEE, Vol. 47 No. 2, pp. 53-65.

Roy, V. and Singh, S. (2017), "Mapping the business focus in sustainable production and consumption literature: review and research framework", *Journal of Cleaner Production*, Vol. 150, pp. 224-236.

Saengchai, S. and Jermsittiparsert, K. (2019), "Coping strategy to counter the challenges towards implementation of Industry 4.0 in Thailand: role of supply chain agility and resilience", *International Journal of Supply Chain Management*, Vol. 8 No. 5, pp. 733-744.

Saleem, H., Li, Y., Ali, Z., Ayyoub, M., Wang, Y. and Mehreen, A. (2020), "Big data use and its outcomes in supply chain context: the roles of information sharing and technological innovation", *Journal of Enterprise Information Management*, Vol. 34 No. 4, doi: 10.1108/JEIM-03-2020-0119.

Sanders, N.R. (2016), "How to use big data to drive your supply chain", *California Management Review*, SAGE Publications Sage CA: Los Angeles, CA, Vol. 58 No. 3, pp. 26-48.

Saxena, P., Stavropoulos, P., Kechagias, J. and Salonitis, K. (2020), "Sustainability assessment for manufacturing operations", *Energies*, Multidisciplinary Digital Publishing Institute, Vol. 13 No. 11, p. 2730.

Scavarda, A., Daú, G., Felipe Scavarda, L., Duarte Azevedo, B. and Luis Korzenowski, A. (2020), "Social and ecological approaches in urban interfaces: a sharing economy management framework", *Science of The Total Environment*, Vol. 713, p. 134407.

Schull, A. and Maslan, N. (2018), "On the adoption of big data analytics: interdependencies of contextual factors", *Proceedings of the 20th International Conference on Enterprise Information Systems*, Funchal, Madeira, Portugal, Scitepress - Science and Technology Publications, pp. 425-431.

Seth, D., Rehman, M.A.A. and Shrivastava, R.L. (2018), "Green manufacturing drivers and their relationships for small and medium(SME) and large industries", *Journal of Cleaner Production*, Vol. 198, pp. 1381-1405.

Sharma, V.K., Chandna, P. and Bhardwaj, A. (2017), "Green supply chain management related performance indicators in agro industry: a review", *Journal of Cleaner Production*, Vol. 141, pp. 1194-1208.

Sharma, R., Jabbour, C.J.C. and Lopes de Sousa Jabbour, A.B. (2020), "Sustainable manufacturing and industry 4.0: what we know and what we don't", *Journal of Enterprise Information Management*, Emerald Publishing, doi: 10.1108/JEIM-01-2020-0024.

Sheng, J., Amankwah-Amoah, J., Khan, Z. and Wang, X. (2020), "COVID-19 pandemic in the new era of big data analytics: methodological innovations and future research directions", *British Journal of Management*, Vol. 0 Nos 1-20, pp. 1467-8551, 12441.

Shibin, K.T., Gunasekaran, A. and Dubey, R. (2017a), "Explaining sustainable supply chain performance using a total interpretive structural modeling approach", *Sustainable Production and Consumption*, Vol. 12, pp. 104-118.

Shibin, K.T., Gunasekaran, A. and Dubey, R. (2017b), "Flexible sustainable manufacturing via decision support simulation: a case study approach", *Sustainable Production and Consumption*, Vol. 12, pp. 206-220.

Shin, D.-H. (2016), "Demystifying big data: anatomy of big data developmental process", *Telecommunications Policy*, Vol. 40 No. 9, pp. 837-854.

Singh, R.K. and Kumar, P. (2019), "Measuring the flexibility index for a supply chain using graph theory matrix approach", *Journal of Global Operations and Strategic Sourcing*, Vol. 13 No. 1, pp. 56-69.

Singh, R.K. and Kumar, R. (2020), "Strategic issues in supply chain management of Indian SMEs due to globalization: an empirical study", *Benchmarking: An International Journal*, Emerald Publishing, Vol. 27 No. 3, pp. 913-932.

Singh, R.K., Kumar, P. and Chand, M. (2019), "Evaluation of supply chain coordination index in context to Industry 4.0 environment", *Benchmarking: An International Journal*, Vol. 28 No. 5, BIJ-07-2018-0204.

Singh, R.K., Kumar, A., Garza-Reyes, J.A. and de Sá, M.M. (2020), "Managing operations for circular economy in the mining sector: an analysis of barriers intensity", *Resource Policy*, Vol. 69, 101752.

Sivarajah, U., Kamal, M.M., Irani, Z. and Weerakkody, V. (2017), "Critical analysis of Big Data challenges and analytical methods", *Journal of Business Research*, Elsevier, Vol. 70, pp. 263-286.

Stock, T. and Seliger, G. (2016), "Opportunities of sustainable manufacturing in industry 4.0", *Procedia CIRP*, Vol. 40, pp. 536-541.

Sun, S., Cegielski, C.G., Jia, L. and Hall, D.J. (2018), "Understanding the factors affecting the organizational adoption of big data", *Journal of Computer Information Systems*, Taylor & Francis, Vol. 58 No. 3, pp. 193-203.

Tachizawa, E.M., Alvarez-Gil, M.J. and Montes-Sancho, M.J. (2015), "How 'smart cities' will change supply chain management", edited by Gary Graham, D., *Supply Chain Management: An International Journal*, Emerald Group Publishing, Vol. 20 No. 3, pp. 237-248.

Takata, M., Lin, B.L., Xue, M., Zushi, Y., Terada, A. and Hosomi, M. (2020), "Predicting the acute ecotoxicity of chemical substances by machine learning using graph theory", *Chemosphere*, Vol. 238, 124604.

Thakur, V. and Mangla, S.K. (2019), "Change management for sustainability: evaluating the role of human, operational and technological factors in leading Indian firms in home appliances sector", *Journal of Cleaner Production*, Vol. 213, pp. 847-862.

Tuljak-Suban, D. and Bajec, P. (2020), "Integration of AHP and GTMA to make a reliable decision in complex decision-making problems: application of the logistics provider selection problem as a case study", *Symmetry*, Vol. 12 No. 5, p. 766.

Vargas, J.R.C., Mantilla, C.E.M. and de Sousa Jabbour, A.B.L. (2018), "Enablers of sustainable supply chain management and its effect on competitive advantage in the Colombian context", *Resources, Conservation and Recycling*, Elsevier, Vol. 139, pp. 237-250.

Verma, S. (2017), "The adoption of big data services by manufacturing firms: an empirical investigation in India", *Journal of Information Systems and Technology Management*, Vol. 14 No. 1, pp. 39-68.

Verma, S. and Bhattacharyya, S.S. (2017), "Perceived strategic value-based adoption of Big Data analytics in emerging economy", *Journal of Enterprise Information Management*, Emerald Publishing, Vol. 30 No. 3, pp. 354-382.

Wang, Y. and Wiebe, V.J. (2016), "Big Data analytics on the characteristic equilibrium of collective opinions in social networks", *Big Data: Concepts, Methodologies, Tools, and Applications*, IGI Global, Hershey, Pennsylvania, pp. 1403-1420.

Wang, C.J., Ng, C.Y. and Brook, R.H. (2020), "Response to COVID-19 in taiwan", *JAMA*, Vol. 323 No. 14, p. 1341.

Watson, H.J. (2019), "Update tutorial: big data analytics: concepts, technology, and applications", *Communications of the Association for Information Systems*, Vol. 44 No. 1, p. 21.

Wu, K.-J., Liao, C.-J., Tseng, M.-L., Lim, M.K., Hu, J. and Tan, K. (2017), "Toward sustainability: using big data to explore the decisive attributes of supply chain risks and uncertainties", *Journal of Cleaner Production*, Vol. 142, pp. 663-676.

Yacob, P., Wong, L.S. and Khor, S.C. (2019), "An empirical investigation of green initiatives and environmental sustainability for manufacturing SMEs", *Journal of Manufacturing Technology Management*, Vol. 30 No. 1, pp. 2-25.

Yasmin, M., Tatoglu, E., Kilic, H.S., Zaim, S. and Delen, D. (2020), "Big data analytics capabilities and firm performance: an integrated MCDM approach", *Journal of Business Research*, Vol. 114, pp. 1-15.

Ye, L., Pan, S.L., Wang, J., Wu, J. and Dong, X. (2021), "Big data analytics for sustainable cities: an information triangulation study of hazardous materials transportation", *Journal of Business Research*, Vol. 128, pp. 381-390.

Zhang, Y., Ren, S., Liu, Y. and Si, S. (2017), "A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products", *Journal of Cleaner Production*, Vol. 142, pp. 626-641.

Zhou, Y., Xu, L. and Muhammad Shaikh, G. (2019), "Evaluating and prioritizing the green supply chain management practices in Pakistan: based on Delphi and fuzzy AHP approach", *Symmetry*, Vol. 11 No. 11, p. 1346.

About the authors

Narender Kumar is an Assistant Professor at Amity University Gurgaon, Haryana, India. He has done his B. Tech. in mechanical Engineering from Kurukshetra University Kurukshetra and M. Tech in industrial Engineering from Indian Institute of Technology Delhi (IITD), India. He is pursuing Ph. D from Delhi Technological University Delhi (DTU), India. He is having 10 years of experience in teaching at undergraduate and postgraduate level. His areas of interest include Operation research, total quality management and quality assurance, Manufacturing Engineering, Optimization Technique and supply chain management. He is a member of many national/international societies viz. Indian Society for Technical Education (ISTE), World Metal Forum (WMF), American Solar Energy Society (ASES), and Society of automotive engineers (SAE India).

Dr Girish Kumar is working as an Associate Professor in the department of mechanical, production and industrial, automobile engineering at Delhi Technological University. He published many articles in national and International journals. His research interests are Reliability Engineering, Maintenance management, Stochastic modelling, Optimization, Quality Engineering. He is having 15 years of experience in teaching and research.

Dr Rajesh Kr Singh is Professor in Operations Management Area at Management Development Institute (MDI), Gurgaon, India. His areas of interest include Supply chain management, Industry 4.0, Circular Economy. Currently, he is serving as Associate Editor to International Journal of Consumer Studies (Wiley) and International Journal of Global Business and Competitiveness (Springer). He is also in Editorial Board of Journal of Supply Chain Management Systems and Vision-The Journal of Business

Perspective. He has also worked as Guest Editor for special issues of Resources, Conservation and Recycling, Sustainable Production and Consumption, Competitiveness Review, Journal of Modelling in Management, Administrative Sciences, Global Journal of Operations and Strategic Sourcing etc. He has published research papers in reputed international/national journals and conferences. He has published papers in journals such as Resources, Conservation and Recycling, Journal of Cleaner Production, Production planning and Control, Annals of Operations Research, International Journal of Production Research, Applied Soft Computing, Clean Technologies and Environmental Policy, Industrial Management and Data Systems, Singapore Management Review, International Journal of Productivity and Performance Management, Benchmarking: An International Journal, Journal of Modelling in Management, Competitiveness Review: An International Business Journal. Rajesh Kr Singh is the corresponding author and can be contacted at: rajesh.singh@mdi.ac.in

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: **permissions@emeraldinsight.com**

2021 8th International Conference on Signal Processing and Integrated Networks (SPIN) | 978-1-6654-3564-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SPIN52536.2021.9566146

ANN and ANFIS Based Approach for Very Short Term Load Forecasting: A Step Towards Smart Energy Management System

Arshi Khan
Department of Electrical Engineering
Delhi Technological University
Delhi 110042
India
arshik1995@gmail.com

M. Rizwan
Department of Electrical Engineering
Delhi Technological University
Delhi 110042
India
rizwan@dce.ac.in

Abstract: **For planning of new electrical plant, it is crucial that load forecasting needs to be done more precisely. It also helps in balancing the load flow in the power distribution network. In a liberated market it is required for the electric utility to think about the market load interest for creating close to accurate power. Momentary electric burden gauge is critical for the security administration and cost control of the electrical industry, and subsequently it is vital to improve the veracity. Basically, for the forecaster, the main motto is to collect historical data and determine the pattern hidden in it. In this paper, focus is on short term load forecasting which is substantial for controlling and operation in real time of the power system. The proposition is the use of artificial intelligence techniques like artificial neural network and adaptive neuro-fuzzy inference system. Both models are studied for the set of data considered and the results are analysed. Also, the results obtained from both the methods is compared to the desired output and the mean absolute percentage error is calculated.**

Keywords- Artificial Intelligence, Load Forecasting, Neural Network, Adaptive Neuro-Fuzzy Inference System

I. INTRODUCTION

Load forecasting, for the most part alluding to determining power interest and energy, is being utilized all through in all portions of the electric force industry, including transmission, generation, retail and distribution. Utilizations of forecasting in load is spread across power supply transmission, planning and also in distribution side planning of system, power systems operations and maintenance, rate design, demand side management, financial planning and so on. Because of the principal job of load forecasting in the utility business tasks or operations, inaccuracy in the data of forecasting of load may bring about monetary weight to or even liquidation of a service organization. While load estimating gives a critical contribution to control frameworks activities and arranging, imprecise of load forecasting can prompt hardware disappointments or even framework wide power outage. In power system numerous vulnerabilities emerges because of maturing of machines, unexpected burden, variances, misfortunes in transmission lines, voltage and recurrence insecurity, change of climate conditions [1]. In conventional methodologies, the models are planned depending on the relation between load power and factor affecting the load flow. They have an advantage of making a simplified model used for prediction for the purpose of forecasting.

Nonetheless, since the relation is non-linear, it becomes difficult to use such methods for short term load forecasting (STLF). Various methodologies for STLF have been used by researchers in the years. Some methods used for it are autoregressive models (AR) [2], support vector regression (SVR) [3], linear regression [4], fuzzy logic [5,8,9,10,24], Iterative reweighed least-squares, Adaptive load forecasting [6], Stochastic Modelling [7], Neural Networks, Neuro Fuzzy systems [11-16,25,26]. This paper focus on two methods of artificial intelligence (AI) namely, ANN and ANFIS. The result obtained are compared and conclusion is drawn. Section II focus on methodologies used followed by results obtained. Section IV contains the conclusion.

II. METHODOLOGY

The data taken is the average of a location at every 15 minutes. 6 days of load consumed at the location is historically stored and the load demand for the 7th day is to be predicted by using AI methodologies.

A. Artificial Neural Network

It is also known as Neural Network (NN). It is a machine acts like human brain with learning capacity and speculation as its attributes. ANNs make them learn capacities that empower them to deliver better outcomes as more information opens up. It is the establishment of AI and tackles issues that would demonstrate outlandish or troublesome by human or measurable norms. They are basically non-linear mathematical processing networks. They are being used in fields like image recognition, load forecasting, speech recognition, energy consumption prediction, data retrieval, mine dam water level prediction and monitoring [17-20]. Fig.1 shows the basic block diagram of ANN.

B. Adaptive Neuro-Fuzzy Inference System

It is a combination of both ANN and fuzzy logic (FL). Hence, it has advantages of both the methods overcoming their flaws. FL cannot gain any information from the data. ANN have absence of information representability and logic. ANFIS can also be used in applications like PV plants maximum power point tracking, load forecasting, photovoltaics model optimization [21-23].

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:43:21 UTC from IEEE Xplore. Restrictions apply.

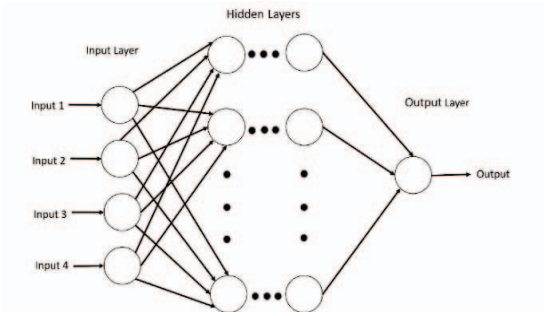


Fig. 1. Basic structure of ANN

Fig. 2 shows the algorithm used while using ANN technique.

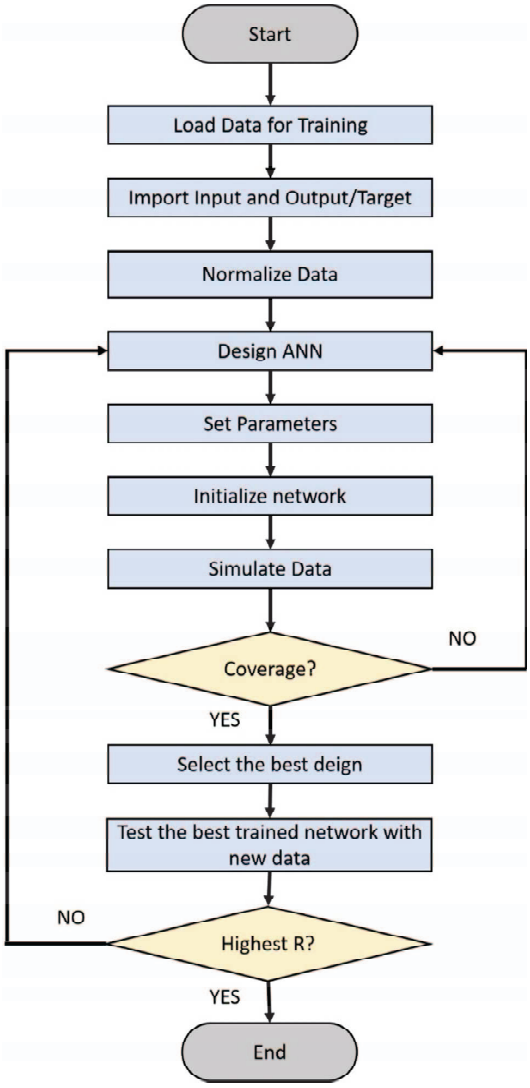


Fig. 2. Flowchart of ANN algorithm

In this paper, the model taken is Takagi-Sugeno. Fig. 3 represents the basic structure of ANFIS. Fig. 4 shows the flowchart for ANFIS.

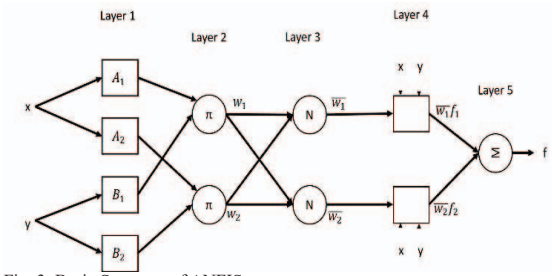


Fig. 3. Basic Structure of ANFIS

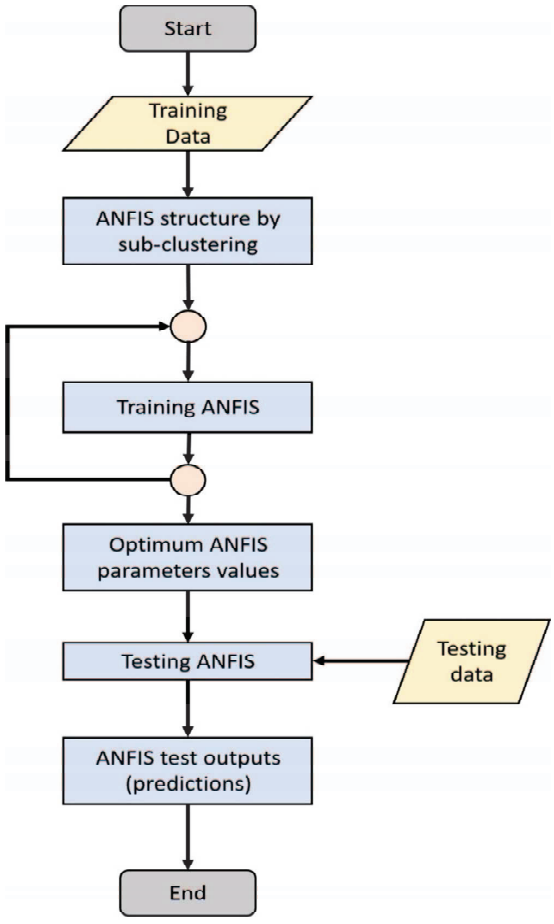


Fig. 4. Flowchart for ANFIS algorithm

III. IMPLEMENTATION AND RESULTS

Using the algorithm discussed through flowchart the forecasted load data for the 7th day for the location is generated through both models of ANN and ANFIS. Table I shows the comparison of results forecasted by both the models and error is calculated.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:43:21 UTC from IEEE Xplore. Restrictions apply.

TABLE I. COMPARING DESIRED WITH FORECASTED LOAD USING ANN AND ANFIS

Time	Actual Load (Day 7)	ANN		ANFIS	
		Forecasted Load	*Error*	*Forecasted Load*	*Error*
00:15	38.7	39.030	0.330	38.692	0.008
00:30	38.7	39.195	0.495	38.648	0.052
00:45	38.7	36.865	1.835	38.690	0.010
01:00	38.7	35.484	3.216	37.717	0.983
01:15	38.7	35.313	3.387	38.250	0.450
01:30	36.88	33.711	3.169	35.052	1.828
01:45	30	32.682	2.682	31.497	1.497
02:00	30	33.792	3.792	33.474	3.474
02:15	30	30.611	0.611	29.997	0.003
02:30	30	30.666	0.666	29.929	0.071
02:45	30	30.649	0.649	29.997	0.003
03:00	30	30.598	0.598	28.394	1.606
03:15	30	27.578	2.422	30.167	0.167
03:30	30	28.388	1.612	30.120	0.120
03:45	30	28.850	1.150	29.741	0.259
04:00	30	29.640	0.360	29.953	0.047
04:15	30	29.999	0.001	30.087	0.087
04:30	30	29.410	0.590	29.958	0.042
04:45	30	29.655	0.345	30.030	0.030
05:00	30	28.240	1.760	29.932	0.068
05:15	30	29.648	0.352	30.375	0.375
05:30	30	30.008	0.008	29.583	0.417
05:45	30	30.698	0.698	30.090	0.090
06:00	30	30.033	0.033	29.900	0.100
06:15	28.28	29.110	0.830	28.330	0.050
06:30	27.7	27.777	0.077	27.715	0.015
06:45	27.7	27.433	0.267	27.694	0.006
07:00	33.13	33.460	0.330	33.131	0.001
07:15	33.66	33.498	0.162	33.645	0.015
07:30	33.9	34.665	0.765	33.916	0.016
07:45	35.05	35.080	0.030	35.087	0.037
08:00	37.66	35.745	1.915	37.644	0.016
08:15	37.86	38.816	0.956	37.872	0.012
08:30	39.67	39.711	0.041	39.662	0.008
08:45	42.48	41.662	0.818	42.503	0.023
09:00	43.79	42.986	0.804	43.784	0.006
09:15	39.79	39.825	0.035	39.789	0.001
09:30	44.12	44.994	0.874	44.126	0.006
09:45	46.3	48.666	2.366	46.298	0.002
10:00	52.83	53.457	0.627	52.829	0.001
10:15	55.35	55.063	0.287	55.351	0.001
10:30	57.13	57.663	0.533	57.128	0.002
10:45	57.9	56.791	1.109	57.900	0.000
11:00	57.56	54.661	2.899	57.559	0.001
11:15	57.52	56.361	1.159	57.523	0.003
11:30	57.8	56.948	0.852	57.798	0.002
11:45	57.8	57.339	0.461	57.798	0.002
12:00	57.31	60.004	2.694	57.31	0
12:15	56.41	57.122	0.712	56.411	0.001
12:30	55.3	55.246	0.054	55.297	0.003
12:45	52.92	52.248	0.672	52.936	0.016
13:00	51.82	51.581	0.239	51.809	0.011
13:15	50.2	49.417	0.783	50.199	0.001
13:30	49.78	52.094	2.314	49.784	0.004
13:45	50.2	52.136	1.936	50.184	0.016
14:00	50.03	51.896	1.866	50.058	0.028
14:15	51.63	53.495	1.865	51.623	0.007
14:30	55.9	55.39	0.51	55.892	0.008
14:45	55.9	57.1	1.2	55.898	0.002
15:00	54.58	56.192	1.612	54.581	0.001
15:15	53.2	53.593	0.393	53.199	0.001
15:30	51.9	51.818	0.082	51.898	0.002
15:45	49.46	49.164	0.296	49.461	0.001
16:00	45.74	45.368	0.372	45.739	0.001
16:15	36.8	39.68	2.88	27.372	9.428
16:30	32.8	31.401	1.399	42.8	10
16:45	31.04	31.972	0.932	-14.501	45.541
17:00	30.7	30.191	0.509	-38.019	68.719
17:15	32.47	32.091	0.379	23.178	9.292
17:30	32.23	32.299	0.069	2.918	29.312
17:45	32.9	32.573	0.327	-29.795	62.695
18:00	32.9	31.899	1.001	-22.324	55.224
18:15	31.47	32.114	0.644	7.26	24.21
18:30	30.73	28.566	2.164	31.263	0.533
18:45	28.97	18.848	10.122	34.072	5.102
19:00	28.9	31.564	2.664	88.65	59.75
19:15	29.7	30.009	0.309	257.177	227.477
19:30	29.87	31.297	1.427	160.565	130.695
19:45	31	30.559	0.441	40.457	9.457
20:00	31.82	31.683	0.137	2.216	29.604
20:15	31.37	30.685	0.685	9.957	21.413
20:30	29.8	31.309	1.509	6.78	23.02
20:45	30.85	32.386	1.536	19.106	11.744
21:00	36.58	36.371	0.209	11.697	24.883
21:15	38.25	38.418	0.168	11.501	26.749
21:30	39.9	39.125	0.775	9.122	30.778
21:45	39.9	38.917	0.983	13.824	26.076
22:00	40.1	39.984	0.116	52.131	12.031
22:15	40	39.12	0.88	60.142	20.142
22:30	40	39.968	0.032	41.733	1.733
22:45	37.97	37.585	0.385	16.837	21.133
23:00	37.7	37.414	0.286	18.239	19.461
23:15	37.7	36.921	0.779	20.684	17.016
23:30	37.7	36.364	1.336	24.243	13.457
23:45	36.87	36.649	0.221	26.701	10.169
00:00	35.7	35.937	0.237	21.034	14.666

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:43:21 UTC from IEEE Xplore. Restrictions apply.

Table II shows the average absolute relative error calculated for both ANN and ANFIS methods used in this paper. Absolute relative error can be calculated using the formula given below:

$$ARE = \frac{P_{desired} - P_{forecasted}}{P_{desired}} \times 100 \quad (2)$$

TABLE II. AVERAGE FORECASTING ERROR

Criterion	ANN	ANFIS
AVG. %ARE	2.913	1.953

Fig. 5 represents the regression plot obtained during ANN model training and testing. Fig. 6 is comparing of actual load anf forecasted load from ANN model. Fig. 7 is the error plot obtained by ANN.

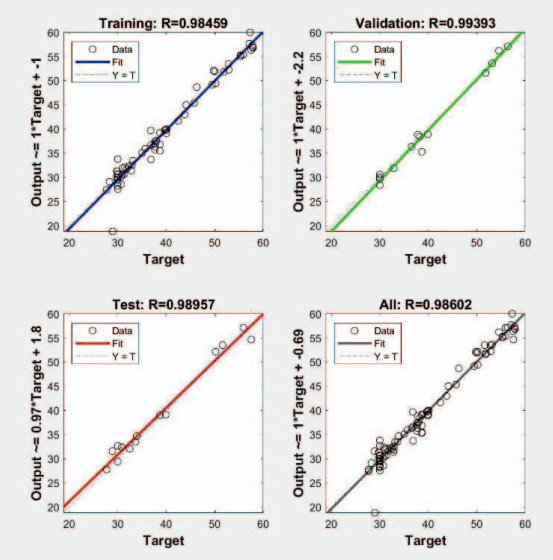


Fig. 5. Regression plot while implementing ANN model

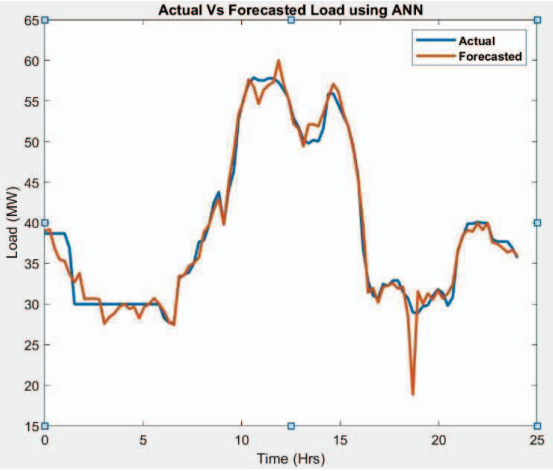


Fig. 6. Actual Vs Forecasted Load through ANN

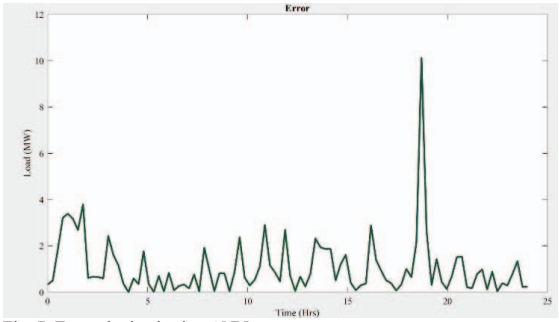


Fig. 7. Error obtained using ANN

In the course of the preparation cycle for the ANFIS structure, rules are executed to be determined the output of the model. The primary layer of the design is called fuzzification. In the subsequent layer, the heaviness of each standard must be registered through a fuzzy AND operation. In the layer 3, it is made the standardization of the qualities and in the layer 4 the defuzzification interaction. At last in layer 5, the general output of the system is acquired. Fig. 8 shows the comparision of actual and forecasted load obtained by ANFIS. Fig. 9 represents the error plot for ANFIS.

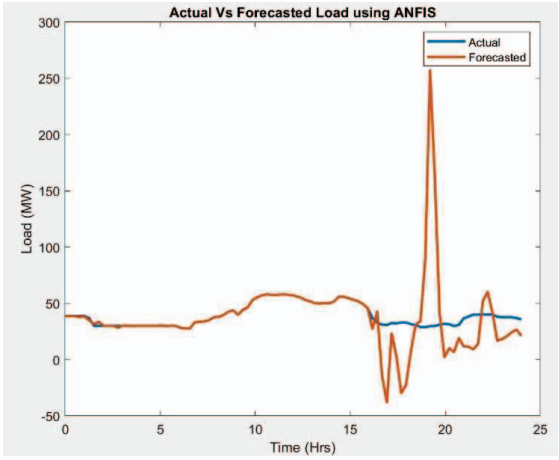


Fig. 8. Actual Vs Forecasted Load through ANFIS

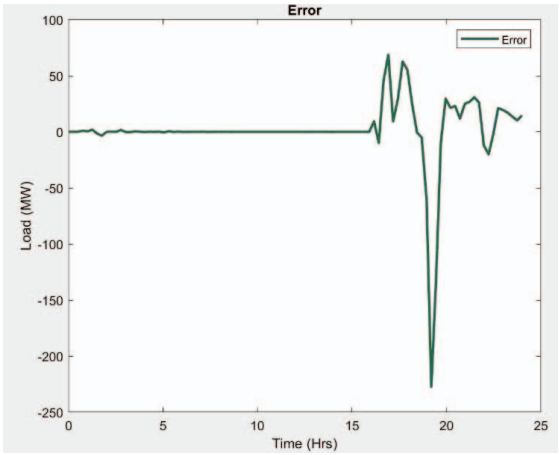


Fig. 9. Error obtained using ANFIS

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:43:21 UTC from IEEE Xplore. Restrictions apply.

IV. CONCLUSION

Efficiency in load forecasting is the issue that needs to be considered as important issue while planning. This paper proposes AI technology to be used for forecasting and also their effectiveness. The study of comparision is conducted to validate the performance of each approach. Error for each method is calculated and then compared. It can be seen that average ARE while using ANN is 2.913% and through ANFIS is 1.953%. It can be concluded that ANFIS model is more efficient than ANN model. So, it can be said that ANFIS model is better than ANN model and it performs with greater accuracy than ANN.

REFERENCES

[1] S. C. Pandian, K. Duraiswamy, C. C. A. Rajan, N. Kanagaraj, "Fuzzy approach for short term load forecasting", Electrical Power System Research, vol. 76, pp. 541-548, 2006.

[2] J. W. Taylor, "Short-term electricity demand forecasting using double seasonal exponential smoothing," J. Oper. Res. Soc., vol. 54, no. 8, pp. 799–805, 2003.

[3] E. E. Elattar, J. Goulermas, and Q. H.Wu, "Electric load forecasting based on locally weighted support vector regression," IEEE Trans. Syst., Man, Cybern., C, vol. 40, no. 4, pp. 438–447, Jul. 2010.

[4] K.-B. Song, Y.-S. Baek, D. H. Hong, and G. Jang, "Short-term load forecasting for the holidays using fuzzy linear regression method," IEEE Trans. Power Syst., vol. 20, no. 1, pp. 96–101, Feb. 2005.

[5] M. Rejc and M. Pantos, "Short-term transmission-loss forecast for the slovenian transmission power system based on a fuzzy-logic decision approach," IEEE Trans. Power Syst., vol. 26, no. 3, pp. 1511–1521, Aug. 2011.

[6] Regression Versus Box Jenkins (Time Series Analysis) case study 2005.

[7] D.W. Bunn, E.D. "Comparative Models for Electrical Load Forecasting" Tohn Wily & Sons 1985.

[8] Kevin M. Passino and Stephen Yurkovich, Fuzzy Control, First Edition, Addison Wesley Longman, Inc., California, 1998.

[9] Zimmermann, H.-J. (1993). Fuzzy set theory - and its applications , second edn, Kluwer, Boston. (1. ed. 1991).

[10] S.J. Kiartzis and A.G. Bakirtzis. A Fuzzy Expert System for Peak Load Forecasting: Application to the Greek Power System. Proceedings of the 10th Mediterranean Electrotechnical Conference, 3: 1097- 1100, 2000.

[11] Jacek M. Zurada, Introduction to Artificial Neural Systems, First Edition, PWS Publishing Company, Boston, 1995.

[12] Jyh-Shing R. Jang, Chuen-Tsai Sun, Eiji Mizutani, Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence, First Edition, Prentice-Hall, Inc., Viacom Company Upper Saddle River, USA, 1997.

[13] H. Khorashadi Zadeh and Zuyi Li, "Transmission Line Distance Protection Using ANFIS and Positive Sequence Components", Proceedings of the iREP Symposium- Bulk Power System Dynamics and Control - VII, Charleston, SC, USA, 2007.

[14] M. Jayabharata Reddy and D.K. Mohanata, "Performance Evaluation of Adaptive Network Based Fuzzy Inference System Approach for Location of Faults on Transmission Lines Using Monte Carlo Simulation", IEEE Transactions on Fuzzy Systems, Volume 16, Issue 4, pp. 909 - 919, Aug. 2008.

[15] R C Bansal, "Overview and Literature Survey of Artificial Neural Networks Applications to Power Systems (1992-2004)", IE (I) Journal-EL, Vol. 86, pp. 282 - 296, March 2006.

[16] A. Seif; "Artificial Intelligence Based Approach Compared With Stochastic Modeling For Electrical Load Forecasting"; M. Sc Dissertation, Office of Graduate Studies, Faculty of Engineering, Cairo University, 2011.

[17] H. D. Block, ``The perceptron: A model for brain functioning," Rev. Modern Phys., vol. 34, no. 1, p. 123, Jan. 1962.

[18] A. N. Hasan, B. Twala, and T. Marwala, ``Predicting mine dam levels and energy consumption using arti_cial intelligence methods," in Proc. IEEE Symp. Comput. Intell. Eng. Solutions, Singapore, Apr. 2013, pp. 171_175.

[19] A. Ali and A. N. Hasan, ``Improving single classi_ers prediction accuracy for underground water pump station in a gold mine using ensemble techniques," in Proc. IEEE Int. Conf. Comput. Tool (EUROCON), Sep. 2015, pp. 1-7.

[20] L. Marwala and B. Twala, ``Univariate modelling of electricity consumption in South Africa: Neural networks and neuro-fuzzy systems," in Proc.n IEEE Int. Conf. Syst., Man Cybern., Oct. 2013, pp. 2238_2243.

[21] L. Marwala and B. Twala, "Forecasting electricity consumption in South Africa: ARMA, neural networks and neuro-fuzzy systems", in Proc. Int. Joint Conf. Neural Netw., Beijing, China, Jul. 2014, pp. 3049-3055.

[22] A. Ali and A. N. Hasan, ``Optimization of PV model using fuzzy-neural network for DC-DC converter systems," in Proc. 9th Int. Renew. Energy Congr. (IREC), Hammamet, Tunisia, Mar. 2018, pp. 1-6.

[23] A. M. Farayola, A. N. Hasan, and A. Ali, ``Curve _tting polynomial technique compared to ANFIS technique for maximum power point tracking," in Proc. 8th Int. Renew. Energy Congr. (IREC), Mar. 2017, pp. 1-6.

[24] D. Karna, A. Vikram, A. Kumar and M. Rizwan, "A Novel Fuzzy based Intelligent Demand Side Management for Automated Load Scheduling", 4th International Conference on Green Energy and Applications (ICGEA), 2020, pp. 171-175.

[25] A. Kumar, M. Rizwan and U. Nangia, "Artificial Neural Network based Model for Short Term Solar Radiation Forecasting considering Aerosol Index", 2018 2nd IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES), 2018, pp. 212-217.

[26] U. Singh and M. Rizwan, "Analysis of Fuzzy Logic, ANN and ANFIS based Models for the Forecasting of Wind Power", 2018 2nd IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES), 2018, pp. 1-7.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:43:21 UTC from IEEE Xplore. Restrictions apply.

J. Inst. Eng. India Ser. B
https://doi.org/10.1007/s40031-021-00685-4

ORIGINAL CONTRIBUTION

Applications of Adaptive Long Short-Term Memory to Active Filtering

Alka Singh[1]

Received: 13 May 2021 / Accepted: 30 September 2021
© The Institution of Engineers (India) 2021

Abstract Power distribution systems rely on the use of active shunt compensators for power quality (PQ) improvement. This paper proposes the application of a new technique based on Long Short-Term Memory (LSTM) to mitigate PQ problems and achieve shunt compensation. The proposed controller comprises a single LSTM cell and the weights of the LSTM are updated online. The output of the LSTM is trained to determine correctly component of load current corresponding to 50 Hz, which is then used for active filtering and compensation. A Simulink-based MATLAB model is developed for the grid connected power distribution network feeding nonlinear loads. A single-phase voltage source converter configured as H-bridge is controlled as a compensator. Additionally, a photovoltaic source of 5 kW rating is integrated at the DC link of this converter. In both the cases discussed, the proposed LSTM network is found to be successfully track the fundamental load current component and utilize the same for compensation. Simulation results justify the ability of the controller developed using LSTM in achieving power factor correction of supply current and reactive and active power injection.

Keywords Compensation · Power quality · Photovoltaic · Neural network · Active filter

Abbreviations

v_s	Supply voltage (V)
i_s	Grid current (A)
i_l	Load current (A)
i_c	Compensator current (A)
P_{pv}	Active power of PV array (kW)
V_{pv}	PV array voltage (V)
I_{pv}	Current injected by PV array (A)
V_{dc}	DC bus voltage (V)
V_{dcref}	Reference voltage at the DC bus (V)
C_{dc}	Capacitance of the inverter (μF)
R_{L1}	Source resistance (Ω)
L_{L1}	Source inductance (mH)
L_f	Filter inductance (mH)
L_b	Inductor of the boost converter (mH)
C_b	Capacitor of the boost converter (mH)
x_t	Input signal
y_t	Output signal
d_t	Desired output (A)
d_{eff}	Effective output (A)
d_{pv}	PV contribution (A)
u_p	Unit template
W	Weight (A)
μ	Learning rate
P, Q	Real and reactive powers (W, var

Introduction

The importance of power quality monitoring, assessment and improvement has acquired new dimension in today's world. Developing countries are also focusing on these aspects due to economic benefits of improved power quality. It is a well-known fact that poor quality of power amounts to financial losses. Hence, the emphasis on new

✉ Alka Singh
alkasingh.eed@gmail.com

[1] Department of Electrical Engineering, Delhi Technological University, New Delhi, India

Published online: 21 October 2021

compensation devices, their control and implementation is carried out by several researchers throughout the world [1].

Shunt compensation [1, 2] is an effective method to improve PQ of power distribution system. It has the ability to correct the power-factor of the supply current and inject the needed reactive power such that the grid is not burdened to supply additional reactive power. Moreover, a PV source can be easily integrated at the compensator ends to inject active power also. This paper discusses both the compensation cases—without and with PV source integrated at the compensator.

Several conventional shunt compensators have been referred in the literature for one-phase, three-phase three-wire and four wire systems. A number of control algorithms have been designed and discussed for these compensators. The literature [2] mentions an Adaline-based adaptive controller and [3] discusses several techniques such as Instantaneous Reactive Power Theory (IRPT), Synchronous Reference Frame Theory (SRFT) and Perfect Harmonic Cancellation method for the control of active filter. The researches [4, 5] discuss the use of multivariable filter and finite impulse response-based filters for extracting the fundamental load current component. Further, the performance of active filter under non-ideal and distorted grid conditions has been discussed in the literature [6, 7] using an additional frequency locked loop. The researchers [8] have discussed the use of bacterial foraging-based proportional integral (PI) controller to optimize the controller gains for a faster dynamic performance.

Several papers have focused on the application of neural networks (NN) in achieving shunt compensation [9–14]. The researchers [10] have proposed a two layered NN for the control of active shunt filter in 1-phase system. Several authors have mentioned different NN techniques such as conductance estimation using NN [11], Chebyshev expansion-based ANN [12] and Recurrent neural network (RNN) for active power filter [13, 14]. In the literature [13, 14] Fai etal have depicted the use of two hidden layers for RNN and used it for purpose of compensation.

Apart from shunt compensation, RNN finds use in forecasting and prediction, image processing and speech recognition. While RNN uses recent past information for estimation purposes, it has certain drawbacks [15]. The two major problems associated with RNN are the vanishing gradient problem when the successive gradient becomes extremely small enough or the gradient may become extremely large enough and explode. Thus, RNN cannot be used for long-term predictions and new modifications in the form of LSTM have been suggested. The main advantage of a LSTM over RNN is the presence of memory gate, which allows it to learn or forget conveniently [15].

LSTM has conventionally been employed in forecasting techniques [16, 17] for electrical load, wind speed prediction, solar power forecasting, etc. Further, all the applications of LSTM discussed in papers use several LSTM cells and layers are placed one after the other interactively. The training and testing sets have huge memory requirement and the network needs to be sufficiently trained first to be used for varied applications.

Recently some authors have suggested modifications to the conventional LSTM employing use of different activation functions [17]. However, the online training of LSTM weights has not been reported in the literature till date. Moreover, the applications of LSTM for control of active shunt filter are also not discussed in the literature.

The objective of this paper is to focus on two aspects viz. present a new, simple and effective structure of LSTM. It realizes a single LSTM cell and all the weights are updated online. The training of LSTM network is carried out using Least Means Square (LMS) approach which is a simple, effective and well-known adaptive filtering technique. The second aspect of the paper focuses on the application of proposed LSTM for shunt compensation. Although different NN techniques have been developed and applied for achieving compensation; however, the use of LSTM in solving PQ problems is not reported yet and is investigated in this paper.

The main highlights of this work are:

i. The present work proposes the application of a single LSTM cell effectively designed for control of PQ problems.
ii. The proposed A-LSTM has adaptive weights which are trained online using LMS technique.
iii. A-LSTM-based controller is developed to track the fundamental load current and use it for compensation.
iv. Results are presented for A-LSTM controller for mitigation of PQ problems in single phase system.
v. The results are presented for two cases—in the absence and presence of PV source.
vi. An appropriate comparison is presented w.r.t conventional PQ control techniques and gives encouraging results.

A-LSTM technique is employed for shunt compensation and two cases are discussed in detail. The first case relates to shunt compensation in a 1-phase grid connected system feeding nonlinear loads at the point of common coupling (PCC). The second case pertains to an additional 5 kW PV source connected to the compensator via a boost converter. The detailed modeling of PV, maximum power point tracking, design of boost converter are not discussed in the paper as some good papers are already available on these topics in the literature. Instead, the focus is on developing a modified LSTM structure and how it can be applied suitably for achieving shunt compensation. The

implementation procedure of A-LSTM technique along with detailed investigations for solving PQ problems are reported.

System Configuration

The system considered in the paper is a single phase power distribution network feeding nonlinear loads as shown in Fig. 1. Nonlinear loads are intentionally connected at the load bus since these loads inject harmonics into the supply system and cause several other power quality problems. A single phase H-bridge configuration is selected as shunt compensator. Provision is also made to switch over a PV source of 5 kW rating at the DC link of the shunt compensator as depicted in Fig. 1. All the design values of the system parameters including the boost converter are mentioned in Appendix.

Mathematical Formulation

This section discusses the conventional and the proposed Adaptive LSTM structure. Thereafter, the application of A-LSTM technique to achieve shunt compensation is discussed.

Basic LSTM Structure

Conventional LSTM configuration for a single unit is shown in Fig. 2. Literature review suggests the complete LSTM comprises a number of such interconnected units, each one placed one after the other. Hence, the information from one unit is passed onto the next. Each of these units or cells comprises several gates such as input gate (i_t), output gate (o_t), forget gate (f_t) and cell state (c_t). The forget gate decides which value to retain sor forget. Appropriate activation functions viz. sigmoid and tanh are selected. The basic underlying equations for a conventional LSTM are

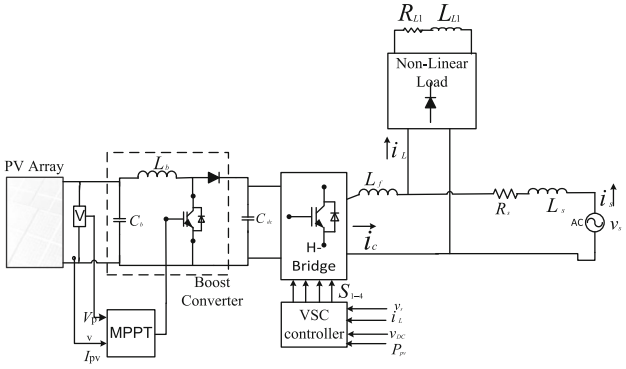


Fig. 1 Schematic diagram of the considered system

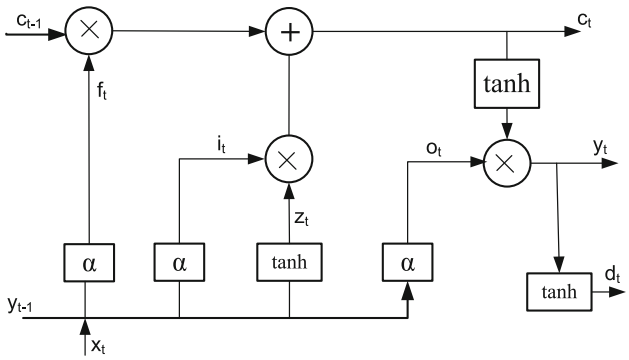


Fig. 2 Conventional LSTM block

$$z_t = \tanh(W_x[x_t; y_{t-1}]) \quad (1)$$

$$i_t = \alpha(W_i[x_t; y_{t-1}]) \quad (2)$$

$$f_t = \alpha(W_f[x_t; y_{t-1}]) \quad (3)$$

where f_t represents the output of forget gate, i_t the input gate and z_t is an intermediate signal marked on the diagram, x_t and y_{t-1} represent the current input and the previous output vector. W_x, W_i and W_f represent the weights to be updated. The new cell state, c_t is computed using c_{t-1} which is the previous cell state and the output of the gates computed above. Element wise multiplication is indicated using the symbol '.'

$$c_t = i_t \cdot z_t + f_t \cdot c_{t-1} \quad (4)$$

The output gate is represented as o_t and involves the updation of weight W_o and α denotes the sigmoidal activation function.

$$o_t = \alpha(W_o[x_t; y_{t-1}]) \quad (5)$$

The output vector y_t is computed using the output gate o_t and tanh activation function

$$y_t = o_t \cdot \tanh(c_{t-1}) \quad (6)$$

Finally, the estimated output (d_t) can be obtained using y_t and the tanh activation function and involves the weight W_d

$$d_t = \tanh(W_d[y_t]) \quad (7)$$

Equations (1)–(7) can be depicted in the form of a block diagram shown in Fig. 2.

Adaptive LSTM (A-LSTM)

This paper proposes a single LSTM unit with adaptive weights. Instead of using multiple LSTM units connected one after the other, the proposed structure is simpler. It is made even more effective by updating the weights online. The Least Means Square (LMS) approach is used for weight updation as discussed in Fig. 3.

Depending on the desired output (d_t), the unit template (u_p) and the computed error signal (e_t), the weights are trained adaptively. The objective of the designed LSTM network is to compute the parameter d_t which is load current component corresponding to 50 Hz. Accordingly, the desired output is subtracted from the target weight component. The target weight is the fundamental value of load current extracted from the input signal x_t and appropriate gain constant. As shown in Fig. 3, the error between the two is computed to be e_t which along with the unit template u_p and the learning rate (μ) are used to update all the weights viz. W_f, W_i, W_z, W_o and W_d. The underlying equation for LMS for single weight updation is shown below

$$W(k+1) = W(k) + \mu e_t(k) u_p(k) \tag{8}$$

It is easily observed that the weight updation is highly dependent on the learning rate, μ which needs to be suitably selected.

The implementation procedure of A-LSTM is presented in Fig. 4. The weights and the biases are initialized and proposed A-LSTM controller is modeled using Eqs. (1)–(7). LMS technique is used to update all the weights using Eq. (8) till convergence is obtained. Thereafter the tracked fundamental load component (d_t) is used further to design the controller for achieving compensation.

Controller Design

The control algorithm is now developed with the proposed A-LSTM technique for PQ improvement. The developed technique is tested for achieving shunt compensation with/without PV source and depicted in Fig. 5. It is important to maintain the DC link voltage (V_{dc}) to its reference value of V_{dcref} for which a proportional integral (PI) regulator is employed. The d_{loss} represents the output of this controller. It is indicative of the switching losses in the compensator. The equation for estimating d_{loss} is

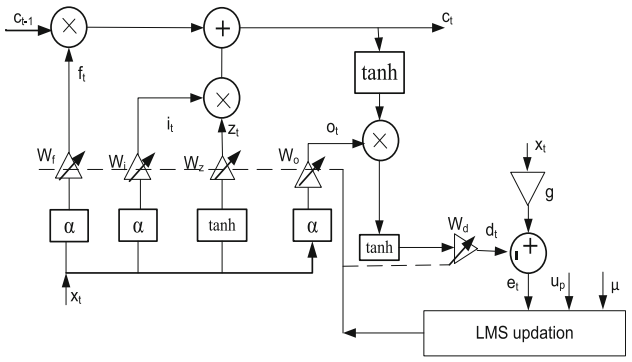


Fig. 3 Schematic layout of proposed A-LSTM technique

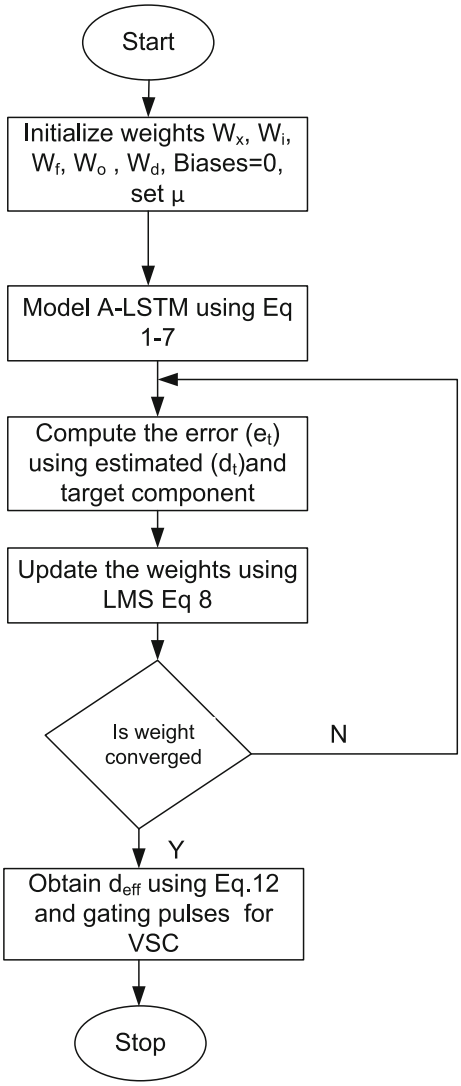


Fig. 4 Flowchart of the proposed controller

$$d_{loss}(k) = d_{loss}(k-1) + k_d[e_{dc}(k) - e_{dc}(k-1)] + k_i e_{dc}(k) \tag{9}$$

where k_d and k_i depict the proportional gain and integral gain of the PI controller, respectively. The error, e_{dc} is computed from the reference and actual V_{dc} as

$$e_{dc}(k) = V_{dcref}(k) - V_{dc}(k) \tag{10}$$

The controller is developed such that the Q demand (reactive power) of the load is met entirely by the compensator. The P demand (active power) of the load and the switching losses are met by the electric grid. However, if a PV source is interfaced at the converter end, the renewable source meets the load demand partially as per its capability. Extra active power may also be injected into the grid, if the load demand is quite less. The

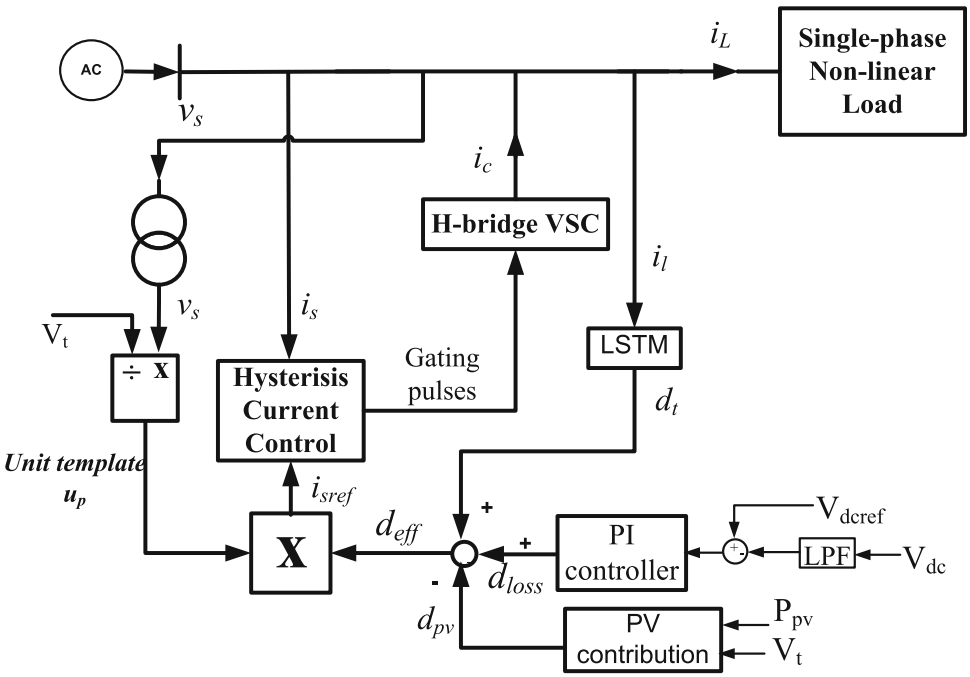


Fig. 5 Schematic layout of the developed controller

contribution of the PV source is included by feed-forward term (d_{pv}) computed as:

$$d_{pv} = \frac{P_{pv}}{V_t} \tag{11}$$

where P_{pv} represents the active power injected by the PV source and V_t denotes the voltage at load bus. The effective weight component (d_{eff}) is computed as

$$d_{eff} = d_t + d_{loss} - d_{pv} \tag{12}$$

The entire control scheme layout with PV source contribution is depicted in Fig. 5. Once the effective weight component is determined, the reference supply current (i_{sref}) is obtained by multiplying d_{eff} with unit template u_p.

$$i_{sref} = d_{eff} * u_p \tag{13}$$

Finally, hysteresis current controller is realized over the reference and actual supply currents (i_{sref} and i_s) to generate the required gating pulses for the compensator.

Results and Discussion

This Section discusses results of the designed controller without PV source followed by results with 5 kW PV source interfaced at the converter end. Figure 6 shows the fundamental weight extracted using the A-LSTM technique under load variations at $t = 0.4$ s and $t = 0.6$ s. Target weight is specified in Ampere and the updated weights tracked with different learning rates $\mu = 0.01$, 0.001 and 0.0001 are shown. It is clear that $\mu = 0.0001$ represents the slowest learning and hence it takes ~ 5 cycles to achieve tracking. Learning with $\mu = 0.01$is quite fast but ripples are observed in the tracked weight. Hence, $\mu = 0.001$ is the optimized learning rate selected for simulation purpose and results.

Without PV Integration

Figure 7 shows the results for shunt compensation and power quality improvement. This figure shows the plots for v_s, i_l, i_s, i_c and v_{dc}. The considered load is nonlinear and load current is distorted. However, the compensator injects appropriate current so as to nullify the harmonics in i_l. Hence, the supply current is quite close to sinusoidal and the PQ of the grid is significantly improved. At $t = 0.4$ s and $t = 0.6$ s, the load is increased. However, the grid current is sinusoidal and it is in phase with the supply voltage. Moreover, the voltage at the dc terminals of the compensator is maintained to the reference value of 400 V. The PI controller realized on the DC terminals of the converter helps to attain this voltage regulation.

Figure 8 shows the signal and harmonic distortion of the supply current reduced to 1.69% and it is much below the allowed limits set by IEEE-519 standard.

With PV Integration

Figure 9 shows the results for shunt compensation and power quality improvement in the presence of 5 kW PV source. The figure shows the plots for v_s, i_l, i_s, i_c and v_{dc}

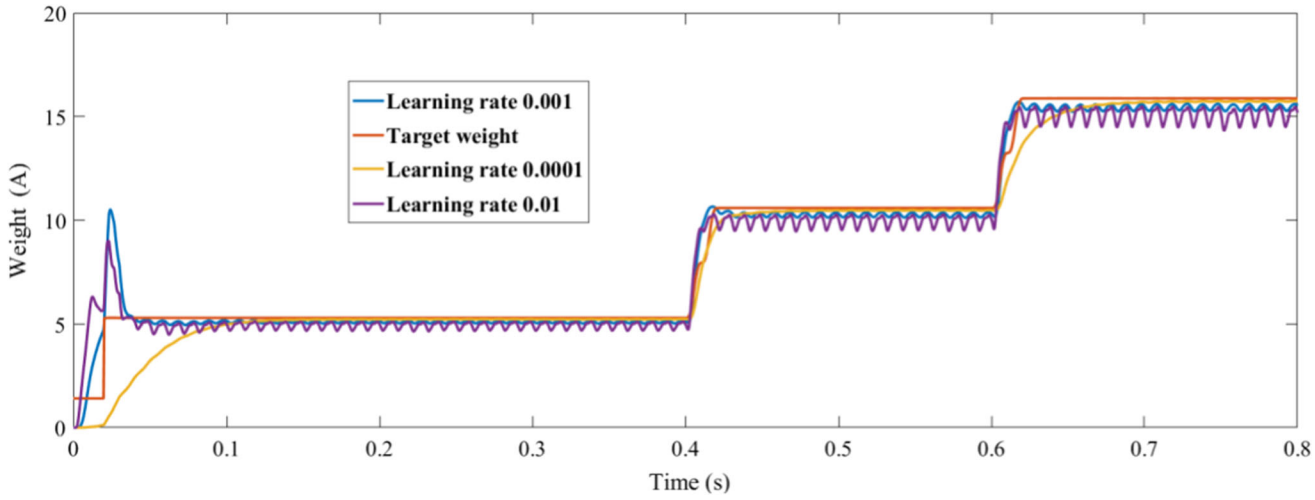


Fig. 6 Variation of fundamental weight with change in learning rate

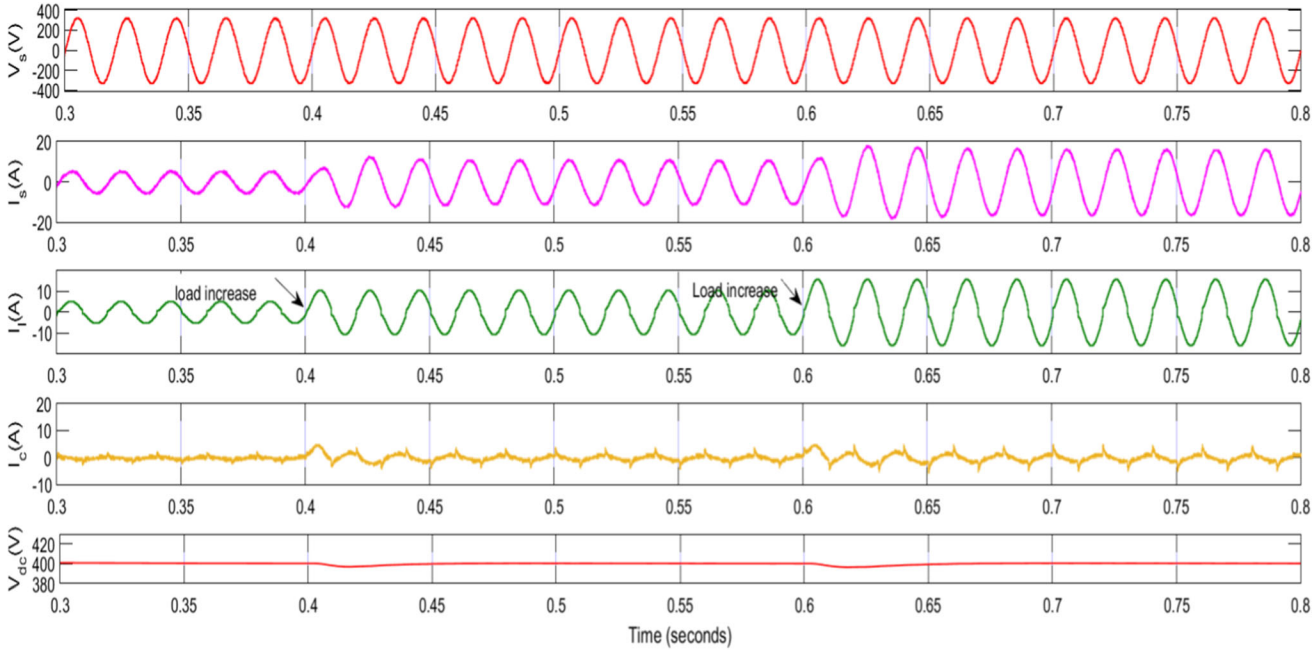


Fig. 7 System performance without PV integration

under similar load variations. The real and reactive power share for the supply, load and compensator are also depicted in Fig. 10.

It is observed from Fig. 9 that there is a considerable phase shift between the voltage and current at the supply end. The two signals are clearly out of phase due to active power injection by PV. Also, the supply current is observed to be sinusoidal even with nonlinear loads. Load increase is performed at $t = 0.4$ s and $t = 0.6$ s. Thus, power quality is improved along with active power injection in the presence of PV source also. Figure 10 shows the respective contributions of real and reactive power of the load, compensator with PV interfaced and the grid. It is observed that the excess PV power is injected to the grid as the load demand is quite less. The load demand is $\sim$ 1 kW while the power injected by PV source is $\sim$ 5 kW.

It is observed that net surplus active power generated by the PV source is fed to the grid. The controller is designed such that the Q demand of the load is completely met by compensator. At $t = 0.4$ s and $t = 0.6$ s, load power is increased, yet P and Q power balance is achieved between the load, compensator and grid.

The performance of the developed technique is tested for weight extraction and compared with Synchronous Reference Frame (SRF) Theory and Second Order Generalized Integrator (SOGI) technique as shown in Fig. 11.

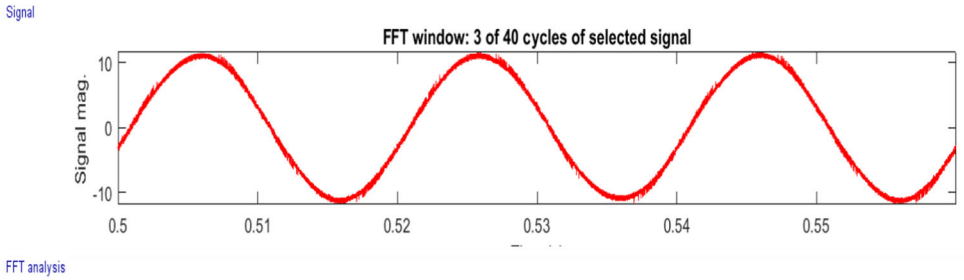
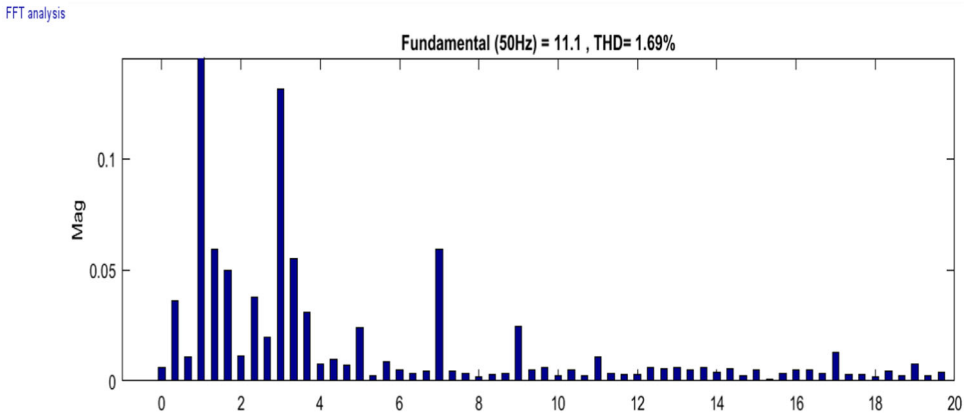


Fig. 8 Supply current and harmonic distortion without PV interconnection



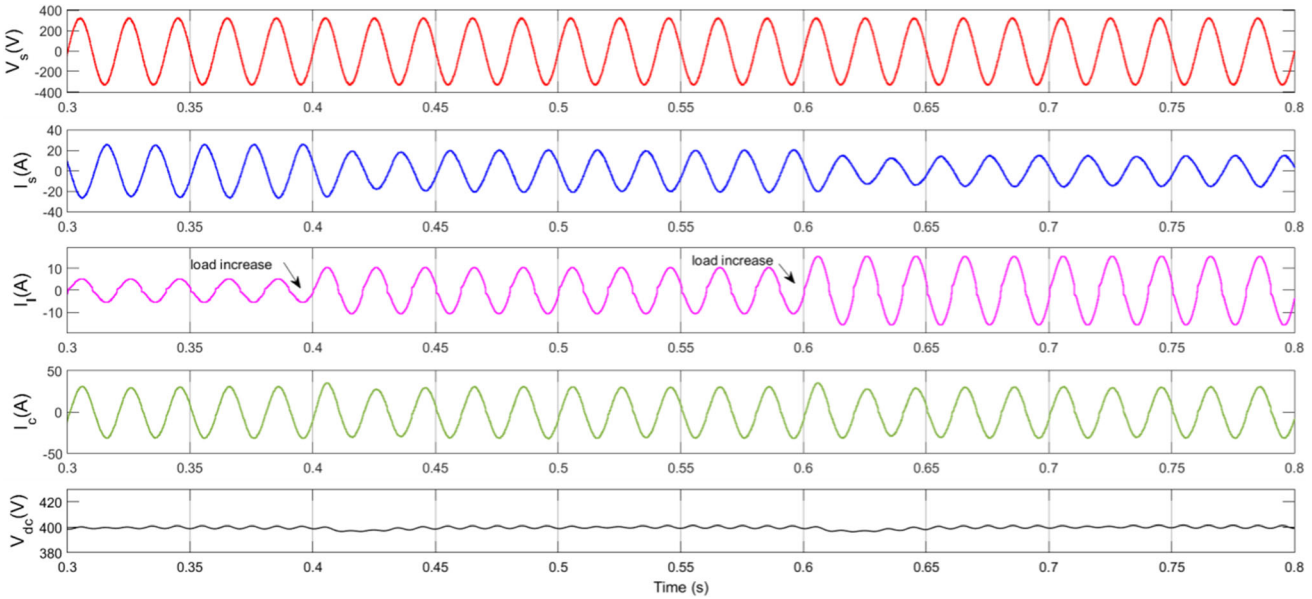


Fig. 9 System performance with PV integration

The result shows that the extracted weight is free from ripples using SOGI and A-LSTM technique. The SRF technique needs additional filters for achieving shunt compensation. Further, a fair comparison of performance features of A-LSTM with SRF and SOGI techniques for shunt compensation is discussed in Table 1. The comparison is made on the basis of convergence, dynamic response, THD of supply current, etc. Features such as fast convergence of weights within 2 cycles with no steady state error and ripples and low THD in supply current are clearly indicated with the developed A-LSTM technique. The proposed technique can be effectively employed for active filtering and improving the PQ of the power distribution system.

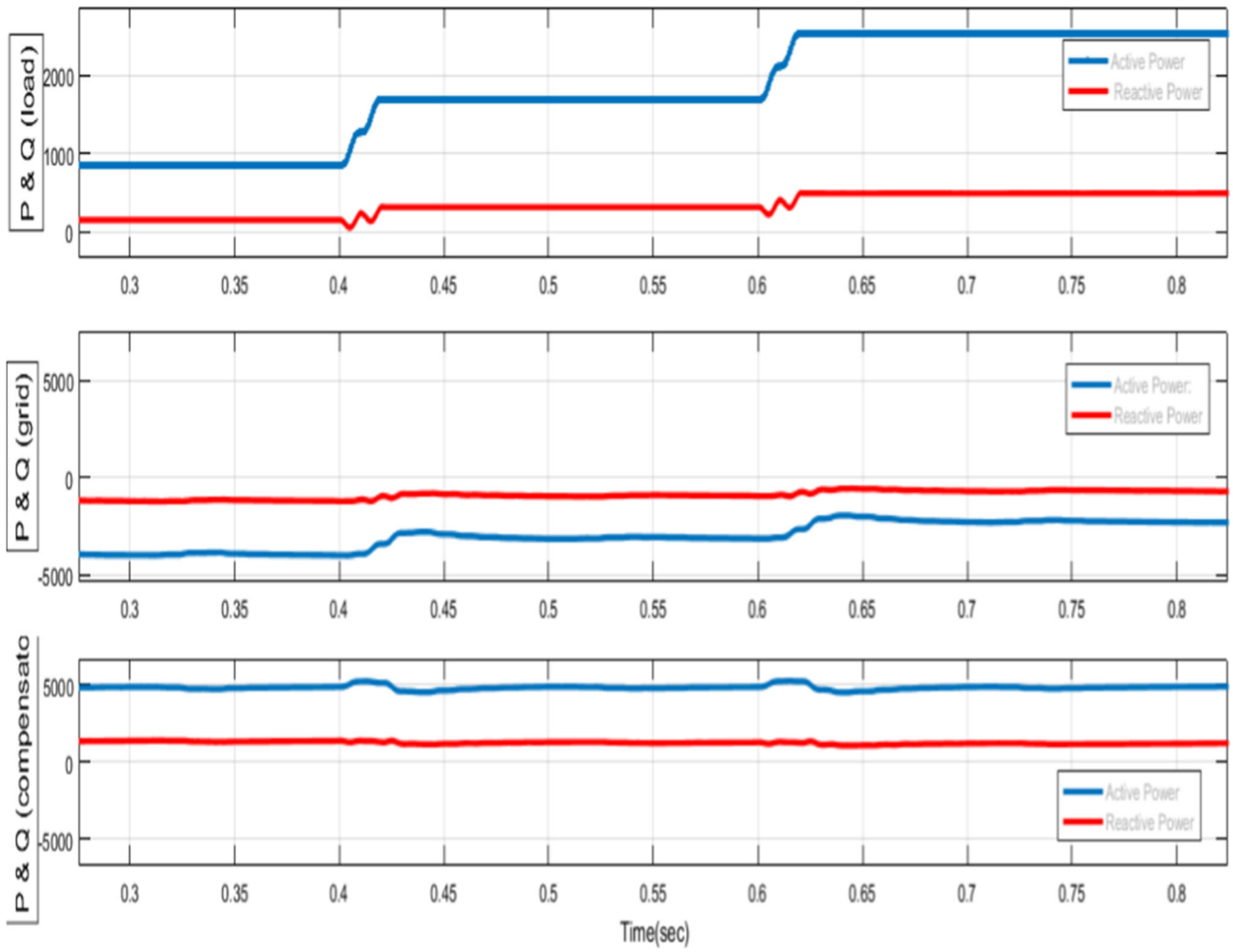


Fig. 10 Real and reactive power flows with PV integration

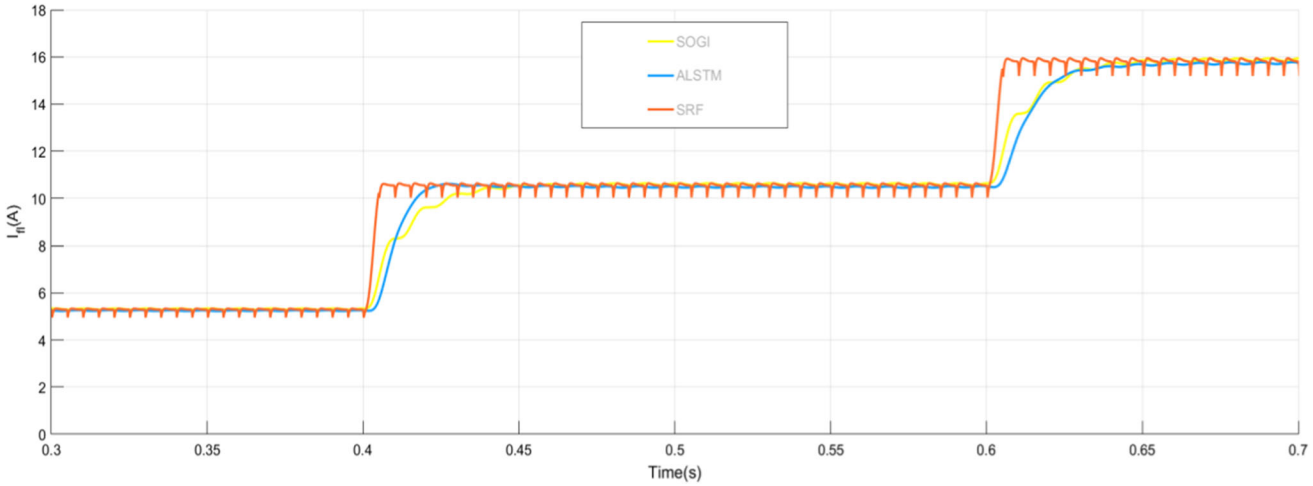


Fig. 11 Weight extraction with A-LSTM, SRF and SOGI techniques

Table 1 Comparison of techniques

S. No	Operations	A-LSTM technique	SRF technique	SOGI technique
1	Control type	Adaptive	Non adaptive	Non adaptive
2	Error in steady state	Low	High	Low
3	Response during dynamics	Fast	Fast but ripples	Moderate
4	PLL requirement	No	Yes	No
5	Additional filtering	No	Yes	No
6	Weight convergence Steady state	Within 2 cycles	in 1–1.5 cycles but ripples	Within 2–3 cycles
7	Harmonic compensation (in i_{sa})	1.69% (low THD)	4.67% (high THD)	3.1% (low THD)
8	Computational burden	Moderate	High	Low

Conclusions

This paper proposes a single layer A-LSTM which is employed for power quality improvement in a single-phase grid connected system. The developed A-LSTM network is trained online to update the weights of the network. The A-LSTM has a simplified structure and has been designed to effectively track the component corresponding to 50 Hz of the load current. Its dynamic response is fast and convergence is achieved in $\sim$ 2 cycles. The tracking performance has been observed to be highly dependent on the learning rate of the LMS technique. Hence, an optimized learning rate has been selected. The trained A-LSTM has been used to achieve shunt compensation in system without/with PV source. In both the cases, the designed controller is found to work satisfactory. Its comparison with conventional controllers highlights several advantages in terms of fast response and convergence along with low harmonic distortion in supply current.

Appendix

AC mains: 220 V, 1-phase, 50 Hz. R_s = 0.05 Ω, L_s = 1 mH. V_{dc} = 400 V. L_f = 3.1 mH. Nonlinear diode rectifier with R_{Ll} = 15 Ω, L_{L1} = 100 mH, 5 kW PV, V_{oc} = 37.5 V, I_{sc} = 8.7 A, L_b = 5 mH, C_b = 500 μF.

Acknowledgements The author is thankful to the Department of Science and Technology, Govt. of India.

Funding No funding was received for the completion of this study.

Declarations

Conflict of interest The author declares there is no conflict of interest.

References

1. B. Singh, A. Chandra, K. Al-Haddad, *Power Quality: Problems and Mitigation Techniques* (Wiley, Chichester, 2015)
2. B. Singh, J. Solanki, An implementation of an adaptive control algorithm for a three-phase shunt active filter. IEEE Trans. Industr. Electron. **56**, 2811–2820 (2009)
3. M.I.M. Montero, E.R. Cadaval, F.B. Gonzalez, Comparison of control strategies for shunt active power filters in three-phase four-wire systems. IEEE Trans. Power Electron. **22**, 229–236 (2007)
4. G. Pathak, D. Mohanty, S.K. Dwivedi, Implementation of MVF-based control technique for 3-Φ distribution static compensator. J. Inst. Eng. India Ser. B **100**, 589–597 (2019)
5. O. Kukrer, H. Komurcugil, R. Guzman, L.G. de Vicuna, A new control strategy for three-phase shunt active power filters based on FIR prediction. IEEE Trans. Industr. Electron. **68**, 7702–7713 (2021)
6. V. Puranik, S.R. Arya, SOGI–FLL based adaptive filter for DSTATCOM under variable supply frequency. J. Inst. Eng. India Ser. B **98**, 423–431 (2017)
7. V.N. Jayasankar, U. Vinatha, Backstepping controller with dual self-tuning filter for single-phase shunt active power filters under distorted grid voltage condition. IEEE Trans. Ind. Appl. **56**, 7176–7184 (2020)
8. S. Mishra, C.N. Bhende, Bacterial foraging technique-based optimized active power filter for load compensation. IEEE Trans. Power Deliv **22**, 457–465 (2007)
9. A. Bhattacharya, C. Chakraborty, A shunt active power filter with enhanced performance using ANN-based predictive and adaptive controllers. IEEE Trans. Industr. Electron. **58**, 421–428 (2011)
10. J. Fei, H. Wang, Y. Fang, Novel neural network fractional-order sliding-mode control with application to active power filter. IEEE Trans. Syst. Man Cybern. Syst. 2021, 1–11
11. S.R. Arya, B. Singh, Neural network based conductance estimation control algorithm for shunt compensation. IEEE Trans. Industr. Inf. **10**, 569–577 (2014)
12. P. Chittora, A. Singh, M. Singh, Chebyshev functional expansion based artificial neural network controller for shunt compensation. IEEE Trans. Industr. Inf. **14**, 3792–3800 (2018)
13. J. Fei, H. Wang, Experimental investigation of recurrent neural network fractional-order sliding mode control of active power filter. IEEE Trans. Circuits Syst. II Express Briefs **67**, 2522–2526 (2020)
14. J. Fei, Y. Chen, Dynamic terminal sliding-mode control for single-phase active power filter using new feedback recurrent neural network. IEEE Trans. Power Electron. **35**, 9904–9922 (2020)

15. F.A. Gers, J. Schmidhuber, F. Cummins, Learning to forget: continual prediction with LSTM. Neural Comput. **12**, 2451–2471 (2000)
16. G. Dudek, P. Pełka, S. Smyl, A hybrid residual dilated LSTM and exponential smoothing model for midterm electric load forecasting. IEEE Trans. Neural Netw. Learn. Syst. 2021, 1–13
17. C.H. Liu, J.-C. Gu, M.-T. Yang, A simplified LSTM neural networks for one day-ahead solar power forecasting. IEEE Access **9**, 17174–17195 (2021)

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

Mathematical Biosciences and Engineering

MBE, 18(6): 8727–8757.
DOI: 10.3934/mbe.2021431
Received: 05 July 2021
Accepted: 09 September 2021
Published: 13 October 2021

http://www.aimspress.com/journal/MBE

Research article

Applied convolutional neural network framework for tagging healthcare systems in crowd protest environment

Gaurav Tripathi[1], Kuldeep Singh[2] and Dinesh Kumar Vishwakarma[3,*]

[1] Department of ECE, Delhi Technological University, Delhi 110042, India
[2] Department of Electronics & Communication Engineering, MNIT, Jaipur
[3] Department of IT, Delhi Technological University, Delhi 110042, India

* **Correspondence:** Email: gauravtripathy@gmail.com.

Abstract: Healthcare systems constitute a significant portion of smart cities infrastructure. The aim of smart healthcare is two folds. The internal healthcare system has a sole focus on monitoring vital parameters of patients. The external systems provide proactive health care measures by the surveillance mechanism. This system utilizes the surveillance mechanism giving impetus to healthcare tagging requirements on the general public. The work exclusively deals with the mass gatherings and crowded places scenarios. Crowd gatherings and public places management is a vital challenge in any smart city environment. Protests and dissent are commonly observed crowd behavior. This behavior has the inherent capacity to transform into violent behavior. The paper explores a novel and deep learning-based method to provide an Internet of Things (IoT) environment-based decision support system for tagging healthcare systems for the people who are injured in crowd protests and violence. The proposed system is intelligent enough to classify protests into normal, medium and severe protest categories. The level of the protests is directly tagged to the nearest healthcare systems and generates the need for specialist healthcare professionals. The proposed system is an optimized solution for the people who are either participating in protests or stranded in such a protest environment. The proposed solution allows complete tagging of specialist healthcare professionals for all types of emergency response in specialized crowd gatherings. Experimental results are encouraging and have shown the proposed system has a fairly promising accuracy of more than eight one percent in classifying protest attributes and more than ninety percent accuracy for differentiating protests and violent actions. The numerical results are motivating enough for and it can be extended beyond proof of the concept into real time external surveillance and healthcare tagging.

Keywords: healthcare; internet of things; deep learning; convolutional neural network; protest;

violence detection

1. Introduction

Internet of things (IoT) is an interconnection of networks where multiple devices are connected through Internet [1]. The devices can range from camera, computers, sensors etc through which information is ingested and processed in the ubiquitous IoT framework. The information can be accessed anywhere, anytime. Wireless Sensor Networks (WSN) [2] propels IoT system framework which has revolutionized the manner of data transmission. Within a WSN the job of the sensors is to sense the respective environment and send it to the appropriate devices such as local host computer or any central cloud server. Modern IoT systems have various smart systems such as smart dustbin [3], smart irrigation system [4], smart healthcare system [5], and smart traffic control [6] to name a few. The focus of this proposed research is on the smart healthcare systems. The focus in the smart healthcare system is usually on the reactive state of healthcare systems. This reactive state deals after an event have happened and the patient has been admitted in the hospital where his vital life signs are monitored, stored and analysed. Smart healthcare is one of the serious challenges that needs to be addressed especially for elderly patients. Smart wearable devices for elderly patients [7] can be included in intelligent framework for smart cities. The devices record the sensor data and transmit it to a secured cloud-based platform. Smart healthcare has multiple attributes such as remote patient monitoring, wearable devices based smart computing and preparing response for emergency services.

Intelligent transport systems in smart cities works assists emergency situations [6].Intelligent transport improves the chances of saving lives in case of accidents, natural disasters and man-made situations such as violent protests. Smart cities must have provisions for dealing emergency situations**.** Proactive emergency response preparation is the key to face crowd protests and violence**.** Smart healthcare systems tag heterogeneous fields and different departments to ensure safety of patients, especially elderly patient as well as health care institutions. Smart heterogeneous applications like smart homes [8] have proven that the convergence of AI and IoT technologies in the best optimized ways during video streaming through wireless micro medical devices (WMMDs) in smart healthcare homes. Constant research in improving surveillance framework for managing crowd protests must be developed. This framework will assist in the deduction of severity of protests and can alert the healthcare systems for the availability of healthcare professionals and specialist.

1.1. Psychology of protest

The psychology of protest has been studied and the researchers have approached the idea of protests and demonstrations using social psychological approach. Van Stekelenburg et al. [9] presented motives for protests and demonstration by humans. The paper presented five unique domains which accelerates the initiation and processing of crowd protest behavior. These five things are as follows:

a) Grievances: Public anger about situations drives the initiations of protests. Joining of any popular political personality can drive agitations and protests [10].

b) Efficacy: It is an individual's person belief that can change their present conditions, policies through protest and demonstrations.

c) Identity: Identification with group, belief or any organization can give the impetus to join the protests.

d) Emotions: Anger is the main emotion that gives the energy to humans to go for protests.

e) Social Embeddedness: Sharing grievances of society as a whole with other fellow members can initiate the drive for protests.

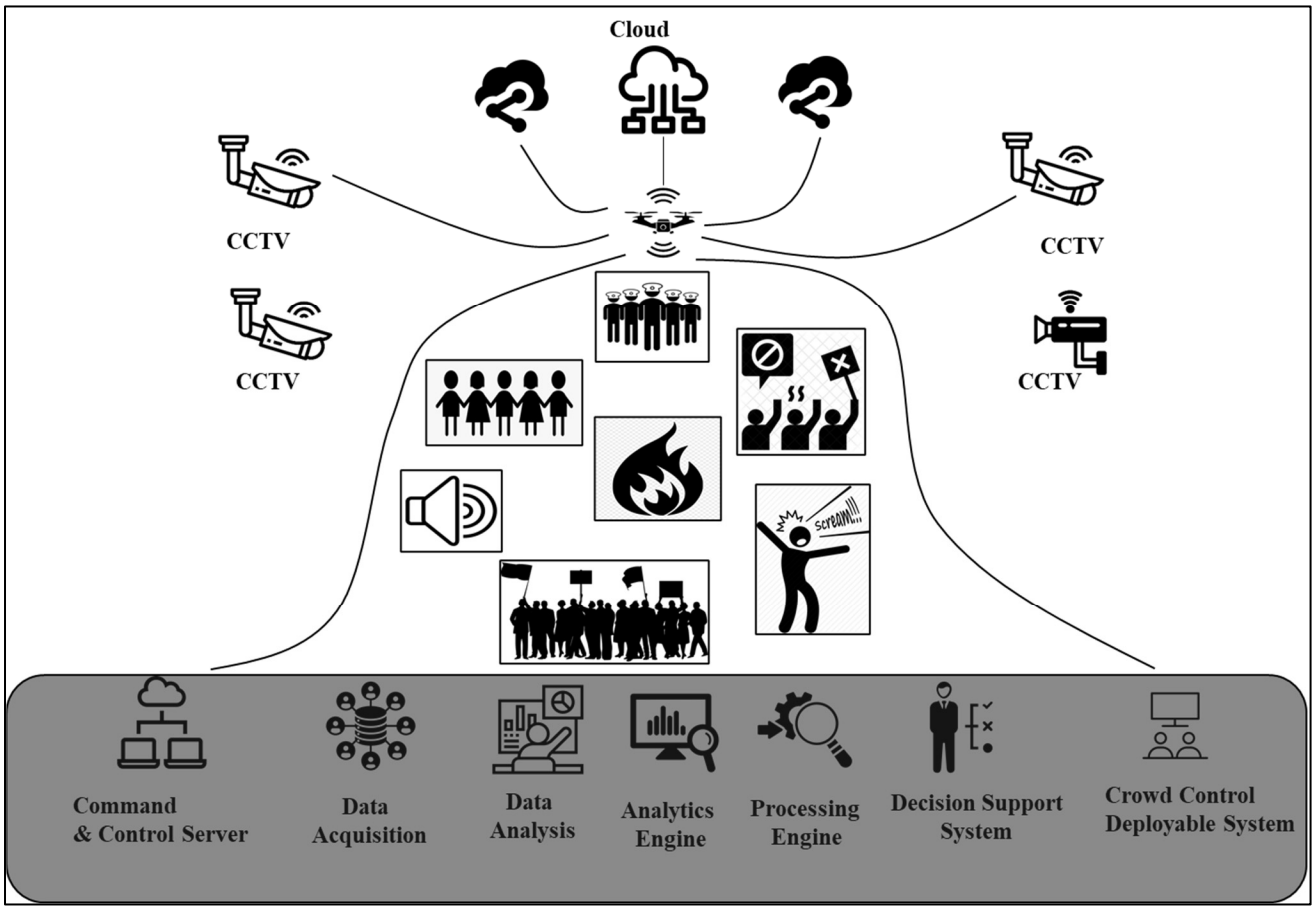


Figure 1. Overview for management of crowd protests using AI and IoT.

As per Figure 1, an overview of the envisioned IoT environment-based framework is presented. The framework is designed for a crowd protest environment. It can be clearly observed that crowd protests activities must be monitored through CCTV and drone cameras in a more sophisticated environment. This is a centrally distributed approach. The feeds are transferred to the central cloud device where the crowd protests feeds are deciphered and classified into multiple stages based on the severity of the protests. Deep learning is used for training the crowd protests models and then further classification is done on the live feeds of the protesting crowd. Based on the severity of crowd protests appropriate tagging is initiated for available healthcare systems and professionals. IoT devices such as CCTV sensors and cloud computing servers would be needed to integrate the seamless transmission of information across the city. The aim of the IoT service is to make sure that primary healthcare infrastructure is secured and upgraded and people dissenting, protesting and indulging in violent activities are attended to in any emergency situation. Thus, an effective response is computed and the specialist healthcare infrastructures and professionals are tagged in such unfortunate event.

1.2. Psychology of violence

The psychology of violence believes that there is a difference that exists in the domain of biology, psychology, sociology, culture, and mass communication which leads to the idea of violence. A general theory breaking down violence is by describing their respective attributes. External as well as internal environment of individual is responsible for motivation to carry out violence. DeKeseredy et al. [11] came up with strong theory on psychology of violence in twelve partitions which include exchange

theory, subcultural theory, resource theory, patriarchal theory, ecological theory, social learning theory, evolutionary theory, sociobiological theory, pathological conflict theory, psychopathological theory, general systems theory, and inequality theory. The same book cites reasons such as feelings of alienation, shame, humiliation by individual or group of people, mortification, rejection by individual or society as a whole, abandonment, denial, mental health ailments leading to depression, anger due to societal injustice, hostility within communities, projection, and displacement of individual or a group of people can cause extreme protests and violence.

1.3. Subsystems of healthcare

Healthcare systems consists of internal as well as external healthcare subsystems. Internal framework systems focus on the individual patient level interactions admitted to the hospital. External subsystem analyzes the external environment to decipher the need of healthcare systems, professionals, specialist doctors, emergency services in a real time situation. These subsystems facilitate the tagging of healthcare systems to the needy people. Figure 2 highlights the subsystems of healthcare framework.

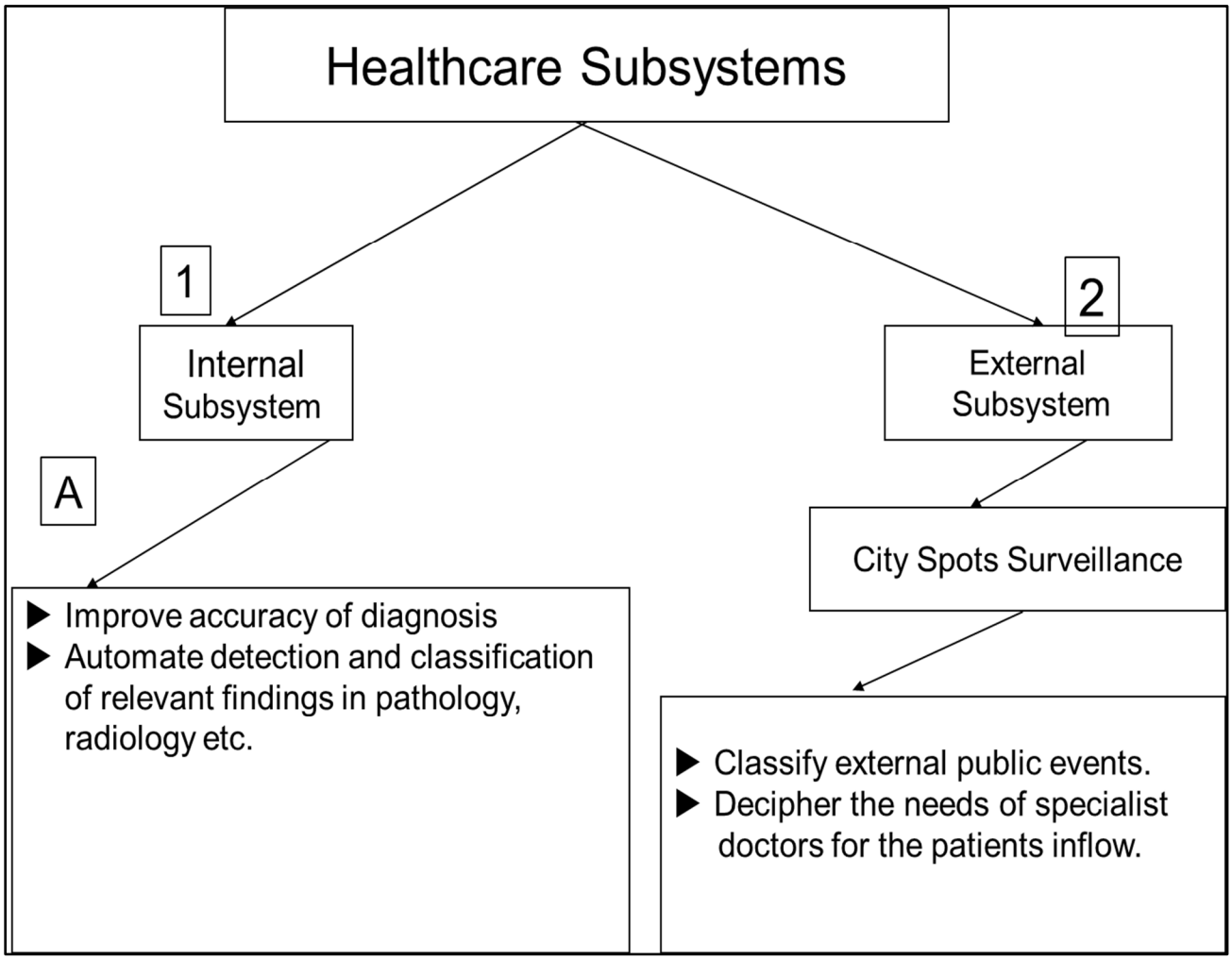


Figure 2. Healthcare system: internal and external subsystems.

The focus of this research is to study external crowd subsystems where crowded places and their

events are observed in a camera based IoT environment. The present research focuses on external subsystem where concentration is on protests sites and violent crowds' sites. Scenarios concerning these two events are included in the current study for tagging healthcare systems.



Figure 3. Protestors sample images.

Crowd protests consists of demonstration, remonstration or remonstrance and it is an expression of disagreement or objection towards an idea or action, typically a political one [12,13]. Protests have been related to cause stir in the mentality of the nation [14,15]. Social movements [16] can trigger protests in countries [17]. Protests can spiral in to violence if the demands of protesters are not met. Aggravated people can resort to violent methods to show their disapproval. Violence classification assumes prime significance in securing human life and public places and properties [18]. Figure 3 gives us the fair idea about how protests is organized using group of people, banners, children, banners,

shouting and so on. Protesting crowds have high energy [19] which can translate itself in to violence laced situation. Interesting cues regarding images and sounds have been proposed in last decade with aim to accelerate the research [20,21]. This research paved the path for more research in computer vision domain.

Convolutional neural networks (CNNs) are a new framework for deep learning solutions. The paper shall refer Convolutional neural networks (CNNs) as ConvNets. ConvNets have been applied for different computer vision applications [22–25]. A sample ConvNets is shown in the figure below.

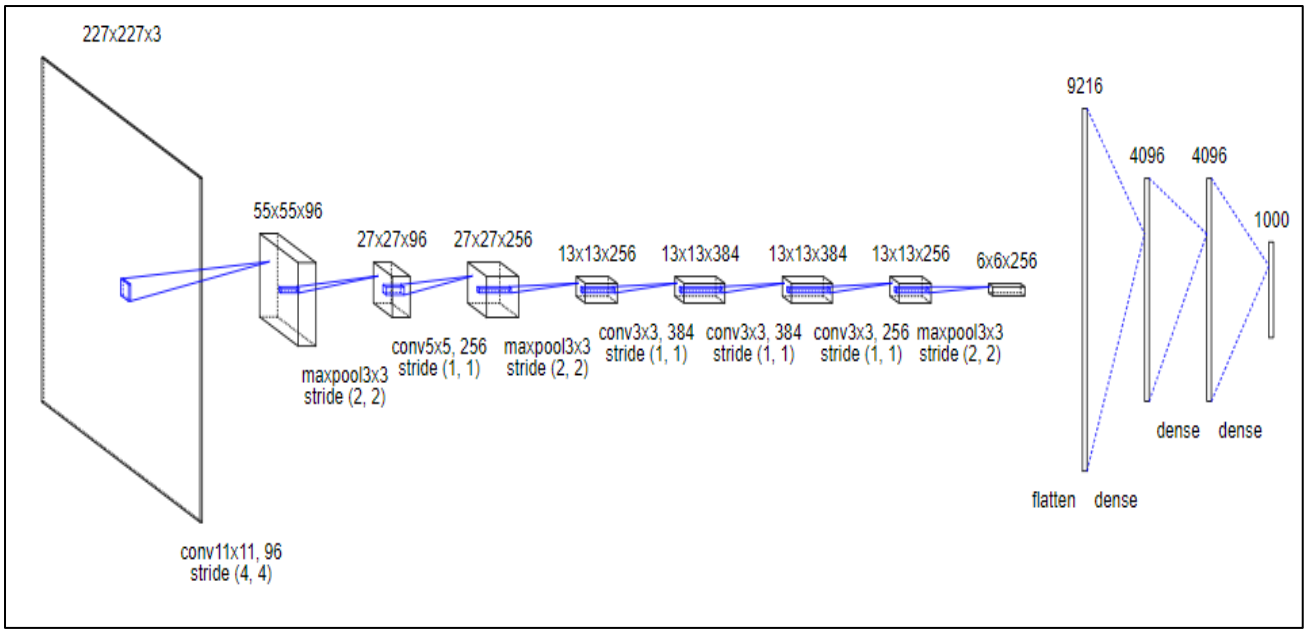


Figure 4. A Standard ConvNet.

The ConvNets have the ability to retain significant image features which it uses to learn the respective classification. The paper presents a novel application of using pre-trained ConvNets to classify stages protests and violence. Popular pretrained models are VGG16 [26], Inception ResNet V2 [27]. Each of these architectures have performed exceptionally well in ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [28]. In the event of moderate protests to extreme protests, it is the responsibility of police and law enforcement agencies to control the protests. The response from the police team for countering protests and violence is use of tear gas, pepper spray, batons, shields, and rubber bullets. The consequences of using these measures presents grave dangers to the protestors and rioters. In the midst of the protests and counter protest processes injuries are observed to the crowds. Most common injuries in case of moderate to extreme protests are ocular trauma and subsequent vision loss [29]. Thus, the paper introduces the framework of crowd behavior situations concerning protests and violence, use of deep learning to classify the situations and popular injuries that happens to the participating crowd in such situation. The contribution of the paper is as follows:

- Utilization of an existing protests datasets [30].
- Curation and utilization of violence dataset from YouTube.
- Explain the rule base to categorize the protests.
- Fuzzification of rules for degree of protests seen in the images/videos.
- Based on the degree of protests tagging of hospitals and healthcare professionals is generated.
- The idea can be spread to various other crowd domains where individual or mass of people gets injured. The whole scenario is envisaged for deployment in an IoT based smart city framework.

Our paper follows the following structure of the paper. Section 2 discusses the brief related work

of the latest protest and violence detection methods. Section 3 discusses the proposed methodology describing the pre-processing philosophy of the input and current proposed ConvNets frameworks. Section 4 describes the experimental setups along with the datasets, Section 5 describes the protests detection performance, Section 6 comes up with the discussion of the paper and Section 7 deals with the conclusion of the paper.

2. Related work

A novel multi-task convolutional neural network was developed for recognizing protesters. This model deciphered the protestors activities using their visual attributes and estimate the perceived violence in the image [30]. Our literature survey specifically focusses on deep learning-based methods for classifying protests or protests like events. Computer vision can be used for deciphering political events related questions using large-scale visual content. A study has been performed for classification of facial features for deciphering their biases towards political movements and leaders using twitter users' photographs [31]. Another study has been performed on the images of political leaders that has been shared on social and multimedia platforms [32] for their perceived emotions and personality traits. Image classification domain depends upon visual features in the input media content. These features can help classify the content in to creative content, interesting content and adult provocative themes [33]. Social event detection has been a constant source of research using computer vision and a deep learning approach has been applied using Graph Convolutional Network [34]. Computer vision has also been used for detecting extreme stages of protests such as arson and stone pelting [35]. A ConvNets and recurrent neural network approach known as collective action from social media (CASM) image data and recurrent neural networks with long short-term memory on text data was used in a two-stage classifier to identify offline collective action from social media data [36]. A novel research study was dedicated for extracting protests news by using Long Short-Term Memory and combined with a deep learning model developed on the inter country datasets [37].Our study also includes related work on violence detection methods which can be a direct result from the protest event. Novel method of Bidirectional Convolutional LSTM has been given by applied by Hanson et.al [38]. A multistream ConvNets architecture using 3D CNN and ConvNets was proposed in [23] and conducted on the CCTV-Fights dataset. A flow mechanism was deployed to detect violent event in crowds using deep learning [39] and is entirely based on the local level conditions of Bangladesh. Song et al. [40] used application of 3D CNN for violence detection approach. 3D CNN was especially used for convolutional operations on extracting the spatio-temporal features of the frames. Xu et al. proposed P3D-LSTM violent video classification [41]. Li et al. [42] synthesized a novel method by using 3D CNN utilizing multiplayer violence detection procedure. Another excellent method using 3D CNN method was proposed by Tran et al. using 3D CNN. The 3D ConvNets were designed to learn the procedure of learning the spatiotemporal features of the input videos and then classifying the videos using machine learning procedures [43]. Zhou et al. applies 3D-CNN for getting deep image features for violence classification [44]. Mandal et al. experimented with discriminative features using fine-tuned deep convolutional neural residual network for classifying crowd behavior [45]. Ammar et al. worked extracted video frames and applied CNN and LSTM layers for classification [46]. Meng et al. experimented with trajectory-based motion features and deep CNN for violence detection [47].

Table 1. Deep learning techniques for Protest and violence detection.

Authors	Methodologies	References
Perez et al.	3D ConvNets	[23]
Won et al.	Visual features for perceived violence.	[30]
Wang et al.	Facial features classification	[31]
Chen et al.	Political leaders' images linkage to protest sentiments.	[32]
Ganguly et al.	Provocative theme features extraction	[33]
Peng et al.	Graph Convolutional Network	[34]
Tripathi et al.	ConvNets used for Arson and stone pelting	[35]
Zhang et al.	ConvNets and recurrent neural network	[36]
Thenmozhi et al.	LSTM based feature extraction	[37]
Hanson et al.	Bidirectional Convolutional LSTM	[38]
Sumon et al.	Flow mechanism for violence detection	[39]
Song et al.	3D ConvNets	[40]
Xu et al.	P3D-LSTM	[41]
Li et al.	Multiplayer violence detection using 3D ConvNets	[42]
Tran et al.	3D ConvNets	[43]
Zhou et al.	3D- ConvNets	[44]
Mandal et al.	Fine-tuned ConvNets for crowd behavior	[45]
Ammar et al.	LSTM based ConvNets	[46]
Meng et al.	Normal ConvNets	[47]

Solving protest environment using ConvNets is a challenging task. A simpler and yet a novel approach is to use pretrained ConvNets subjected to fine tuning which can classify the protests and can then the situation can be classified in to their respective fuzzy categories. The authors believe that the crowd formation itself presents an opportunity to tag healthcare system although then the threat level is very less and tagging is not necessary in the normal crowd situation. The approach presented in the current paper is novel, simple and easy to tag the healthcare systems.

3. Proposed methodology

We approach the problem of deciphering two domains and the convergence of one domain in to one another. Crowd protests can transform themselves in to crowd violence and in some case crowd violence reverts back to crowd protests. As long as the protests are under control, peaceful, local law and order authorities have no problems with the demonstrators. This research article tries to decipher the various stages of convergence of protests in to violence using pretrained ConvNets. The idea of this research is to assess the protesting site for any need of the specialist healthcare workers/doctors. Escalation of protest stage to violent stage can have serious ramifications. To prove a point pretrained ConvNets are utilized in the current research. We utilize the cluster of pretrained ConvNets architecture and combined them for deducing the protests-violence classification using VGG16 [26] and Inception ResNet V2 [27]. It is a deeper network and pre-trained on more than a million images on the famous ImageNet database [48]. These ConvNets have been chosen as they have shown better performance in tackling the problem of vanishing gradient [48].

Table 2. Subtle differences in existing methodology and proposed methodology.

Authors	Methodologies	Proposed Methodology
Perez et al. [23]	3D ConvNets	3D ConvNets are known to acquire higher computational resources. Pretrained models can be easily trained and used effectively. It can also be used for mobile based devices.
Won et al. [30]	Visual features for perceived violence.	Features are extracted by pretrained model itself.
Wang et al. [31]	Facial features classification	The present approach focusses on the whole image and not just the facial features classification.
Chen et al. [32]	Political leaders' images linkage to protest sentiments.	General public images and their actions are classified for clearer protests stage deciphering.
Ganguly et al. [33]	Provocative theme features extraction	Highly dependent on the 12 features that is being classified. Has the ability to add more attributes and analyze protests more comprehensively.
Peng et al. [34]	Graph Convolutional Network	Pretrained Convnets
Tripathi et al. [35]	ConvNets used for Arson and stone pelting	Covers more protests attributes rather than just two scenarios.
Zhang et al. [36]	ConvNets and recurrent neural network	Pretrained Convnets
Henmozhi et al. [37]	LSTM based feature extraction	Pretrained Convnets
Hanson et al. [38]	Bidirectional Convolutional LSTM	Pretrained Convnets can be combined with Bidirectional Convolutional LSTM in future protests scenarios
Sumon et al. [39]	Flow mechanism for violence detection	Pretrained Convnets
Song et al. [40]	3D ConvNets	Pretrained Convnets
Xu et al. [41]	P3D-LSTM	Pretrained Convnets
Li et al. [42]	Multiplayer violence detection using 3D ConvNets	Pretrained Convnets for crowd behavior analysis.
Tran et al. [43]	3D ConvNets	Pretrained Convnets for crowd behavior analysis.
Zhou et al. [44]	3D- ConvNets	Pretrained Convnets for crowd behavior analysis.
Mandal et al [45]	Fine-tuned ConvNets for crowd behavior	Efficient pretrained models are used for protests as well as violence classification
Ammar et al. [46]	LSTM based ConvNets	Pretrained Convnets for crowd behavior analysis.
Meng et al. [47]	Normal ConvNets	Pretrained Convnets for crowd behavior analysis.

The previous section has highlighted the methods that is being applied for the protests and violence detection methods. Latest literature has few methods that directly deals with crowd protests. Also, violence detection using ConvNets are also fewer. The methods that are being touched here utilizes the pretrained Convnets. The main objective of choosing this method is to make sure that this proof of concept transforms itself in to real time program which can be used for Android and web-based devices equally. Table 2 presents a simpler difference approach between existing methodologies and proposed methodology.

Let us assume that X to be the of the protest dataset $|X|$ The error $\in$ is calculated as:

$$\in (W, X) = \frac{1}{|\mathrm{D}|}\sum_{i=1}^{|\mathrm{D}|} \mathcal{L}\left(\mathcal{F}(X_i, w), \widehat{c_i}\right) \tag{2}$$

where X_i is the i^{th} image of training dataset X, $\mathcal{F}(X_i, w)$ is the ConvNet function that predicts the class c_i of X_i given the respective weights w,and $\widehat{c_i}$ is the ground-truth class of the i^{th} image, and $\mathcal{L}(c_i, \widehat{c_i})$ is a penalty function for predicting c_i instead of $\widehat{c_i}$. We can set $\mathcal{L}$ to the logistic loss function [49]. In the proposed method we Adaptive Gradient Descent (AGD/AdaGrad) [50].The cost function to compute the optimization error is as follows:

$$\mathcal{L} = -\frac{1}{|X|}\sum_{i}^{|\mathrm{D}|} \ln\left(p(c_i|X_i)\right) \tag{3}$$

$p(c_i|X_i)$ denotes the probability by which X_i is correctly classified in the protest dataset. If W_h^t denotes the weights in h^{th} hidden convolutional layer at iteration t, and $\hat{\mathcal{L}}$ denotes the cost over a mini-batch of size M, then the updated weights in the next iteration are computed as below:

$$\eta^t = \eta^{\left\lceil \frac{tM}{|\mathrm{X}|} \right\rceil} \tag{4}$$

$$U_h^{t+1} = \varrho U_h^t - \eta^t \alpha_h \frac{\partial \hat{\mathcal{L}}}{\partial W_h} \tag{5}$$

where α_h is the learning rate for the h^{th} layer, ϱ is the momentum that signifies the contribution of the last weight update in current iteration and η is the scheduling rate that decreases α at the end of each epoch. The Adaptive Gradient Descent was computed for optimizations vide the following equations with respect to the mini batches formed similar to that used in SGD.

$$U_h^{t+1} = \varrho U_h^t - \frac{\eta}{\sqrt{K_{h,h}^t + \beta}} . \alpha_h \frac{\partial \hat{\mathcal{L}}}{\partial W_h} \tag{6}$$

$$W_h^{t+1} = W_h^t + U_h^{t+1} \tag{7}$$

where $K_{h,h}^t \in R^{d \times d}$ represents a matrix such that the diagonal elements (h, h) equals the sum of gradients.

Following the above calculation, the weights in the fine-tuning architecture were updated for optimization as indicated by Eq (7). We can clearly infer from Eqs (3)–(7), the back-propagation activity was carried out iteratively to fine tune the pretrained model. Following two pretrained ConvNets are used in deciphering the multiclass protests and then used for computing severity of the protests.

3.1. VGG16 ConvNets

VGG16 is one of the popular pretrained ConvNets and had a good classification accuracy in ImageNet large scale visual recognition challenge (ILSVRC) [28]. This is simple to use and is popular in the transfer learning domain. In the process of fine tuning VGG16, the fully connected top layers are removed and new fully-connected layers are added for our classification of tasks. The model utilizes a set of weights pre-trained on ImageNet. VGG16 uses gradient descent combined with a loss function over a sufficiently large dataset propels the ConvNets to learn the patterns in the complex visual space data that is fed to the network. VGG16 consists of 16 convolutional layers. These layers help the architecture in feature extraction. VGG16 has three main constituents: convolution, pooling, and fully connected layers. VGG16 proceeds with two convolution layers. Then a pooling layer is added. The pattern is repeated again. Then three convolutions are repeated with pooling attached. Finally, three FC (Fully Connected) layers are added.

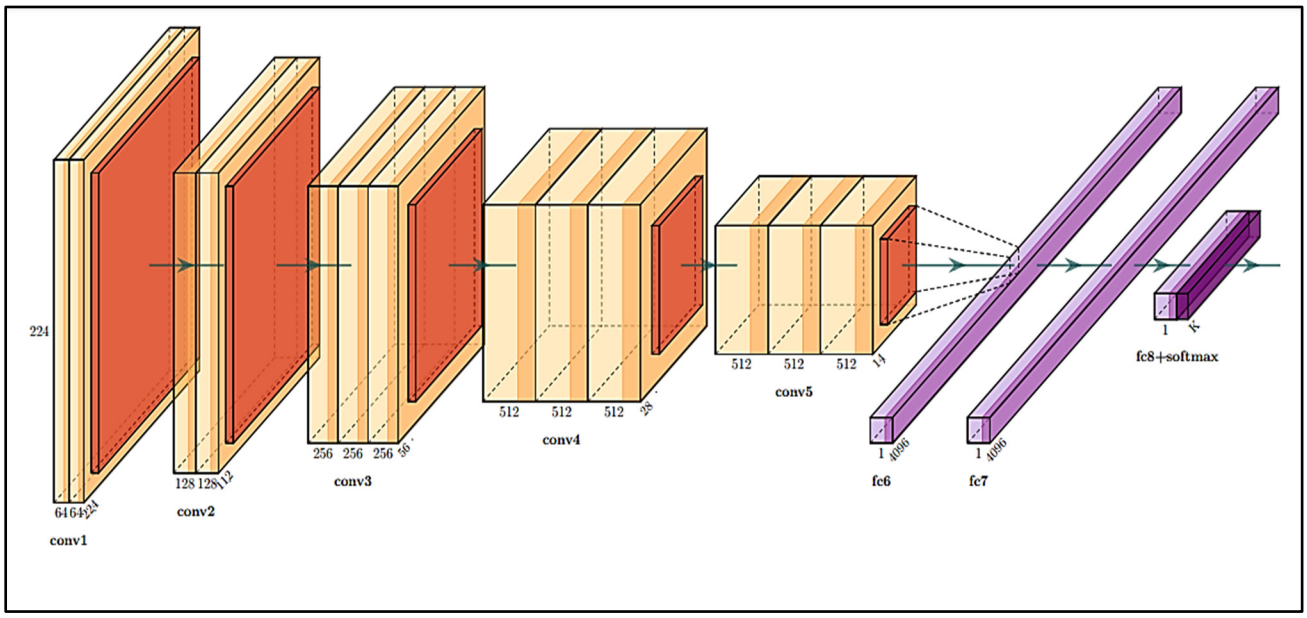


Figure 5. VGG16 Architecture.

Figure 5 presents the VGG16 architecture which is used in the current research to highlights the working of ConvNets and is generated using [51].The above architecture has single global max pooling layers, single fully connected layers with 512 units, single drop out layer and single softmax activation layers of violence enabled inputs. The layers in VGG16 are frozen and feature extraction is performed for the additional FC layer.

3.2. Inception ResNet ConvNets

With the introduction of residual connections with normal inception blocks has yielded good performance in 2015 ILSVRC challenge and the new architecture is known as Inception ResNet V2 [27]. These residual connections facilitate the shortcuts in the model to train deeper neural networks, Hence the performance has increased for inception blocks.

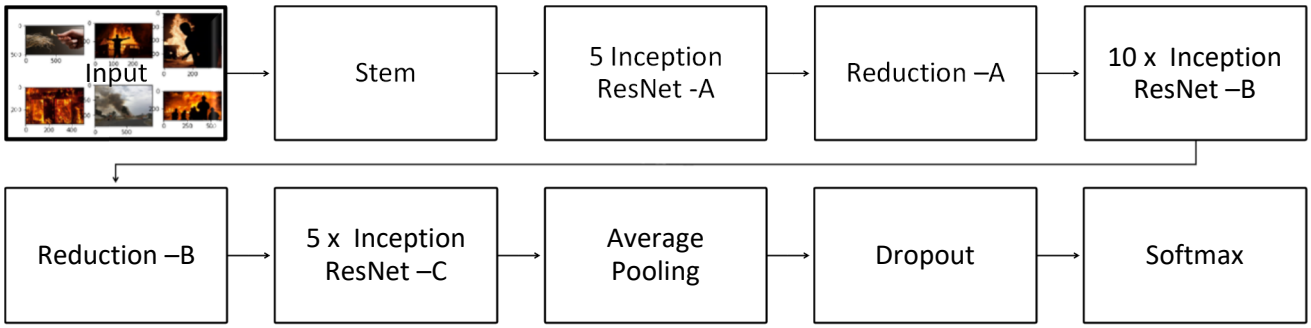


Figure 6. Inception-ResNet architecture.

Figure 6 shows the steps of the input images to the final classification of the protests. This model reduces the representational bottleneck by introducing factorization of large block in to two smaller blocks. Also, it expanded the filter banks horizontally rather than vertically. This ensured less loss of information. The above steps were incorporated to build three different types of inception modules known as Figure 6. Instead of training the model from scratch the paper relies upon the transfer learning concept and does fine tuning of the Inception-ResNet V2 on the protest dataset.

4. Experimental approach

We have evaluated our approach on standard imbalanced protests datasets [30] and have tried to classify various classes of protests dataset using transfer learning. To further prove the research analysis, we have curated a violence related dataset mostly from YouTube and other related videos. Protest dataset [30] have been used extensively for classification using three pretrained ConvNets. The number of images in this dataset is 40,000 images. We have added an additional category of loudspeakers for protest dataset and removed violence category from this dataset. Instead, we have curated our own images and added to the violence category. The approach is twofold. One is to analyse protests from the angle of Table 2 and find out the severity of the protest events. The second is to classify protests and violence. We are focussing on protests dataset and their combinations among them for the fuzzification of the severity of protest. We design a rule base for identifying the stages of protests and its severity. Table 3 presents the rule base in a tabular form:

Table 3. Rule base for severity of protests.

Situation Scenario	children	fire	flags	20 people	100 people	Night	Photo	Police	Protest	Shouting	Sign	Loud speakers	Fuzzy Protest
1.	–	√	–	–	–	√	–	√	√	–	–	–	Medium
2.	√	–	–	–	–	√	–	√	√	√	–	–	Severe
3.	–	√	–	–	√	√	–	–	–	√	–	√	Severe
4.	–	√	–	–	–	√	–	–	–	√	–	–	Severe
5.	–	√	√	–	√	–	√	–	√	√	–	√	Severe
6.	–	√	√	–	√	–	–	√	–	√	–	√	Severe
7.	–	√	–	–	–	–	–	√	–	–	–	√	Medium
8.	–	√	√	–	√	–	–	–	√	√	√	√	Severe
9.	–	√	–	–	√	–	–	–	–	√	–	√	Severe
10.	–	√	–	–	√	–	–	√	–	√	√	√	Severe
11.	–	√	–	–	√	–	–	√	–	–	√	√	Medium
12.	–	√	–	–	√	–	–	√	–	√	√	√	Severe
13.	–	–	√	–	√	–	–	–	√	–	√	√	Normal
14.	√	√	√	–	√	–	√	–	√	–	√	√	Normal
15.	√	√	√	–	–	–	–	–	–	–	–	–	Normal

The above table gives a fair idea about how to interpret the table. There are twelve attributes which decide the situation. Based on the presence of few or more attributes beyond a threshold, the protest situation is classified in to different normal, medium and extreme level protests. For example, if there is a presence of night, children, fire, shooting and police presence the protests situation is in severe category and appropriate solution is suggested. One thing we noted that combinations of fire, night, police, children, shouting and loudspeakers and group of more than 100 denoted severe protests in to combining in to violence. These combinations have inherently more chances of transformations of protests in to violence. As soon as these combinations are detected an alert notification is pushed to the all the available health care centres of the city. Based on the severity categories an alert for specialist healthcare personnel is generated. Based on the presence of these twelve attributes of any crowd protest event, healthcare systems are tagged and a ready alert is generated. A report by Amnesty international [52] discussed about the police actions on the protestors. The report discussed about the actions taken by police in tackling unarmed protestors and use of weapons to neutralize them. Serious injuries amounting to loss of lives have been reported. As stated in the literature, the standard police response to control protests is to use multitude of weapons, such as tear gas, pepper spray, rubber bullets, batons, riot shields, and TASERS [53]. Similarly violent protestors display their potential by using rocks, sticks, knives, guns, and Molotov cocktails [54]. Taser devices are used by law enforcement agencies to control violent protests and riots. These devices are known to reduce the risk of injury and death by minimizing the use of lethal force. Spectrum of injuries reported by usages of taser [55] is shown in the Table 4:

Table 4. Spectrum of injuries reported by usages of taser [55].

S.No	Type of injuries reported by usage of Tasers	Types of healthcare professional needed
1.	A basilar skull fracture, right subarachnoid haemorrhage, and left-sided epidural haemorrhage necessitating craniotomy	✓ Neurosurgeon Surgeon [56]
2.	A concussion, facial laceration, comminuted nasal fracture, and orbital floor fracture	✓ Maxillofacial Surgeon [57]
3.	Penetration of the outer table and cortex of the cranium by a Taser probe with seizure-like activity reported by the officer when the Taser was activated	✓ Plastic Surgeon [58]
4.	A forehead hematoma and laceration	

Hence locating these surgeons nearby to the protests and violence can help save many lives and prevent major injury losses. Below table is a research article that focusses on the types of injuries that is faced by the crowd protestors.

Table 5. Type of injuries possible during protests [59].

Protests Response	Ocular Injuries	Chest Injuries	Fractures	Skin Injuries	Bullet Injuries
Tear gas	√	–	–	√	–
Pepper spray	√	–	–	√	–
Rubber bullets	√	√	–	–	√
Batons	√	√	√	–	–
Riot shields	√	–	–	–	–
TASERS	√	√	√	√	√

Table 5 describes the types of serious injuries that happens in severe protests and violence. Table 6 describes the types of doctors needed for the kind of injuries reported in Table 4.

Table 6. Protest's injuries and their respective specialist.

Typical Injuries	Type of specialist needed
Ocular Injuries	Ophthalmic Surgeon [60]
Chest Injuries	Orthopaedic Surgeons [61]
Fractures	Orthopaedic Surgeons [61]
Skin Injuries	Plastic Surgeon [58]
Bullet Injuries	Orthopaedic Surgeons [61]

The above tables clearly describe the types of police response used, types of injuries inflicted and types of specialists needed in case of severe protests. Thus, a proper relationship between protests, response, injuries and specialist healthcare professionals has been established in this research. Now based on the severity of protests, general physicians as well as specialists' doctors would always be available in the hospitals during protests. For the implementation of proposed protests classification and their respective typical injuries detection in crowded scene algorithm, we used standard Keras

application with Python on Ubuntu operating system and experiments were conducted on a standard machine (Intel Core i5,8 GB RAM) equipped with NVIDIA TITAN X GP.

5. Protest detection performance

We deployed the above-defined architecture. The results are divided in to two approaches:

- The first approach uses pretrained ConvNets for multiclass classification of protests datasets. We use VGG16 and Inception Resnet V2 for the protest's classification.
- A rule based is defined in the previous section. Based on this rule base an alert is generated which can be sent to the nearest healthcare centers demanding the readiness of specialist's doctors as described in the previous sections. Sending the alerts can be automatic or through a human in the loop [62]. This research believes that the human is necessary to broadcast the protest level to all the healthcare centers in the vicinity of the protest's sites. The whole research is dedicated to the protest's behavior and its respective severity.

Total dataset is revised by adding loudspeakers to the protests datasets and violence category is separated from the original dataset. Then the utility is invoked for displaying of dataset images. The image is as follows:



Figure 7. Protest images in the actual dataset.

Then the experiment proceeded with VGG16 and Inception Resent V2 pretrained ConvNets. The experiment utilizes k-fold cross Validation technique. Cross-validation is a unique procedure used for evaluation of deep learning models on the limited data sample. As specified in the literature [63] ,the choice of k is normally taken to be 5 or 10. Arbitrary increase in the value of k cannot guarantee better

results as resampling subsets gets smaller. The experimental procedure is as follows:

1. Protest dataset is shuffled randomly.
2. Split the dataset into k = 5 groups
3. For each such group:
   - ✓ Take the group as a hold out or test data set
   - ✓ Take the remaining groups as a training data set
   - ✓ Fit a model on the training set and evaluate it on the test set
   - ✓ Retain the evaluation score and discard the model
4. Model evaluation scores and respective confusion matrix is described thereafter.

Evaluation metrics: The performance of the proposed research concept was tested and on the parameters of accuracy, specificity, sensitivity, precision, and f1-score. The parameters used in the following equations are true-positive N_{TP}, number of true negative N_{TN}, number of false positive N_{FP}, and number of false negatives N_{FN} to measure predicted classes which are based on the evaluation metric mentioned in Eqs (8–12):

$$\text{Accuracy}= \frac{N_{TP}+N_{TN}}{N_{TP}+N_{TN}+N_{FP}+N_{FN}} \tag{8}$$

$$\text{Specificity}= \frac{N_{TN}}{N_{TN}+N_{FP}} \tag{9}$$

$$\text{Sensitivity}= \frac{N_{TP}}{N_{TP}+N_{FN}} \tag{10}$$

$$\text{Precision}= \frac{N_{TP}}{N_{TP}+N_{FP}} \tag{11}$$

$$\text{F1- Score}= 2 \times \frac{\text{recall}\times precision}{precision+recall} \tag{12}$$

VGG16 confusion matrix and classification results are as follows. The respective confusion matrix is as follows:

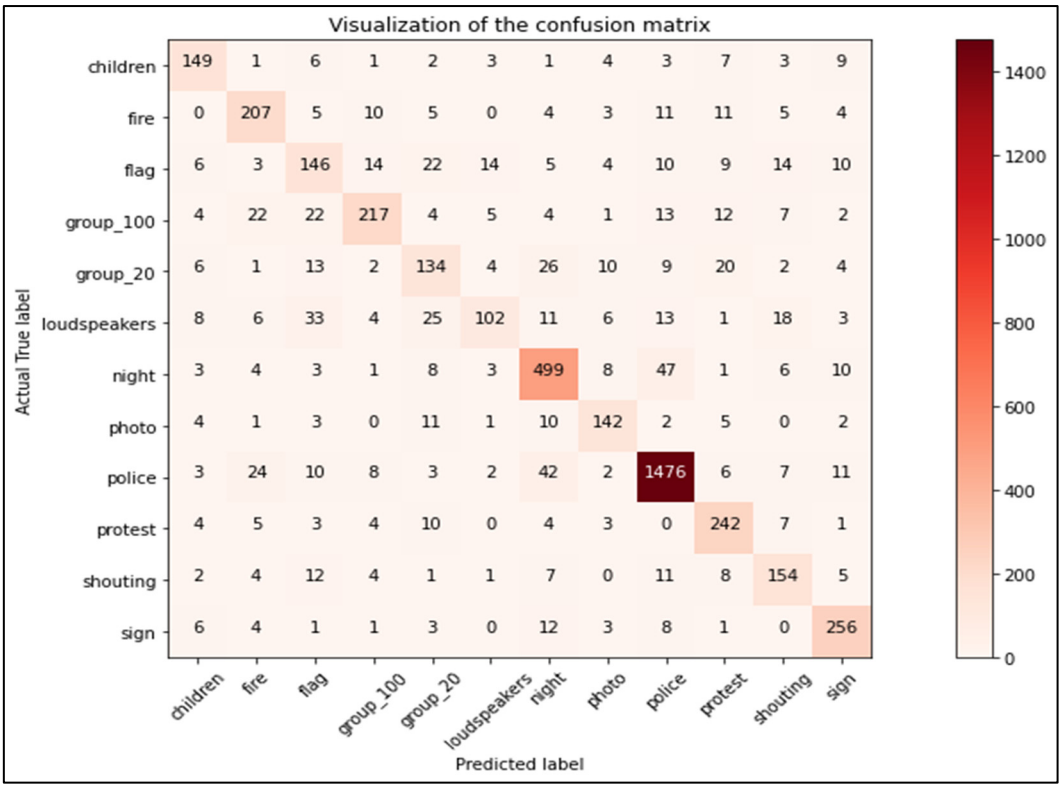


Figure 8. VGG16 confusion matrix for protest dataset.

The training accuracy is a follow:

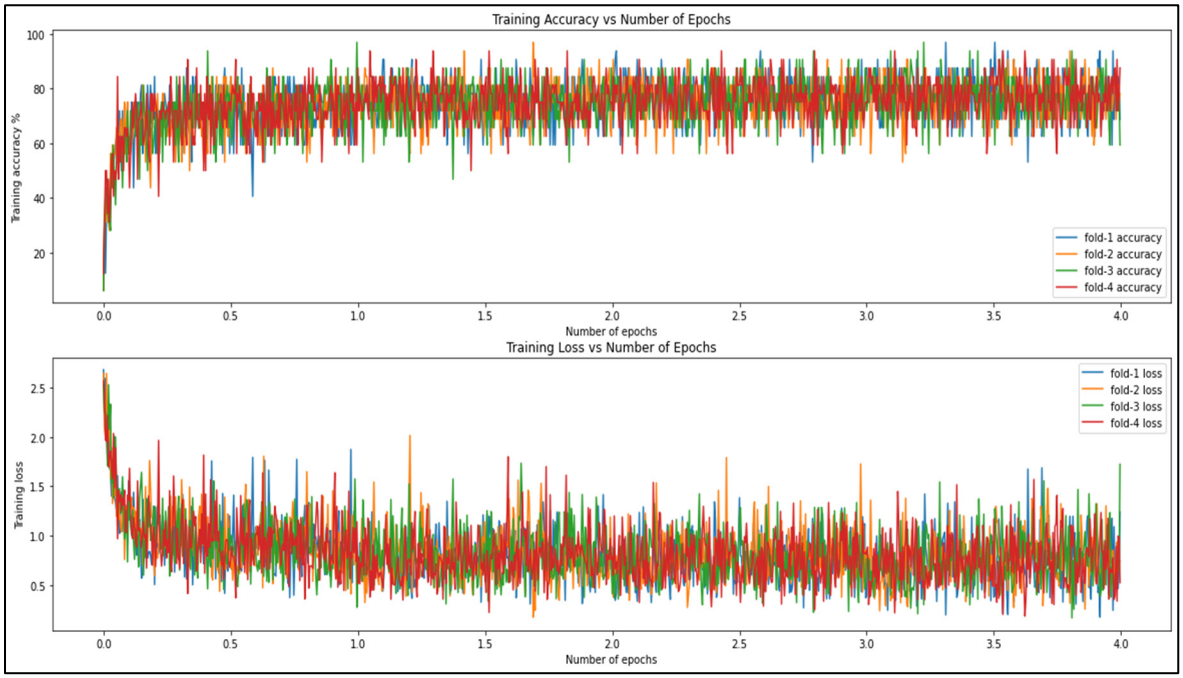


Figure 9. Fivefold accuracy training accuracy vs number of epochs for VGG16.

The results of Inception Resnet V2 in the form of confusion matrix and five-fold accuracy graphs is as follows:

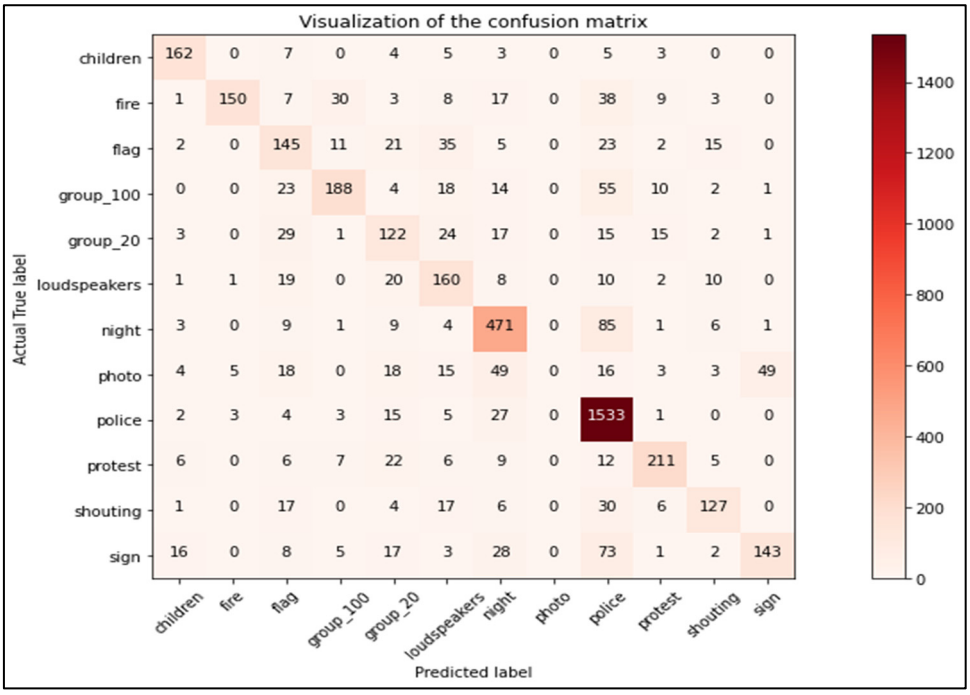


Figure 10. Confusion matrix for protests datasets using Inception Resnet V2.

The 5-fold accuracy graph for all the three approaches is as follows:

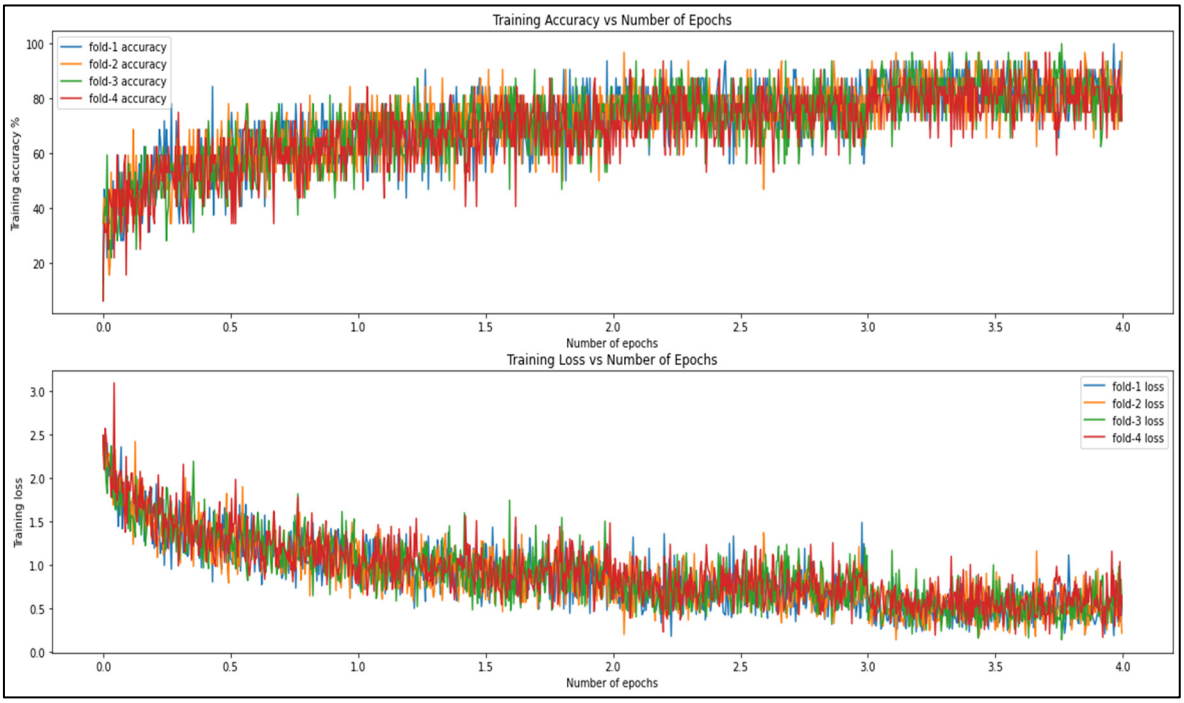


Figure 11. Five-fold training accuracy vs number of epochs graph.

Table 7. Accuracies comparison table.

	Accuracy	F-score	Recall	Precision
VGG16 [26]	80.259	0.738	0.739	0.746
Inception ResNet V2 [27]	82.259	0.758	0.789	0.776

From the above Table 7, it becomes clear that our proposed framework achieves substantial performance on the protest dataset proving that our association strategy of pretrained ConvNets for protest detection have given practical results. The results are promising and they can be further improved using more balanced dataset by improving protest dataset. The results in the multiclass classification are returned in the probability. The returned probabilities of the classes beyond a threshold are taken as the entity present in the input images. Multiclass classification helps in classification of image datasets and helps in classifying the images. Based on this multi class classification various rule set can be framed and a decision support system can be easily made for assisting humans. Further a simple binary classification for classification of protests and violence category is also complied in this research. Here a binary classification between protest and violence dataset is undertaken. The aim for such a classification is that the detection of violence in the input images/videos can make the alert system more perfect for alerting healthcare systems personnel.

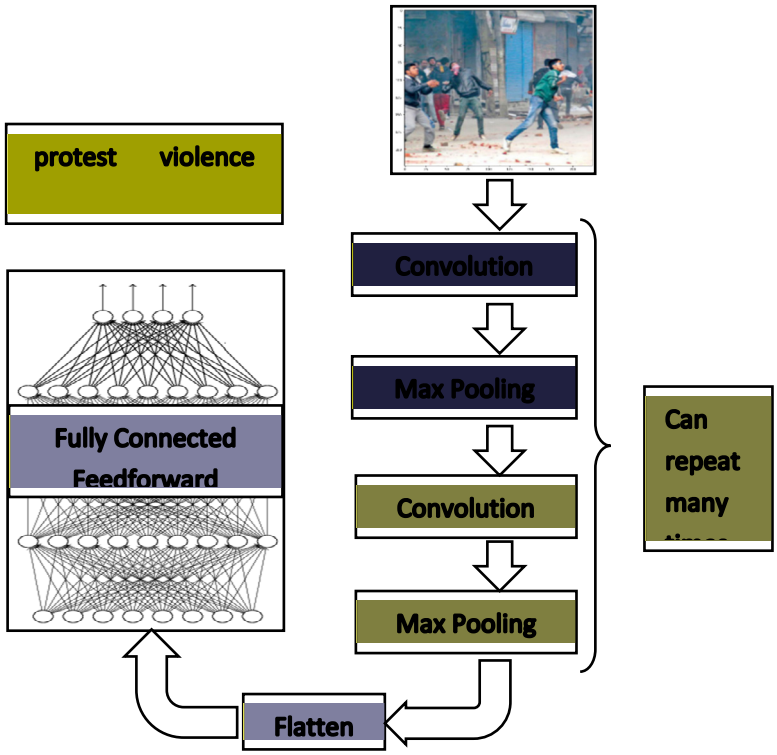


Figure 12. A simplistic diagram of ConvNets module for image classification.

The protest and violence dataset were subsampled to total of 9000 images each. A classification scheme using pretrained ConvNets was utilized. We use VGG16 for the classification purpose. Finetuning of VGG16 is undertaken. The top fully connected layers are removed newly fully connected layer is added to the existing pretrained ConvNets. This new layer consisted of global max pooling layers, one fully connected layer, one drops out layer with 0.49 rate and one softmax activation function for 2 types of categories. Feature extraction is performed for the newly added FC layer. This ensure that the weights in the system is not random and gradient would be under control. Post feature extraction the final ConvNets layer is unfreeze and fine tuning of the model starts for father iteration. The learning is rate is 0.001 and the Ada Grad optimizer is applied on the classification algorithm.

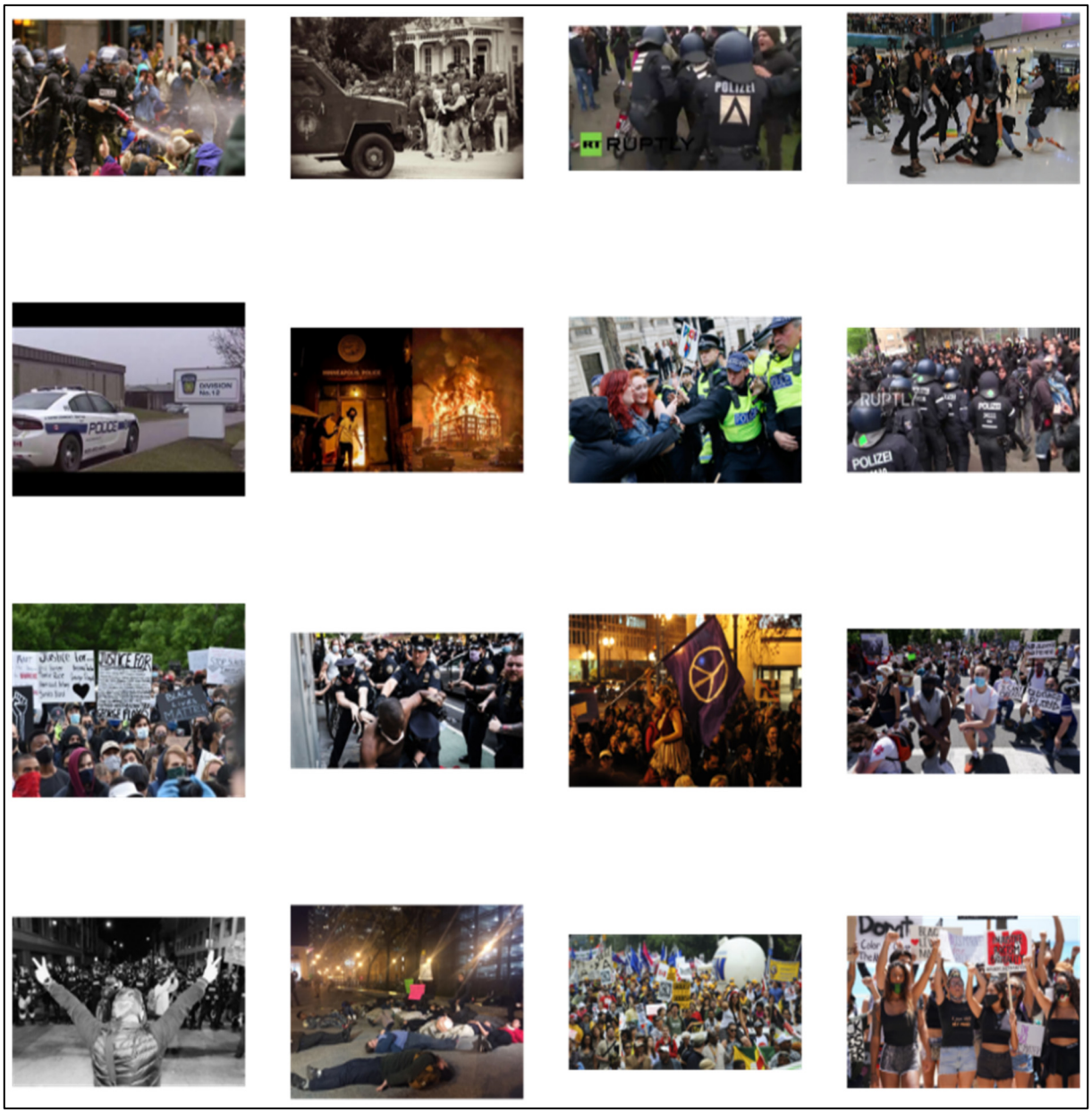


Figure 13. Sample images of protests and violence.

The results of the system are as follows:

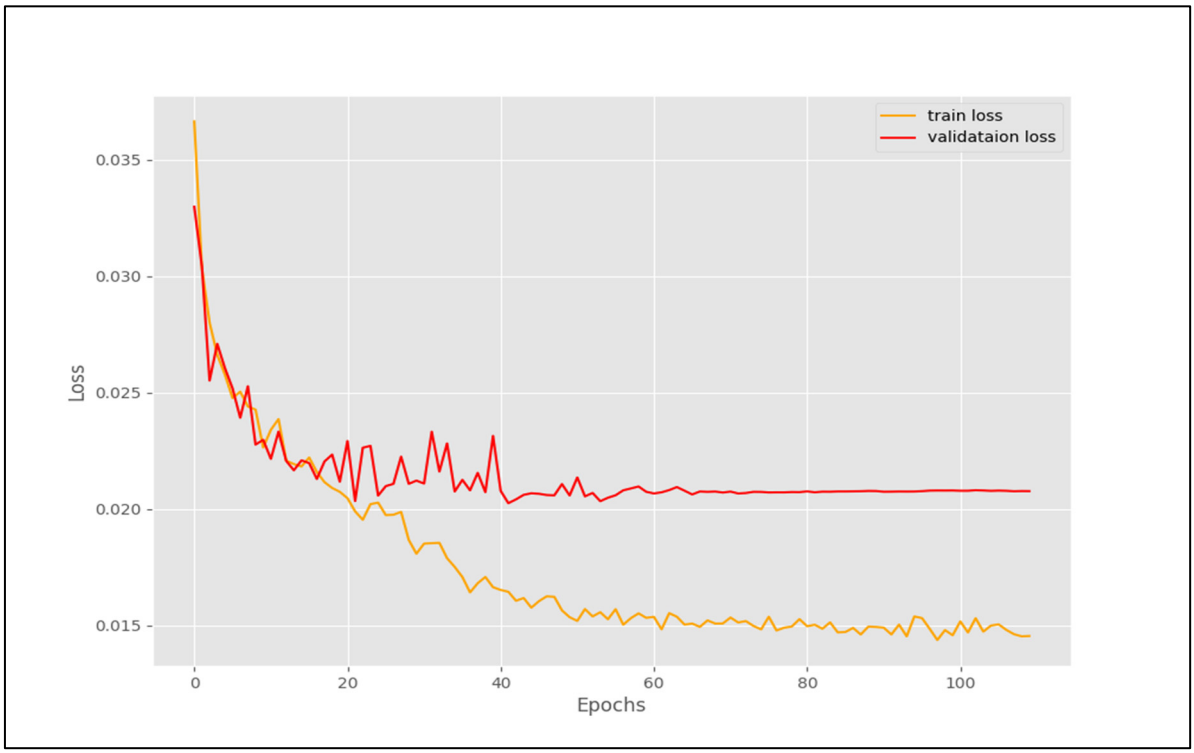


Figure 14. Training loss and validation for protest and violence classification.

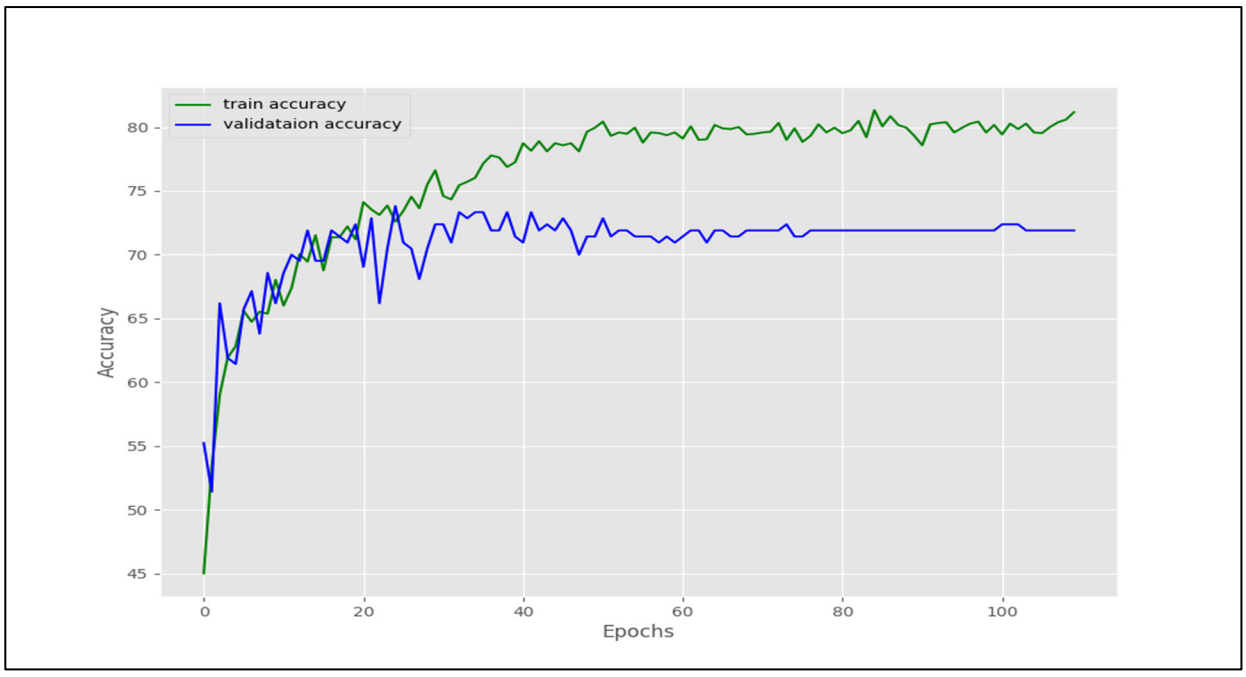


Figure 15. Training accuracy and validation accuracy for protest and violence classification.

Fine tuning the whole model using transfer learning gives better result and the model converge faster.

Table 8. Accuracy of the protests and violence dataset.

Protests and Violence dataset	Accuracy
VGG16 [26]	92.789%

Figure 15 and Table 8 shows the accuracy and loss diagrams of the VGG16 approach for classifying protest and violence. To the best of knowledge, this is a unique classification scheme of classifying image categories of protests and violence. Violent activities automatically qualify for tagging of specialist healthcare system and surgeons. Figures 14 and 15 describes the result and shows overfitting results. The sample is size is limited which may not generalize well on unseen data. As per Figure 12 a simplistic framework for classification is shown where convolution and pooling layers combinations extend in to the deep learning architecture. The approach to tackle overfitting is to optimize and reduce the deep learning capacity by removing convolutional layers as shown in Figure 12 and reducing the number of elements in the hidden layers. Application of regularization and usage of dropout layers can help in reducing the overfitting in the model performance. A recent research article in the medical field has proposed a novel deep learning model for more efficient and precise recognition of Tetralogy of Fallot (TOF) [64] has discussed a novel method of deploying stochastic pooling methods to counter overfitting models. A rank-based stochastic pooling (RSP) for breast cancer classification has been recently published [65] . A future version of this paper can be extended by combining two advanced neural networks namely Graph convolutional network (GCN) and ConvNets (CNN).

6. Discussion

Protest detection for crowd management is a challenging job. This research has presented a way to link crowd behaviour to deep learning and tagging healthcare system to the safety of people. There are various reasons which affects individuals and group of people which can trigger the cause of protests. Protest can manifest itself due to various reasons which includes anger, emotions, identity to social embeddedness. The idea percolated in this research is that protests are collective crowd behaviour. If not handled properly, protests can spiral in to violence. The paper explores post protests experiences and safety of the people who are in protesting, police who are controlling the protestors and normal crowd who are just bystanders, onlookers who are witnessing the protesting events. Thus, a lot of people gets involved, voluntarily or involuntarily. The paper explores this crowd behaviour using protest dataset. Using multiple attributes of the protest dataset this research classifies the severity of crowd protests using rule set belonging to the respective attributes of protest dataset. Converging of protests in to violence can cause stern retaliatory action from the law enforcement agencies. The significance of this research is to come out with a safety plan of people who may get injured in the protests. This research highlighted the common injuries that happens during severe protests. These injuries are sustained by the protestors as well as the police forces. Even the bystanders can get fatally injured in these protests. The highlight of the research is to make a protest storyboard. People protesting would definitely face police controllable actions. Protesting people are united either by anger or identity issues, they usually turn violent to make their point of view heard by the government system and authorities. Amidst all this chaos, if the healthcare systems, professionals, ambulances, medicine

stores etc are ready for protest and violence situation, the objective of this research is achieved. At the end of the day no human lives must be lost due to violent protests. Crowd behaviour like protests and riots is a very hard problem to solve using any images datasets. Further the current research links the type of injuries that can happen during protests and severe protests. Each such injury is mapped to a specialist health care personnel to specific surgeons. Also, it is imperative to view the tagging of healthcare systems from two point of view. The first view is dependent solely on the multiclass protest classification. Based on their respective threshold probability score, the rule sets are fired and an alert is generated for tagging healthcare systems. The second view is to generate the classification of each frame of the real time of the input video in to protests and violent categories. This research believes that if the violent state is detected then it automatically is qualified for tagging specialist healthcare systems and doctors to be at alert for saving human lives. This research can be a game changer in the IoT enabled smart cities and devices. Future versions of this can be Android complied and an app can be made functional for directly pushing notifications to the healthcare professionals. All the medicines, injections, anti-biotics, oxygen cylinders needed for such scenarios could be readied beforehand and the human lives can be saved even in the midst of sever protests and violence.

6.1. Contexts, scenarios and solution

People believe in the power of protests. The time frame of acceptance of the protesters' demands may vary. In the initial stages of protests, the event looks like a failure but the power of protests becomes visible in the long term. A system is envisaged which interconnects the subsystems of healthcare tagging. This system is connected to the videos / live stream of the protesting sites. These frames from these feeds are extracted and pumped into the trained model which classifies the protest attributes. Each combination of protest attributes forms a rule which is triggered respectively. There are usually two types of healthcare centers. Primary healthcare centers are controlled by the government and the private hospitals. All the primary health care centers are geo mapped. As soon as the protests' severity begins, their respective classification and notifications regarding the state of the protests begin to generate. These notifications are pushed through the pipeline and concerned authorities are informed about the ongoing situations. Following prototype is envisaged to prove our points of tagging healthcare amidst protests and violence. There are eight camera sub screen which can receive the video feeds from the CCTV cameras via RTSP protocols. For testing purpose, the videos have been taken from Internet and a proof-of-concept procedure is developed. The images shown are the camera feeds of the protesting and violent crowds and the respective police presences. Protest severity is determined and push notifications to the ten primary health care centers are sent to be ready for receiving patients with external injuries. This research can be used in different protests scenarios. For example, in a developing country like India, two recent protests have created a widespread headline all over the world. These protests are farmers protests [66] which is still in continuing and protests against the Citizenship Amendment Act [67] have given the impetus to authors to analyze the protests scenarios and minimize injuries to people, police and innocent bystanders. Below there are samples of images that highlights the proof of system images where feeds can be received on line from a connected CCTV / camera. As it has already been mentioned in Figure 1 of introduction section, the concept is to make a full fledge system where a command-and-control center deals with all the feeds. It is accompanied by the preprocessing and decision support system. All these operations converge onto crowd controlling deployable system using the deep learning framework.

The video streams undergo deep learning based analysis and severity of protests is ascertained. The final results goes to the decision support system which helps the human operator to take a final call on the prevailing situation.

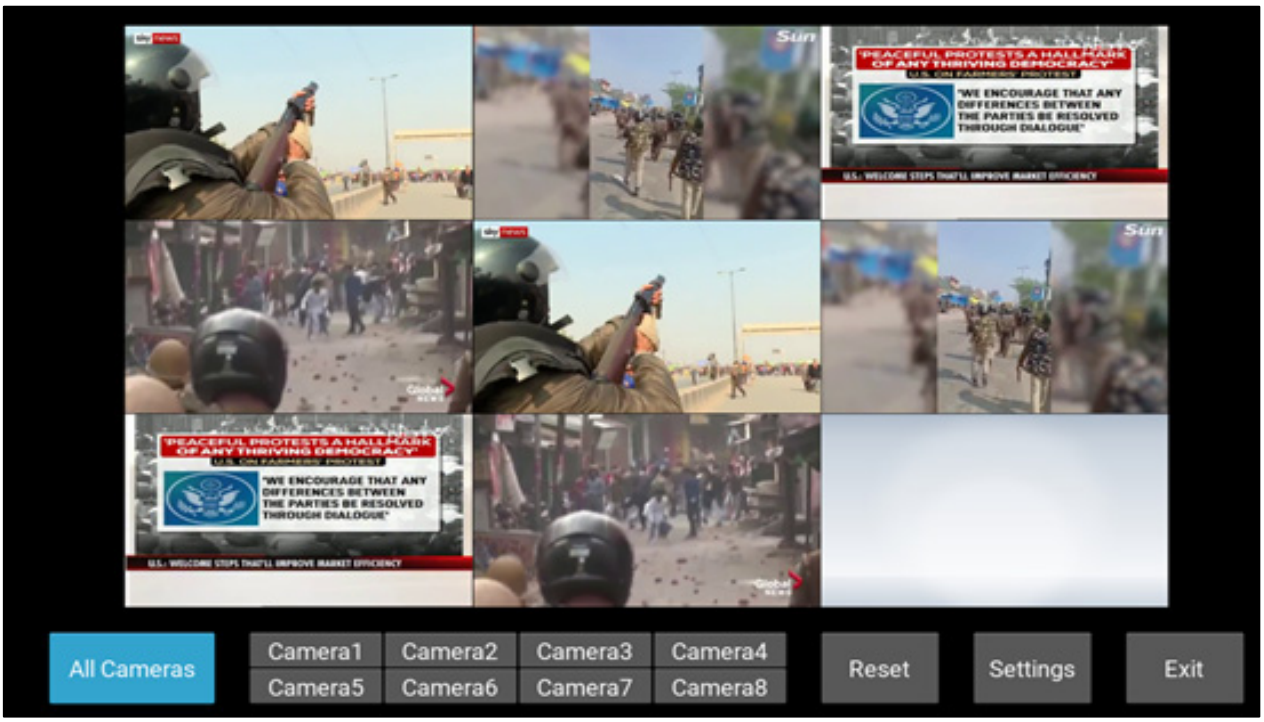


Figure 16. Main screen for the video feeds from the protest's sites.

The system shall take feeds from various CCTV feeds. For testing purpose, the system is shown to have eight screens. It can be extended to multiple screens. The system can see all feeds in one screen or if a user wants to explore the feeds, individual camera screen can be accessed.

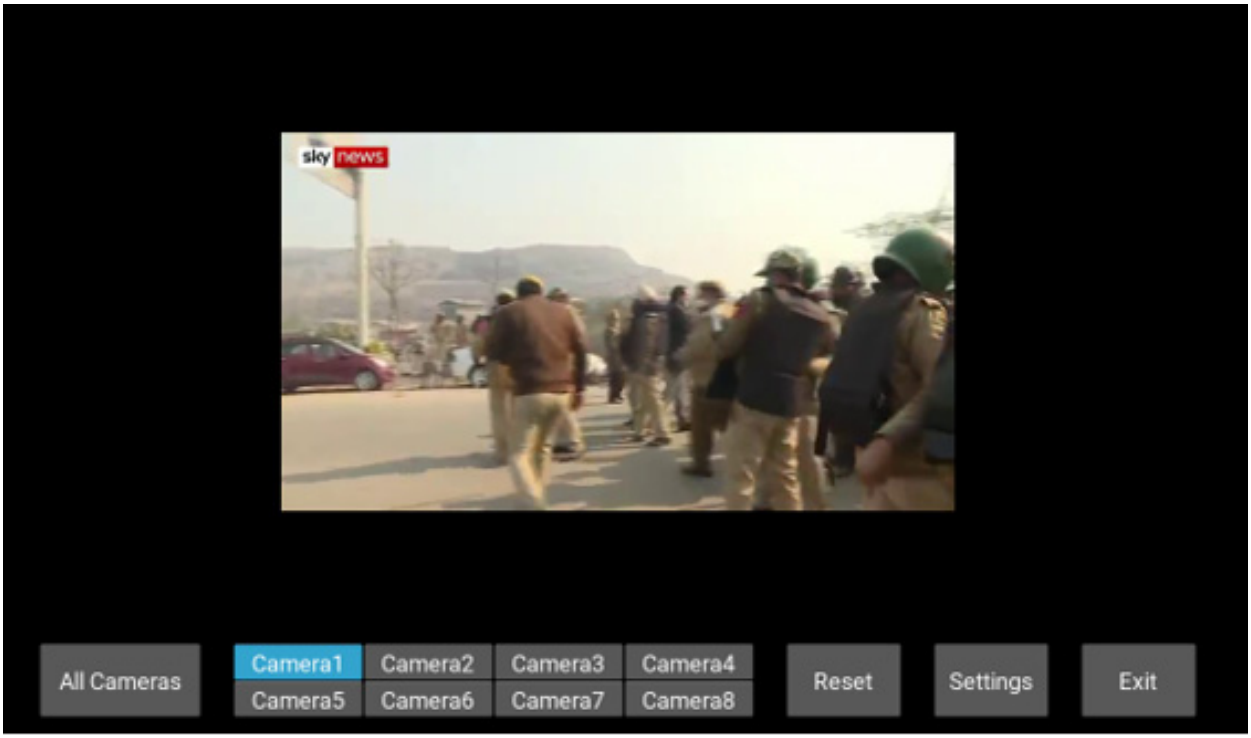


Figure 17. Camera 1 screen for the video feeds from the protest's sites**.**



Figure 18. Camera 2 screen for the video feeds from the protest's sites.

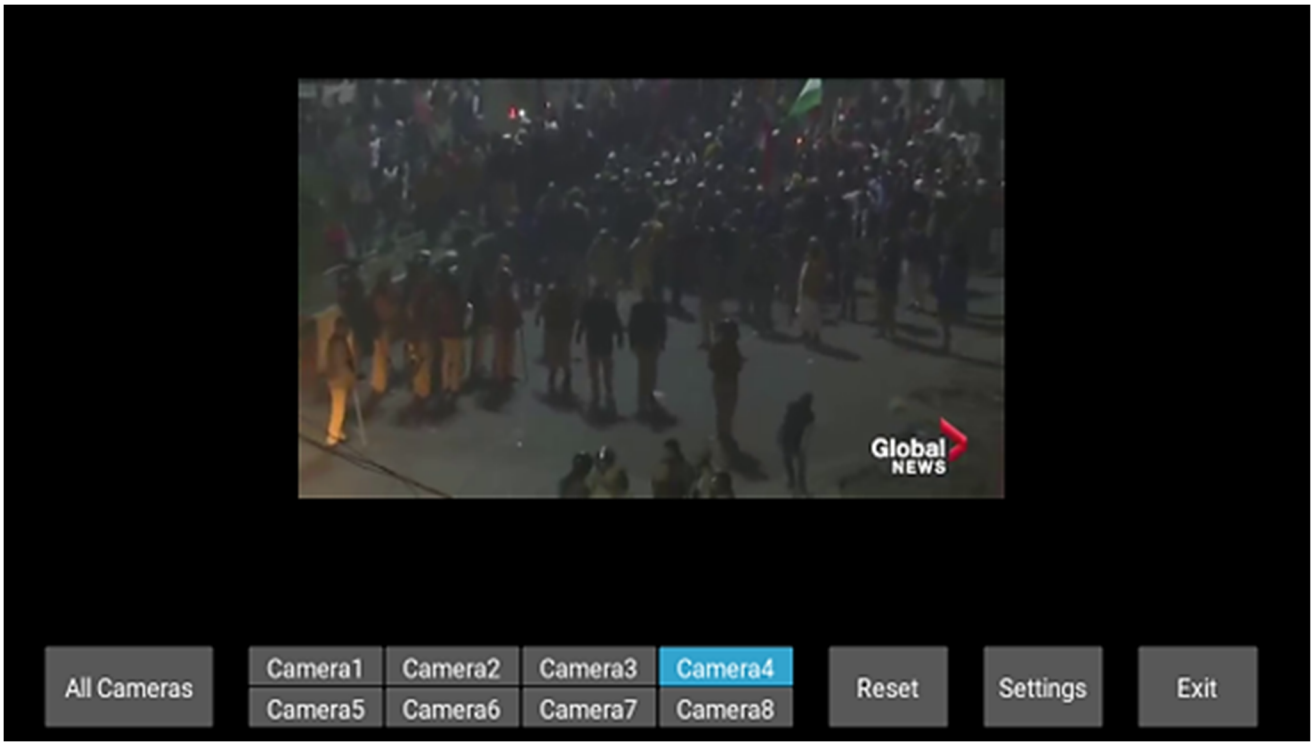


Figure 19. Camera 3 screen for the video feeds from the protest's sites.

The feeds that are received is ingested by the trained model. The model classifies it presence of various protests attributes. As soon as the protests attributes are received in the single frame, the rules are fired. Based on the rules, the ten nearest primary healthcare centers are identified and notifications to these centers are pushed via messaging application. The researchers assume that there are two levels of healthcare centers involved in this strategy. First the Primary Health Care centers of the Government, and top government hospitals and secondly all private hospitals which are well equipped to handle severe traumatic cases. The assumption is that in the actual implementation details all the health care centers would be marked on a map and a central authority would be responsible for sending the grievances messages to the nearest healthcare centers. Based on their responses, actual actions would be initiated to inform the driver of the ambulance which in turn can take the injured persons to the nearest healthcare centers.

7. Conclusions

This research is a unique attempt to link the protest storyboard to healthcare systems. This research uses deep learning for multiclass classification of the protest datasets. Based on the respective thresholds of the protest's attributes, a rule set has been presented which classifies the severity of protests in to normal, medium and severe category. Also, it shows an exclusive differentiation between protests and violence. In case of violence a direct tagging of healthcare infrastructures and doctors can be raised by the human in the loop. This is a novel study to alert the specialist healthcare doctors to be on standby in case the severity of protests breaches the thresholds. This framework can be integrated further with an IoT environment and Android application so that respective messaging push notification can be send to directly to all the concerned specialist healthcare personnel. The researchers

have assumed that it is full fledge Internet based system with linkage to CCTV cameras, the video streaming, the geo-location of all the healthcare centres situated in the city on a map, and a human in the loop operator which can guide the messaging of injured person to the nearest health care centres. As soon as the healthcare institution accepts the injured person, the system moves on the next phase of taking care of the injured persons with specialist doctors, surgeons, equipment and so on. The main objective is to save human lives of the protestors as well as the police personnel. Automatic surveillance becomes a lot easier and any sort of protests in public places can be tactfully handled. A real time framework of assessing the protests site can help prepare for any eventuality. The current research can be a big boon for tagging healthcare systems and can ensure availability of specialist doctors as per the severity of doctors, medicines, ambulances and other healthcare entities.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. D. Singh, G. Tripathi, A. J. Jara, A survey of Internet-of-Things: Future vision, architecture, challenges and services, in *2014 IEEE world forum on Internet of Things (WF-IoT)*, 2014.
2. C. Guy, Wireless sensor networks, in *Sixth international symposium on instrumentation and control technology: Signal analysis, measurement theory, photoelectronic technology, and artificial intelligence*, (2006), 635711.
3. N. S. Kumar, B. Vuayalakshmi, R. J. Prarthana, A. Shankar, IOT based smart garbage alert system using Arduino UNO, in *IEEE Region 10 International Conference TENCON*, 1028–1034.
4. M. S. Munir, I. S. Bajwa, A. Ashraf, W. Anwar, R. Rashid, Intelligent and Smart Irrigation System Using Edge Computing and IoT, *Complexity*, 2021.
5. R. Rajavel, S. K. Ravichandran, K. Harimoorthy, P. Nagappan, K. R. Gobichettipalayam, IoT-based smart healthcare video surveillance system using edge computing, *J. Ambient Intell. Humaniz Comput.*, **3** (2021), 1-13
6. F. Malik, M. A. Shah, H. A. Khattak, Intelligent transport system: An important aspect of emergency management in smart cities, in *2018 24th International Conference on Automation and Computing (ICAC)*, 2018.
7. A. H. Sodhro, M. S. Obaidat, A. Gurtov, N. Zahid, S. Pirbhulal, L. Wang, et al., Towards wearable sensing enabled healthcare framework for elderly patients, in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, 2020.
8. A. H. Sodhro, A. Gurtov, N. Zahid, S. Pirbhulal, L. Wang, M. Rahman, et al., Toward convergence of AI and IoT for energy-efficient communication in smart homes, *IEEE Int. Things J.*, **8** (2020), 9664–9671.
9. J. V. Stekelenburg, B. Klandermans, The social psychology of protest, *Curr. Sociol.*, (2013), 886–905
10. J. Joo, W. Li, F. F. Steen, S. C. Zhu, Visual persuasion: Inferring communicative intents of images, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.
11. W. S. DeKeseredy, B. Perry, Advancing critical criminology: theory and application, *Lexington Books*, 2006.

12. https://www.merriam-webster.com/dictionary/protest, Available: https://www.merriam-webster.com/dictionary/protest. [Accessed 13.2.2021 2 2021.]
13. https://www.macmillandictionary.com/dictionary/british/protest_1, Available: https://www.macmillandictionary.com/dictionary/british/protest_1. [Accessed 13 02 2021.]
14. Z. C. Steinert-Threlkeld, Spontaneous collective action: Peripheral mobilization during the Arab Spring, *Am. Polit. Sci. Rev.*, (2017), 379–403
15. Z. C. Steinert-Threlkeld, D. Mocanu, A. Vespignani, J. Fowler, Online social networks and offline protest, *EPJ Data Sci.*, (2015), 1–9
16. S. González-Bailón, J. Borge-Holthoefer, Y. Moreno, Broadcasters and hidden influentials in online protest diffusion, *Am. Behav. Sci.*, (2013), 943–965.
17. R. Enikolopov, A. Makarin, M. Petrova, Social media and protest participation: Evidence from Russia, *Econometrica*, (2020), 1479–1514.
18. T. Senst, V. Eiselein, A. Kuhn, T. Sikora, Crowd Violence Detection Using Global Motion-Compensated Lagrangian Features and Scale-Sensitive Video-Level Representation, in *IEEE Transactions on Information Forensics and Security*, (2017), 2945–2956.
19. J. Lofland, Protest: studies of collective behaviour and social movements, *Routledge*, 2017.
20. H. Grabner, F. Nater, M. Druey, G. L. Van, Visual interestingness in image sequences, in *Proceedings of the 21st ACM international conference on Multimedia*, 2013.
21. M. Redi, N. O'Hare, R. Schifanella, M. Trevisiol, A. Jaimes, 6 seconds of sound and vision: Creativity in micro-videos, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
22. A. Krizhevsky, I. Sutskever, E. H. Geoffrey, Imagenet classification with deep convolutional neural networks, *Adv. Neur. Inf. Process. Syst.*, 2012.
23. M. Perez, A. C. Kot, A. Rocha, Detection of Real-world Fights in Surveillance Videos, *Process. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019.
24. K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, *Adv. Neur. Inf. Process. Syst.*, 2014.
25. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, *Comput. Sci.*, 2014.
26. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint, arXiv:1409.1556.
27. S. Christian, I. Sergey, V. Vincent, A. Alexander, Inception-v4 inception-resnet and the impact of residual connections on learning, *Comput. Sci.*, 2017.
28. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, et al., Imagenet large scale visual recognition challenge, *Int. J. Comput. Vis.*, (2015), 211–252
29. C. Lum, J. Schallhorn, F. Lum, S. Ramanathan, J. Oatts, A. G. de A. Campomanes, et al., Ocular injuries and blindness caused by crowd control measures in civil protests, *Ophthalmology*, 2020.
30. D. Won, Z. C. Steinert-Threlkeld, J. Joo, Protest activity detection and perceived violence estimation from social media images, in *Proceedings of the 25th ACM international conference on Multimedia*, 2017.
31. Y. Wang, Y. Li, J. Luo, Deciphering the 2016 μs presidential campaign in the twitter sphere: A comparison of the trumpists and clintonists, in *Tenth International AAAI Conference on Web and Social Media*, 2016.

32. D. Chen, K. Park, J. Joo, Understanding Gender Stereotypes and Electoral Success from Visual Self-presentations of Politicians in Social Media, *Adv. Comput. Sci. Profess.*, 2020.
33. D. Ganguly, M. H. Mofrad, A. Kovashka, Detecting sexually provocative images, in *IEEE Workshop on Applications of Computer Vision (WACV)*, 2017.
34. H. Peng, J. Li, Y. Song, R. Yang, R. Ranjan, P. S. Yu, Streaming Social Event Detection and Evolution Discovery in Heterogeneous Information Networks, *ACM Trans. Knowl. Discov. Data*, (2021), 1–33
35. G. Tripathi, K. Singh, D. K. Vishwakarma, Detecting Arson and Stone Pelting in Extreme Violence: A Deep Learning Based Identification Approach, in *International Conference on Intelligent Human Computer Interaction*, 2020.
36. H. Zhang, J. Pan, Casm: A deep-learning approach for identifying collective action events with text and image data from social media, *Sociol. Methodol.*, (2019), 1–57.
37. D. Thenmozhi, C. Aravindan, A. Shyamsunder, A. Viswanathan, A. K. Pujari, Extracting Protests from News Using LSTM models with different Attention Mechanisms, in *CLEF (Working Notes)*, 2019.
38. A. Hanson, K. Pnvr, S. Krishnagopal, L. Davis, Bidirectional Convolutional LSTM for the Detection of Violence in Videos, in *European Conference on Computer Vision*, 2018.
39. S. A. Sumon, M. T. Shahria, M. R. Goni, N. Hasan, A. Almarufuzzaman, R. M. Rahman, Violent Crowd Flow Detection Using Deep Learning, in *Asian Conference on Intelligent Information and Database Systems*, 2019.
40. W. Song, D. Zhang, X. Zhao, J. Yu, R. Zheng, A. Wang, A Novel Violent Video Detection Scheme Based on Modified 3D Convolutional Neural Networks, *IEEE Access*, (2019), 39172–39179
41. X. Xu, X. Wu, G. Wang, H. Wang, Violent Video Classification Based on Spatial-Temporal Cues Using Deep Learning, in *2018 11th International Symposium on Computational Intelligence and Design (ISCID)*, 2018.
42. C. Li, L. Zhu, D. Zhu, J. Chen, Z. Pan, X. Li, et al., End-to-end Multiplayer Violence Detection based on Deep 3D CNN, in *Proceedings of the 2018 VII International Conference on Network, Communication and Computing*, 2018.
43. D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, *Proc. IEEE Int. Conf. Comput.*, 2015.
44. Z. Zhou, M. Zhu, K. Yahya, Violence Behavior Detection Based on 3D-CNN, *Comput. Syst. Appl.*, (2017), 34.
45. B. Mandal, J. Fajtl, V. Argyriou, D. Monekosso, P. Remagnino, Deep residual network with subclass discriminant analysis for crowd behavior recognition, in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018.
46. S. Ammar, M. R. Anjum, T. Rounak, M. Islam, Using deep learning algorithms to detect violent activities, *Semant. Schol.*, 2019.
47. Z. Meng, J. Yuan, Z. Li, Trajectory-Pooled Deep Convolutional Networks for Violence Detection in Videos, in *International Conference on Computer Vision Systems*, 2017.
48. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
49. R. E. Fan, K. W. Chang, C. J. Hsieh, X. R Wang, C. J. Lin, LIBLINEAR: A library for large linear classification, *J. Mach. Learn.Res.*, **9** (2008), 1871–1874.
50. J. Duchi, E. Hazan, Y. Singer, Adaptive Subgradient Methods for Online Learning and Stochastic Optimization, *J. Mach. Learn. Res.*, **12** (2011), 2121–2159.

51. H. Iqbal, Harisiqbal88/plotneuralnet v1. 0.0, URL: https://doi. org/10.5281/Zenodo, 2018.
52. I. Amnesty, Police-brutality, *Amnesty Int.*, 2021.
53. S. Venkatesan, The Riot Trauma: What Injuries Should You Expect From Non-Lethal Police Weapons and Protests?, *Trauma, Soc. EM, EMS*, 2020.
54. M. A. Kadivar, N. Ketchley, Sticks, stones, and Molotov cocktails: Unarmed collective violence and democratization, in *Socius: Sociological Research for a Dynamic World*, **4** (2018), 2378023118773614.
55. B. E. Mangus, L. Y. Shen, S. D. Helmer, J. Mahe, R. S. Smith, Taser and Taser associated injuries: a case series, *Am. Surg.*, **74** (2008), 862–865.
56. K. M. Buchanan, L. J. Elias, G. B. Goplen, Differing perspectives on outcome after subarachnoid hemorrhage: the patient, the relative, the neurosurgeon, *Neurosurgery*, **46** (2000), 831–840.
57. D. K. Raja, D. R. Raja, V. N. Shakul, The role of maxillofacial surgeon in the management of skull base tumors," *IP Int. J. Maxillofac. Imaging*, **2** (2021), 107–109.
58. A. Younas, I. Shah, T. C. Lim, M. Figari, G. Louie, D. Matic, et al., Evaluating an International Facial Trauma Course for Surgeons: Did We Make a Difference?, in *Craniomaxillo-facial Trauma Reconstruction Open*, **6** (2021), 24727512211019245.
59. E. Y. Y. Chan, K. K. C. Hung, H. H. Y. Hung, C. A. Graham, Use of tear gas for crowd control in Hong Kong, *The Lancet*, **394** (2019), 1517–1518.
60. K. Pauley, R. Flin, A. Azuara-Blanco, Intra-operative decision making by ophthalmic surgeons, *Br. J. Ophthalmol.* **97** (2013), 1300–1307.
61. J. W. Eikelboom, G. Karthikeyan, N. Fagel, J. Hirsh, American Association of Orthopedic Surgeons and American College of Chest Physicians guidelines for venous thromboembolism prevention in hip and knee arthroplasty differ: what are the implications for clinicians and patients?, *Chest*, **135** (2009), 513–520.
62. S. Budd, E. C. Robinson, B. Kainz, A survey on active learning and human-in-the-loop deep learning for medical image analysis, *Med. Image Anal.*, (2021), 102062.
63. M. Kuhn, K. Johnson, Applied Predictive Modelling, 2013.
64. S. H. Wang, K. Wu, T. Chu, S. L. Fernandes, Q. Zhou, Y. D. Zhang, et al., SOSPCNN: Structurally Optimized Stochastic Pooling Convolutional Neural Network for Tetralogy of Fallot Recognition, *Wireless Commun. Mobile Comput.*, **2021** (2021).
65. Y. D. Zhang, S. C. Satapathy, D. S. Guttery, Improved breast cancer classification through combining graph convolutional network and convolutional neural network, *Inf. Process. Manage.*, **58** (2021),102439.
66. S. Narayanan, Understanding farmer protests in India, *Academ. Stand. Against Povert.*, **1** (2020), 1.
67. B. Bhushan, South atlantic quarterly, *S. Atl. Q.*, **120** (2021), 201–208.

AIMS Press

©2021 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)

Kragujevac Journal of Mathematics
Volume 48(4) (2024), Pages 629–646.

APPROXIMATION BY A COMPOSITION OF APOSTOL-GENOCCHI AND PĂLTĂNEA-DURRMEYER OPERATORS

NAV SHAKTI MISHRA[1] AND NAOKANT DEO[1]

Abstract. The present paper deals with the Durrmeyer construction of operators based on a class of orthogonal polynomials called Apostol-Genocchi polynomials. For the proposed operators, we first establish a global approximation result followed by its convergence estimate in terms of usual, r-th and weighted modulus of continuity. We further study the asymptotic type results such as the Voronovskaya theorem and quantitative Voronovskaya theorem. Moreover, we estimate the rate of pointwise convergence of the proposed operators for functions of bounded variation defined on the interval $(0,\infty)$. Finally, the results are validated through graphical representations and an absolute error table.

1. Introduction

Recently, Prakash et al. [20] proposed the following positive linear sequence of operators:

$$G_n^{\alpha,\lambda}(f;x)=\sum_{k=0}^{\infty}s_{n,k}^{\alpha,\lambda}(x)f\left(\frac{k}{n}\right),\quad x\in[0,\infty), \tag{1.1}$$

where $s_{n,k}^{\alpha,\lambda}(x)=e^{-nx}\left(\frac{1+\lambda e}{2}\right)^{\alpha}\frac{g_k^{\alpha}(nx;\lambda)}{k!}$ and $g_k^{\alpha}(x;\lambda)$ is the generalized Apostol-Genocchi polynomials of order α, which belong to the class of orthogonal polynomials. These polynomials were defined for a complex variable z, $|z|<\pi$ in [16]. However, in this study we limit ourselves to a real variable $t\in[0,\infty)$. The generalized Apostol

Key words and phrases. Apostol-Genocchi polynomials, Păltănea basis, generating functions, special functions, functions of bounded variation.

2020 *Mathematics Subject Classification.* Primary: 41A25. Secondary: 41A36, 11B83.

DOI

Received: April 02, 2021.

Accepted: August 06, 2021.

Genocchi polynomial of order α, i.e., $g_k^{\alpha}(x;\lambda)$ can be estimated with the help of following generating function:

$$\left(\frac{2t}{1+\lambda e^t}\right)^{\alpha} e^{xt} = \sum_{k=0}^{\infty} g_k^{\alpha}(x;\lambda)\frac{t^k}{k!}. \tag{1.2}$$

The more explicit form of $g_k^{\alpha}(x;\lambda)$ was proposed by Luo and Srivastava in [17]. They presented some elementary properties of these polynomials and derived explicit series representation of $g_k^{\alpha}(x;\lambda)$ in terms of hypergeometric function defined by Gauss. The series is given as follows:

$$\begin{aligned} g_k^{\alpha}(x;\lambda) =& 2^{\alpha}\alpha!\binom{k}{\alpha}\sum_{n=0}^{k-\alpha}\binom{k-\alpha}{n}\binom{\alpha+n-1}{n}\frac{\lambda^n}{(1+\lambda)^{\alpha+n}} \\ &\times\sum_{j=0}^{n}(-1)^j\binom{n}{j}j^n(x+j)^{k-n-\alpha}{}_2F_1\left(\alpha+n-k,n;n+1;\frac{j}{x+j}\right), \end{aligned}$$

where $\{k,\alpha\}\in\mathbb{N}\cup\{0\}$, $\lambda\in\mathbb{R}\setminus\{-1\}$, $x\in\mathbb{R}$ and ${}_2F_1(a,b;c;t)$ denotes the Gaussian hypergeometric function defined by

$${}_2F_1(a,b;c;t) = {}_2F_1(b,a;c;t) = \sum_{k=0}^{\infty}\frac{(a)_k(b)_k}{(c)_k}\cdot\frac{t^k}{k!}.$$

In particular, for $\alpha=1$ and $\lambda=1$, these operators reduce to classical Genocchi polynomials which are obtained by the following generating function:

$$\frac{2te^{xt}}{e^t+1} = \sum_{k=0}^{\infty} g_k(x)\frac{t^k}{k!}, \quad |t|<2\pi,\ x\in\mathbb{R},$$

where $g_k(x) = g_k(x;1)$. It can be clearly seen that $g_k(x)$ are the k^{th}-degree polynomials, few terms of which are given as follows:

$$g_1(x)=1, \quad g_2(x)=2x-1, \quad g_3(x)=3x(x-1),$$
$$g_4(x)=4x^3-6x^2+1, \quad g_5(x)=5x^4-10x^3+5x,\ldots$$

For the case $x=0$, one can obtain the so-called Genocchi numbers g_k using the relation:

$$g_k(x) = \sum_{i=0}^{k}\binom{k}{i}g_i x^{k-i}.$$

Genocchi numbers can be defined in many ways depending on the field where they are intended to be applied. They find a wide range of application in numerical analysis, combinatorics, number theory, graph theory etc. Luo [15,16] defined Apostol-Genocchi polynomials of higher order and also introduced q-Apostol-Genocchi polynomials. He studied the relationship of these polynomials with Zeta function. In the last two decades, a surprising number of papers appeared studying Genocchi numbers, their combinatorial relations, Genocchi polynomials and their generalisations along with their various expansions and integral representations. To the readers, we suggest following articles [4, 18, 21] and references therein.

In the recent past, much work has been dedicated towards the Durrmeyer type modification of linear positive operators. For instance, Dhamija and Deo [8] introduced the Durrmeyer form of Jain operators based on inverse Pólya-Eggenberger distribution. They studied its moments with the aid of Vandermonde convolution formula and analysed other approximation properties. Heilmann and Raşa [12] studied a link between Baskakov-Durrmeyer type operators and their corresponding classical Kantorovich variants. Acu and Radu [3] introduced and studied a class of operators which link α-Bernstein operators and genuine α-Bernstein Durrmeyer operators. To see more work relevant to this area, one may refer [2, 5, 7, 10, 11, 13].

Inspired by above stated researches, we now consider a Durrmeyer type modification of Apostol-Genocchi operators based on Păltănea basis on positive real line. For $f \in C[0,\infty)$ and $\rho > 0$, the operators are defined as follows:

$$\mathcal{M}_n^{\alpha,\lambda}(f;x) = \sum_{k=0}^{\infty} s_{n,k}^{\alpha,\lambda}(x) \int_0^{\infty} l_{n,k}^{\rho}(t) f(t)\, dt, \quad x \in [0,\infty), \tag{1.3}$$

where $l_{n,k}^{\rho}(t) = n\rho e^{-n\rho t} \frac{(n\rho t)^{k\rho-1}}{\Gamma(k\rho)}$ and $s_{n,k}^{\alpha,\lambda}(x)$ is defined in (1.1).

The outline of the study is as follows. We consider a Durrmeyer type construction of Apostol-Genocchi operators based on the basis function due to Păltănea [19] with real parameters α, λ and ρ. We establish approximation estimates such as a global approximation theorem and rate of approximation in terms of usual, r-th and weighted modulus of continuity. We further study asymptotic formulae such as Voronovskaya theorem and quantitative Voronovskaya theorem. The last theorem is an application of the proposed operators for the functions whose derivatives are of bounded variation. Moreover, the approximation and the absolute error therein has been shown graphically by varying the values of various parameters using Mathematica software.

2. Preliminaries

Before proceeding to our main results, we state some general lemmas which are useful throughout this paper. In addition, we have used Mathematica software wherever necessary for complex and tedious calculations such as for moments and central moments etc.

Lemma 2.1. *For $e_s(t) = t^s$, $s \in \mathbb{N} \cup \{0\}$ and $\rho > 0$, we have*

$$\int_0^{\infty} l_{n,k}^{\rho}(t) t^s = \frac{(k\rho+s-1)!}{(n\rho)^s (k\rho-1)!} = \frac{(k\rho)_s}{(n\rho)^s},$$

where the symbol $(\beta)_n = \beta(\beta+1)(\beta+2)\cdots(\beta+n-1)$, $(\beta)_0 = 1$ denotes the rising factorial.

Lemma 2.2. *For operators (1.3), the moments are obtained as follows:*

$\mathcal{M}_n^{\alpha,\lambda}(e_0;x) = 1,$

$$\mathcal{M}_n^{\alpha,\lambda}(e_1;x) = x + \frac{\alpha}{n(1+\lambda e)},$$

$$\mathcal{M}_n^{\alpha,\lambda}(e_2;x) = x^2 + \frac{x}{n}\left[\frac{1+2\alpha+\lambda e}{(1+\lambda e)} + \frac{1}{\rho}\right] + \frac{1}{n^2}\left[\frac{\alpha^2 - 2\alpha\lambda e - \alpha e^2\lambda^2}{(1+\lambda e)^2} + \frac{\alpha}{\rho(1+\lambda e)}\right],$$

$$\begin{aligned}
&\mathcal{M}_n^{\alpha,\lambda}(e_3;x)\\
=&x^3 + \frac{x^2}{n}\left[\frac{3(\alpha+\lambda e+1)}{(1+\lambda e)} + \frac{3}{\rho}\right]\\
&+ \frac{x}{n^2}\left[\frac{3\alpha^2 - 3\alpha\lambda^2e^2 - 3\alpha\lambda e + 3\alpha + \lambda^2e^2 + 2\lambda e + 1}{(1+\lambda e)^2} + \frac{3(2\alpha+\lambda e+1)}{\rho(1+\lambda e)} + \frac{2}{\rho^2}\right]\\
&+ \frac{1}{n^3}\left[\frac{\alpha^3 - 3\alpha^2\lambda^2e^2 - 6\alpha^2\lambda e - \alpha\lambda^3e^3 - 4\alpha\lambda^2e^2 - 5\alpha\lambda e}{(1+\lambda e)^3} + \frac{2\alpha}{\rho^2(1+\lambda e)}\right.\\
&\left.+\frac{3(\alpha^2 - \alpha\lambda^2e^2 - 2\alpha\lambda e)}{\rho(1+\lambda e)^2}\right],
\end{aligned}$$

$$\begin{aligned}
&\mathcal{M}_n^{\alpha,\lambda}(e_4;x)\\
=&x^4 + \frac{x^3}{n}\left[\frac{2\alpha+3\lambda e+3}{(1+\lambda e)} + \frac{6}{\rho}\right]\\
&+ \frac{x^2}{n^2}\left[\frac{25 + 12\alpha + 6\alpha^2 + 50\mathrm{e}\lambda + 25\mathrm{e}^2\lambda^2 - 6\alpha\mathrm{e}^2\lambda^2}{(1+\lambda e)^2} + \frac{18(1+\alpha+\mathrm{e}\lambda)}{(1+\lambda e)\rho} + \frac{11}{\rho^2}\right]\\
&+ \frac{x}{n^3}\left[\frac{7 + 20\alpha + 63\alpha^2 + 2\alpha^3 - 6\alpha^2e^2\lambda^2 - 9\alpha^2\lambda e}{(1+\lambda e)^3}\right.\\
&+ \frac{-5\alpha e^3\lambda^3 + 4\alpha e^2\lambda^2 + 24\alpha\lambda e + 7e^3\lambda^3 + 21\lambda e}{(1+\lambda e)^3}\\
&\left.+\frac{6(3\alpha^2 - 3\alpha e^2\lambda^2 - 3\alpha\lambda e + 3\alpha + e^2\lambda^2 + 2\lambda e + 1)}{(1+\lambda e)^2\rho} + \frac{11(1+2\alpha+\lambda e)}{(1+\lambda e)\rho^2} + \frac{6}{\rho^3}\right]\\
&\times\frac{1}{n^4}\left[\frac{\alpha^4 - 6e^2\alpha^3\lambda - 12e\alpha^3\lambda - 16e^4\alpha\lambda^4 + 8e^3\alpha\lambda^3 - 82e^3\alpha\lambda^2 - 118e^2\alpha\lambda^2 - 66\alpha + \lambda e}{(1+\lambda e)^4}\right.\\
&+ \frac{6(\alpha^3 - 3e^2\alpha^2\lambda^2 - 6e\alpha^2\lambda - e^3\alpha\lambda^3 - 4e^2\alpha\lambda^2 - 5e\alpha\lambda)}{(1+\lambda e)^3\rho}\\
&\left.+ \frac{11(\alpha^2 - \alpha e^2\lambda^2 - 2\alpha\lambda e)}{(1+\lambda e)^2\rho^2} + \frac{6\alpha}{(1+\lambda e)\rho^3}\right].
\end{aligned}$$

Proof. In the proposed operators (1.3), for $s = 0, 1, 2$ respectively we have 1.

2. For $s = 1$, again using Lemma 2.1 we have

$$\mathcal{M}_n^{\alpha,\lambda}(e_1;x) = \frac{e^{-nx}}{n}\left(\frac{1+\lambda e}{2}\right)^{\alpha}\sum_{k=0}^{\infty}\frac{g_k^{\alpha}(nx;\lambda)}{k!}k. \tag{2.1}$$

Differentiating both sides of (1.2) with respect to t and taking limits $t \to 1$ and $x \to nx$, we have

$$\sum_{k=0}^{\infty} \frac{g_k^{\alpha}(nx;\lambda)}{k!} k = 2^{\alpha} e^{nx} \left(\frac{1}{1+\lambda e}\right)^{1+\alpha} (\alpha + nx(1+\lambda e)).$$

Making use of this value in equation (2.1), we obtain the first moment.

3. Similarly for $s = 2$, we have

$$\mathcal{M}_n^{\alpha,\lambda}(e_2;x) = \frac{e^{-nx}}{n^2} \left(\frac{1+\lambda e}{2}\right)^{\alpha} \sum_{k=0}^{\infty} \frac{g_k^{\alpha}(nx;\lambda)}{k!} k^2. \tag{2.2}$$

On differentiating both sides of (1.2) with respect to t and taking limits $t \to 1$ and $x \to nx$, we have

$$\begin{aligned}\sum_{k=0}^{\infty} \frac{g_k^{\alpha}(nx;\lambda)}{k!} k(k-1) =& 2^{\alpha} e^{nx} \left(\frac{1}{1+\lambda e}\right)^{1+\alpha} \{2ns\alpha(1+\lambda e) \\ &+ n^2x^2(1+\lambda e)^2 - (\lambda e(3+\lambda e) - \alpha + 1)\}.\end{aligned}$$

Combining this with the first order moment and equation (2.2) we obtain the third moment.

We can obtain the higher order moments in a similar way. □

Lemma 2.3. *Let us define* $\delta_n^s(x) = \mathcal{M}_n^{\alpha,\lambda}(\Phi_s;x)$, *where* $\Phi_s(t) = (e_1 - x)^s$ *and* $s = 1, 2$. *Then, from Lemma* 2.2 *we have*

$$\begin{aligned}\delta_n^{(1)}(x) &= \frac{\alpha}{n(1+\lambda e)}, \\ \delta_n^{(2)}(x) &= \frac{x}{n}\left[1 + \frac{1}{\rho}\right] + \frac{1}{n^2}\left[\frac{\alpha^2 - 2\alpha\lambda e - \alpha e^2\lambda^2}{(1+\lambda e)^2} + \frac{\alpha}{\rho(1+\lambda e)}\right].\end{aligned}$$

Furthermore,

$$\begin{aligned}\lim_{n\to\infty} n\mathcal{M}_n^{\alpha,\lambda}(\Phi_1;x) &= \frac{\alpha}{(1+\lambda e)}, \\ \lim_{n\to\infty} n\mathcal{M}_n^{\alpha,\lambda}(\Phi_2;x) &= \left(1 + \frac{1}{\rho}\right)x, \\ \lim_{n\to\infty} n^2\mathcal{M}_n^{\alpha,\lambda}(\Phi_4;x) &= \frac{3(1 + (6+8\alpha)\rho - (3+8\alpha)\rho^2 + \lambda e(1 + 6\rho - 3\rho^2))x^2}{(1+\lambda e)\rho^2}\end{aligned}$$

and

$$\lim_{n\to\infty} n^3\mathcal{M}_n^{\alpha,\lambda}(\Phi_6;x) = \frac{(1+\rho)(2+\rho+3\alpha\rho+\lambda e(2+\rho))x^5}{(1+\lambda e)^3\rho^4}.$$

Remark 2.1. Since fourth and sixth central moments are too lengthy and unnecessarily space consuming, we are omitting their values here. Instead we choose to write their limiting values, which is useful in the proofs of our main theorems.

3. Main Theorems

Theorem 3.1. *For any $f \in C_B[0,\infty)$, where $C_B[0,\infty)$ is the class of all continuous and bounded functions, we have*

$$\mathcal{M}_n^{\alpha,\lambda}(f(t);x) = f(x)$$

uniformly on any compact subset of $[0,\infty)$.

Proof. Taking into account Lemma 2.2, we can easily see that $\mathcal{M}_n^{\alpha,\lambda}(e_r;x) \to x^r$ for each $r = 0,1,2$ and hence using the well known Korovkin's theorem due to [14], operators $\mathcal{M}_n^{\alpha,\lambda}$ converge uniformly on each compact subset of $[0,\infty)$. □

3.1. **Global approximation.** Let us denote $B_f[0,\infty)$ the space of all functions f on positive real axis that satisfy the condition $|f(x)| \le H_f(1+x^2)$ where H_f is a constant depending only on f but independent of x.

Let $C_f[0,\infty)$ be the subspace of $B_f[0,\infty)$ containing all continuous f on $[0,\infty)$. The norm in $C_f[0,\infty)$ is defined by

$$||f||_2 = \sup_{x\in[0,\infty)} \frac{|f(x)|}{1+x^2}.$$

Also, let $C_f^l[0,\infty) := \left\{ f \in C_f[0,\infty) : \lim_{x\to\infty} \frac{|f(x)|}{1+x^2} \text{ is finite} \right\}$.

Theorem 3.2. *For each $f \in C_f^l[0,\infty)$, we have*

$$\lim_{n\to\infty} \left\| \mathcal{M}_n^{\alpha,\lambda}(f;x) - f \right\|_2 = 0.$$

Proof. The proof of this theorem can be given by application of Korovkin theorem [9] on the interval $[0,\infty)$. Therefore, it would suffice if we prove that

$$\lim_{n\to\infty} \left\| \mathcal{M}_n^{\alpha,\lambda}(e_l;x) - e_l \right\|_2 = 0, \quad l = 0,1,2. \tag{3.1}$$

For $l = 0$, condition (3.1) holds as operators $\mathcal{M}_n^{\alpha,\lambda}$ preserve constant functions. Next, we can write

$$\left\| \mathcal{M}_n^{\alpha,\lambda}(e_1;x) - x \right\|_2 \le \sup_{x\in[0,\infty)} \frac{\alpha}{n(1+\lambda e)(1+x^2)} \to 0,$$

for adequately large n. Therefore, the condition (3.1) is satisfied for $l = 1$.

Finally, we write

$$\left\| \mathcal{M}_n^{\alpha,\lambda}(e_2;x) - x^2 \right\|_2 \le \sup_{x\in[0,\infty)} \frac{1}{(1+x^2)} \left\{ \frac{x}{n}\left(1+\frac{1}{\rho}\right) + \frac{1}{n^2}\left(\frac{\alpha^2 + 2\alpha\lambda e - \alpha e^2\lambda^2}{(1+\lambda e)^2} + \frac{\alpha}{\rho(1+\lambda e)} \right) \right\},$$

which suggests

$$\lim_{n\to\infty} \left\| \mathcal{M}_n^{\alpha,\lambda}(e_2;x) - e_2 \right\|_2 = 0.$$

Hence, the theorem follows. □

Let $C_B[0,\infty)$ be the class of all continuous and bounded real valued functions. We define the r-th order modulus of continuity by $\omega_r(f,\delta)$ and define it as

$$\omega_r(f,\delta) = \sup_{x\in[0,\infty)} \sup_{0\le h\le\delta} |\Delta_h^r f(x)|,$$

where Δ denotes the forward difference. In particular, the usual modulus of continuity is defined for $r=1$ and is denoted by $\omega(f,\delta)$. Moreover, we define the norm as $\|f\| = \sup_{x\in[0,\infty)} |f(x)|$.

Also, the Peetre's K-functional for the function $g \in C_B^2[0,\infty)$ is defined as:

$$K_2(f;\delta) = \inf_{g\in C_B^2[0,\infty)} \left\{ \|f-g\| + \delta \|g''\| : g \in C_B^2[0,\infty) \right\},$$

where

$$C_B^2[0,\infty) = \{g \in C_B[0,\infty) : g', g'' \in C_B[0,\infty)\}.$$

The next theorem establishes the degree of approximation of the operators $\mathcal{M}_n^{\alpha,\lambda}$ in terms of the usual and second order modulus of continuity for the functions in the space $C_B[0,\infty)$.

Theorem 3.3. *For $\hbar \in C_B^2[0,\infty)$, define the auxiliary operators $\widetilde{\mathcal{Q}}_n^{\alpha,\lambda}$ as*

$$\widetilde{\mathcal{Q}}_n^{\alpha,\lambda}(\hbar;x) = \mathcal{M}_n^{\alpha,\lambda}(\hbar;x) - \hbar\left(x + \frac{\alpha}{n(1+\lambda e)}\right) + \hbar(x). \tag{3.2}$$

Then there exists a constant $C>0$ such that

$$|\mathcal{M}_n^{\alpha,\lambda}(\hbar,x) - \hbar(x)| \le C\omega_2(\hbar,\sqrt{\delta}) + \omega(\hbar,\delta_n^{(1)}(x)),$$

where

$$\delta = \delta_n^{(2)}(x) + \left(\frac{\alpha}{n(1+\lambda e)}\right)^2.$$

Proof. Using Lemma 2.2, one can easily observe that $\widetilde{\mathcal{Q}}_n^{\alpha,\lambda}((t-x);x) = 0$.

Let $f \in C_B^2[0,\infty)$, then by Taylor's expansion we have

$$f(t) = f(x) + (t-x)f'(x) + \int_x^t (t-u)f''(u)du.$$

Moreover, we can write

$$\begin{aligned}
\left|\widetilde{\mathcal{Q}}_n^{\alpha,\lambda}(f;x) - f(x)\right| &= \left|\widetilde{\mathcal{Q}}_n^{\alpha,\lambda}\left(\int_x^t (t-u)f''(u)du, x\right)\right| \\
&\le \left|\mathcal{M}_n^{\alpha,\lambda}\left(\int_x^t (t-u)f''(u)du, x\right)\right| \\
&\quad + \left|\int_x^{\left(x+\frac{\alpha}{n(1+\lambda e)}\right)} \left(\left(x + \frac{\alpha}{n(1+\lambda e)}\right) - u\right) f''(u)du\right|
\end{aligned}$$

$$\leq \left(\mathcal{M}_n^{\alpha,\lambda} \left((t-x)^2; x \right) + \left(\frac{\alpha}{n\left(1+\lambda e\right)} \right)^2 \right) \|f''\| . \tag{3.3}$$

Since we know that

$$\left| \mathcal{M}_n^{\alpha,\lambda} \left(\hbar; x \right) \right| \leq \|\hbar\| ,$$

therefore

$$\left| \widetilde{\mathcal{Q}}_n^{\alpha,\lambda}(\hbar; x) \right| \leq \left| \mathcal{M}_n^{\alpha,\lambda} \left(\hbar; x \right) \right| + \left| \hbar \left(x + \frac{\alpha}{n\left(1+\lambda e\right)} \right) \right| + |\hbar(x)| \leq 3 \|\hbar\| . \tag{3.4}$$

Finally, combining equations (3.2), (3.3) and (3.4), we get

$$\begin{aligned}
\left| \mathcal{M}_n^{\alpha,\lambda} \left(\hbar; x \right) - \hbar(x) \right| \leq & \left| \widetilde{\mathcal{Q}}_n^{\alpha,\lambda}(\hbar - f); x) - (\hbar - f)(x) \right| + \left| \widetilde{\mathcal{Q}}_n^{\alpha,\lambda}(f; x) - f(x) \right| \\
& + \left| \hbar(x) - \hbar \left(x + \frac{\alpha}{n(1+\lambda e)} \right) \right| \\
\leq & 4 \|\hbar - f\| + \left(\mathcal{M}_n^{\alpha,\lambda} \left((t-x)^2; x \right) + \left(\frac{\alpha}{n\left(1+\lambda e\right)} \right)^2 \right) \|f''\| \\
& + \left| \hbar(x) - \hbar \left(x + \frac{\alpha}{n(1+\lambda e)} \right) \right| \\
\leq & C \left\{ \|\hbar - f\| + \left(\delta_n^{(2)}(x) + \left(\frac{\alpha}{n\left(1+\lambda e\right)} \right)^2 \right) \|f''\| \right\} \\
& + \omega \left(\hbar, \frac{\alpha}{n\left(1+\lambda e\right)} \right) .
\end{aligned}$$

Taking infimum over all $f \in C_B^2[0,\infty)$ and using the result $K_2(f,\delta) \leq \omega_2\left(f, \sqrt{\delta}\right)$ due to [6], we get the desired outcome. □

Theorem 3.4. *Let $\hbar \in C_B(0,\infty)$, then for any $r > 0$, $x \in [0,r]$ and adequately large n, we have*

$$\left| \mathcal{M}_n^{\alpha,\lambda} \left(\hbar; x \right) - \hbar(x) \right| \leqslant 4 H_\hbar \left(1 + x^2 \right) \frac{D}{n} + 2\omega_{r+1} \left(\hbar, \sqrt{\frac{D}{n}} \right) ,$$

where D is a positive constant.

Proof. If $x \in [0,r]$ and $t > r+1$, then $t - x > 1$. Therefore, we have the following inequality:

$$|\hbar(t) - \hbar(x)| \leqslant 4 H_\hbar \left(1 + x^2 \right) (t-x)^2 .$$

Again for $x \in [0,r]$ and $t \in [0, r+1]$ and using the well known inequality $\omega(f, \beta\delta) \leqslant (\beta+1)\,\omega(f,\delta)\,, \beta \in (0,\infty)$, one can obtain

$$|\hbar(t) - \hbar(x)| \leqslant \left(1 + \frac{|t-x|}{\delta} \right) \omega_{r+1}(\hbar, \delta) .$$

From (3.5) and (3.5), we can write

$$|\hbar(t)-\hbar(x)| \leqslant 4H_{\hbar}\left(1+x^2\right)(t-x)^2+\left(1+\frac{|t-x|}{\delta}\right)\omega_{r+1}(\hbar,\delta).$$

Applying operator $\mathcal{M}_n^{\alpha,\lambda}$ in the above relation and making use of Cauchy-Schwarz inequality, we get

$$\begin{aligned}\left|\mathcal{M}_n^{\alpha,\lambda}(\hbar;x)-\hbar(x)\right| \leqslant & 4H_{\hbar}\left(1+x^2\right)\mathcal{M}_n^{\alpha,\lambda}\left((t-x)^2;x\right)\\ &+\left(1+\frac{1}{\delta}\mathcal{M}_n^{\alpha,\lambda}(|t-x|;x)\right)\omega_{r+1}(\hbar,\delta)\\ \leqslant & 4H_{\hbar}\left(1+x^2\right)\mathcal{M}_n^{\alpha,\lambda}\left((t-x)^2;x\right)\\ &+2\omega_{r+1}\left(\hbar,\sqrt{\mathcal{M}_n^{\alpha,\lambda}\left((t-x)^2;x\right)}\right).\end{aligned}$$

Since $\mathcal{M}_n^{\alpha,\lambda}\left((t-x)^2;x\right) \leqslant \frac{D}{n}$, where D is a positive constant, it follows that for adequately large n, we have

$$\left|\mathcal{M}_n^{\alpha,\lambda}(\hbar;x)-\hbar(x)\right| \leqslant 4H_{\hbar}\left(1+x^2\right)\frac{D}{n}+2\omega_{r+1}\left(\hbar,\sqrt{\frac{D}{n}}\right),$$

which is the required result. □

3.2. **Quantitative Voronovskaya theorem and Voronovoskaya theorem.** Let $C_B[0,\infty)$ be the subspace of $B_f[0,\infty)$ containing all continuous and bounded functions f for which $\lim\limits_{x\to\infty}|f(x)|\left(1+x^2\right)^{-1}$ is finite.

The weighted modulus of continuity $\Omega(f,\delta)$ due to [1] for each $f \in C_B[0,\infty)$ is defined as

$$\Omega(f,\delta)=\sup_{x\in[0,\infty),|h|<\delta}\frac{|f(x+h)-f(x)|}{\left(1+h^2+x^2+h^2x^2\right)}.$$

In the next theorem, we discuss the quantitative Voronovskaya theorem for the proposed operators (1.3) and derive a Voronovoskaya asymptotic result as a resulting corollary, making use of the following properties of weighted modulus of continuity. For every $f \in C_f^l[0,\infty)$,

(a) $\Omega(f,\delta)\to 0$ for $\delta\to 0$;
(b) $|f(t)-f(x)| \leq \left(1+(t-x)^2\right)\left(1+x^2\right)\Omega(f,|t-x|)$.

Theorem 3.5. *Let $\hbar'' \in C_2^l[0,\infty)$ and $x\in[0,\infty)$. Then we have*

$$\begin{aligned}&\left|\mathcal{M}_n^{\alpha,\lambda}(\hbar;x)-\hbar(x)-\frac{\alpha}{n(1+\lambda e)}\hbar'(x)\right.\\ &\left.-\frac{1}{2}\left(\frac{x}{n}\left(1+\frac{1}{\rho}\right)+\frac{1}{n^2}\left(\frac{\alpha^2+2\alpha\lambda e-\alpha e^2\lambda^2}{(1+\lambda e)^2}+\frac{\alpha}{\rho(1+\lambda e)}\right)\right)\hbar''(x)\right|\end{aligned}$$

$$\leq \frac{8\left(1+x^2\right)}{n}\Omega\left(\hbar'', \frac{1}{\sqrt{n}}\right).$$

Proof. By Taylor's expansion, we may write

$$\begin{aligned}\mathcal{M}_n^{\alpha,\lambda}(\hbar;x)-\hbar(x) &= \mathcal{M}_n^{\alpha,\lambda}\left(\left(\hbar(t)-\hbar(x)\right);x\right)\\ &= \mathcal{M}_n^{\alpha,\lambda}\left(\left(t-x\right)\hbar'\left(x\right)+\frac{\left(t-x\right)^2}{2}\hbar''\left(x\right)+\lambda\left(t,x\right)\left(t-x\right)^2;x\right),\end{aligned}$$

where $\lambda\left(t,x\right)=\left(\hbar''\left(\varsigma\right)-\hbar''\left(x\right)\right)/2$ is a continuous function which tends to zero at 0 and ς lies between x and t. Using Lemma 2.3, we get

$$\begin{aligned}&\left|\mathcal{M}_n^{\alpha,\lambda}(\hbar;x)-\hbar(x)-\frac{\alpha}{n\left(1+\lambda e\right)}\hbar'\left(x\right)\right.\\ &\left.-\frac{1}{2}\left(\frac{x}{n}\left(1+\frac{1}{\rho}\right)+\frac{1}{n^2}\left(\frac{\alpha^2+2\alpha\lambda e-\alpha e^2\lambda^2}{\left(1+\lambda e\right)^2}+\frac{\alpha}{\rho\left(1+\lambda e\right)}\right)\right)\hbar''\left(x\right)\right|\\ \leq&\mathcal{M}_n^{\alpha,\lambda}\left(\left|\lambda\left(t,x\right)\right|\left(t-x\right)^2;x\right).\end{aligned}$$

With simple manipulations in property (b) of weighted modulus of continuity and using $|\varsigma-x|\leq|t-x|$, we can write

$$\left|\lambda\left(t,x\right)\right|\leq 8\left(1+x^2\right)\left(1+\frac{\left(t-x\right)^4}{\delta^4}\right)\Omega\left(\hbar'',\delta\right),$$

which implies that

$$\left|\lambda\left(t,x\right)\right|\left(t-x\right)^2\leq 8\left(1+x^2\right)\left(\left(t-x\right)^2+\frac{\left(t-x\right)^6}{\delta^4}\right)\Omega\left(\hbar'',\delta\right).$$

Therefore, in view of Lemma 2.3, we can write

$$\mathcal{M}_n^{\alpha,\lambda}\left(\left|\lambda\left(t,x\right)\right|\left(t-x\right)^2;x\right)\leq 8\left(1+x^2\right)\Omega\left(\hbar'',\delta\right)\left\{\delta_n^{(2)}\left(x\right)+\frac{1}{\delta^4}\delta_n^{(6)}\left(x\right)\right\},$$

as $n\to\infty$. Choosing $\delta=\frac{1}{\sqrt{n}}$, we get the desired outcome. □

Corollary 3.1. *Let f be a bounded and integrable function on the interval $[0,\infty)$ such that the second derivative of f exists at a fixed point $x\in[0,\infty)$. Then*

$$\lim_{n\to\infty} n(\mathcal{M}_n^{\alpha,\lambda}(f;x)-f(x))=\frac{\alpha}{(1+\lambda e)}f'(x)+x\left(1+\frac{1}{\rho}\right)f''(x).$$

3.3. **Functions of derivatives of bounded variation.** Next we estimate the rate of convergence of the operators (1.3) for functions with derivatives of bounded variation defined on $[0,\infty)$.

Let $f\in \mathrm{DBV}_\tau[0,\infty)$ be the class of functions whose derivatives are of bounded variation on any finite subinterval of $[0,\infty)$ and satisfy the growth condition $|f(t)|\leq$

Kt^{τ}, $\tau > 0$ for all $t > 0$ and constant $K > 0$. For such functions, let us represent our proposed operators (1.3) in the following form

$$\mathcal{M}_n^{\alpha,\lambda}(f;x) = \int_0^{\infty} \mathfrak{q}_{n,\rho}^{\alpha,\lambda}(x;t) f(t) dt, \tag{3.5}$$

where $\mathfrak{q}_{n,\rho}^{\alpha,\lambda}(x;t) = \sum_{k=0}^{\infty} s_{n,k}^{\alpha,\lambda}(x) l_{n,k}^{\rho}(t)$.

Lemma 3.1. *For $x \in [0,\infty)$ and adequately large n, we have*

(i) *if $0 \le y < x$, then*

$$\vartheta_n(x,y) = \int_0^y \mathfrak{q}_n^{\alpha,\rho}(x;t)\, dt \le \frac{K\delta_n^2(x)}{n(x-y)^2};$$

(ii) *if $x < z \le \infty$, then*

$$1 - \vartheta_n(x,z) = \int_z^{\infty} \mathfrak{q}_n^{\alpha,\rho}(x;t)\, dt \le \frac{K\delta_n^2(x)}{n(z-x)^2}.$$

Proof. (i) Taking into account Lemma 2.2 and proposed operators (1.3), we have

$$\begin{aligned}
\vartheta_n(x,z) &= \int_0^y \mathfrak{q}_n^{\alpha,\rho}(x;t)\, dt \\
&\le \int_0^y \mathfrak{q}_n^{\alpha,\rho}(x;t) \left(\frac{x-t}{x-y}\right)^2 dt = \frac{1}{(x-y)^2} \int_0^y (t-x)^2 \mathfrak{q}_n^{\alpha,\rho}(x;t)\, dt \\
&\le \frac{1}{(x-y)^2} \mathcal{M}_n^{\alpha,\lambda}\left((t-x)^2;x\right) \le \frac{K\delta_n^2(x)}{n(x-y)^2}.
\end{aligned}$$

Proof of (ii) is similar to (i). □

Theorem 3.6. *Consider a function f of bounded variation on every sub-interval of $[0,\infty)$ that satisfies the growth condition $|f(t)| \le Kt^{\tau}$ for some absolute constant K and $\tau > 0$. If there exists an integer γ, $(2\gamma \ge \tau)$ such that $f(t) \le O(t^{\gamma})$ for every $t > 0$, then for $\gamma > 0$, $x \in [0,\infty)$ and sufficiently large n, we have*

$$\begin{aligned}
\left|\mathcal{M}_n^{\alpha,\lambda}(f;x) - f(x)\right| \le & \frac{1}{2}\left(f'(x+) + f'(x-)\right)\delta_n^1(x) \\
& + \sqrt{\frac{K\delta_n^2(x)}{4n}}\left|f'(x+) - f'(x-)\right| + \frac{x}{\sqrt{n}} \bigvee_{x-\frac{x}{\sqrt{n}}}^{x+\frac{x}{\sqrt{n}}} (f'_x) \\
& + \frac{K\delta_n^2(x)}{nx} \sum_{k=1}^{[\sqrt{n}]} \bigvee_{x-\frac{x}{k}}^{x+\frac{x}{\sqrt{k}}} (f'_x) + \frac{K\delta_n^2(x)}{nx^2}\left|f(2x) - f(x) - xf'(x)\right| \\
& + \wp(\gamma,\tau,x) + \frac{K\delta_n^2(x)}{nx^2}\left|f(x)\right| + \sqrt{\frac{K\delta_n^2(x)}{n}} f'(x+),
\end{aligned}$$

where $\bigvee_{b}^{a}(f)$ *denotes the total variation of* f *on any finite subinterval* $[a,b]$ *of* $[0,\infty)$ *and* $\wp(\gamma,\tau,x) := 2^{\gamma}\text{Ç}\left(\int_0^{\infty}(t-x)^{2\tau}\mathfrak{q}_{n,\rho}^{\alpha,\lambda}(x;t)\,dt\right)^{\frac{\gamma}{2\tau}}$.

Proof. For $x\in[0,\infty)$, we can write for our proposed operators (3.5) that

$$\begin{aligned}\mathcal{M}_n^{\alpha,\lambda}(f;x)-f(x) &= \int_0^{\infty}\mathfrak{q}_{n,\rho}^{\alpha,\lambda}(x;t)\,(f(t)-f(x))\,dt\\ &= \int_0^{\infty}\mathfrak{q}_{n,\rho}^{\alpha,\lambda}(x;t)\left(\int_x^t f'(x)\,du\right)dt. \qquad (3.6)\end{aligned}$$

Also for any $f\in DBV_{\gamma}[0,\infty)$, equality (3.7) holds true, i.e.,

$$\begin{aligned}f'(u) =& \frac{1}{2}\left(f'(x+)+f'(x-)\right)+f'_x(u)+\frac{1}{2}\left(f'(x+)-f'(x-)\right)\operatorname{sgn}(u-x)\\ &+\delta_x(u)\left(f'(u)-\frac{1}{2}\left(f'(x+)+f'(x-)\right)\right), \qquad (3.7)\end{aligned}$$

where

$$\delta_x(u)=\begin{cases}1, & u=x,\\ 0, & u\neq x.\end{cases}$$

It can be easily verified that:

$$\int_0^{\infty}\left(\int_x^t\left(f'(u)-\frac{1}{2}\left(f'(x+)+f'(x-)\right)\right)\delta_x(u)du\right)\mathfrak{q}_{n,\rho}^{\alpha,\lambda}(x;t)\,dt=0.$$

Now in view of our proposed operators (3.5), we may write

$$\begin{aligned}&\int_0^{\infty}\left(\int_x^t\left(\frac{1}{2}\left(f'(x+)+f'(x-)\right)\right)du\right)\mathfrak{q}_{n,\rho}^{\alpha,\lambda}(x;t)\,dt\\ =&\frac{1}{2}\left(f'(x+)+f'(x-)\right)\mathcal{M}_n^{\alpha,\lambda}((t-x);x).\end{aligned}$$

Moreover

$$\begin{aligned}&\int_0^{\infty}\left(\int_x^t\left(\frac{1}{2}\left(f'(x+)-f'(x-)\right)\operatorname{sgn}(u-x)\right)du\right)\mathfrak{q}_{n,\rho}^{\alpha,\lambda}(x;t)\,dt\\ \leq&\frac{1}{2}\left|f'(x+)-f'(x-)\right|\int_0^{\infty}|t-x|\,\mathfrak{q}_{n,\rho}^{\alpha,\lambda}(x;t)\,dt\\ \leq&\frac{1}{2}\left|f'(x+)-f'(x-)\right|\left(\mathcal{M}_n^{\alpha,\lambda}\left((t-x)^2;x\right)\right)^{1/2}. \qquad (3.8)\end{aligned}$$

Making use of equations (3.7)–(3.8) and Lemma 2.2 in equation (3.6), we get

$$\mathcal{M}_n^{\alpha,\lambda}(f;x)-f(x)\leq\frac{1}{2}\left(f'(x+)+f'(x-)\right)\mathcal{M}_n^{\alpha,\lambda}((t-x);x)$$

$$
\begin{aligned}
&+\frac{1}{2}\left|f'\left(x+\right)-f'\left(x-\right)\right|\left(\mathcal{M}_n^{\alpha,\lambda}\left(\left(t-x\right)^2;x\right)\right)^{1/2}\\
&+\int_0^\infty\left(\int_x^t f'_x\left(x\right)du\right)\mathfrak{q}_{n,\rho}^{\alpha,\lambda}\left(x;t\right)dt\\
\leq&\frac{1}{2}\left(f'\left(x+\right)+f'\left(x-\right)\right)\mathcal{M}_n^{\alpha,\lambda}\left(\left(t-x\right);x\right)\\
&+\sqrt{\frac{K\delta_n^2\left(x\right)}{4n}}\left|f'\left(x+\right)-f'\left(x-\right)\right|\\
&+\int_0^\infty\left(\int_x^t f'_x\left(x\right)du\right)\mathfrak{q}_{n,\rho}^{\alpha,\lambda}\left(x;t\right)dt.
\end{aligned}
$$

Taking absolute values on both sides and rewriting equation we have:

$$
\begin{aligned}
\left|\mathcal{M}_n^{\alpha,\lambda}\left(f;x\right)-f\left(x\right)\right|\leq&\frac{1}{2}\left(f'\left(x+\right)+f'\left(x-\right)\right)\mathcal{M}_n^{\alpha,\lambda}\left(\left(t-x\right);x\right)\\
&+\sqrt{\frac{K\delta_n^2\left(x\right)}{4n}}\left|f'\left(x+\right)-f'\left(x-\right)\right|+P_{n_1}\left(x\right)+P_{n_2}\left(x\right),
\end{aligned}
\tag{3.9}
$$

where

$$
P_{n_1}\left(x\right)=\left|\int_0^x\left(\int_x^t f'_x\left(x\right)du\right)\mathfrak{q}_{n,\rho}^{\alpha,\lambda}\left(x;t\right)dt\right|
$$

and

$$
P_{n_2}\left(x\right)=\left|\int_x^\infty\left(\int_x^t f'_x\left(x\right)du\right)\mathfrak{q}_{n,\rho}^{\alpha,\lambda}\left(x;t\right)dt\right|.
$$

Integrating by parts after applying Lemma 3.1, and taking $y=x-\frac{x}{\sqrt{n}}$, we obtain

$$
P_{n_1}\left(x\right)\leq\int_0^{x-\frac{x}{\sqrt{n}}}\vartheta_n(x;t)\left|f'_x\left(t\right)\right|dt+\int_{x-\frac{x}{\sqrt{n}}}^x\vartheta_n(x;t)\left|f'_x\left(t\right)\right|dt.
$$

Since $f'_x(x)=0$ and $\vartheta_n(x;t)\leq 1$, it implies

$$
\begin{aligned}
\int_{x-\frac{x}{\sqrt{n}}}^x\vartheta_n(x;t)\left|f'_x\left(t\right)\right|dt&=\int_{x-\frac{x}{\sqrt{n}}}^x\vartheta_n(x;t)\left|f'_x\left(t\right)-f'_x\left(x\right)\right|dt\\
&\leq\int_{x-\frac{x}{\sqrt{n}}}^x\bigvee_t^x\left(f'_x\right)dt\leq\frac{x}{\sqrt{n}}\bigvee_{x-\frac{x}{\sqrt{n}}}^x\left(f'_x\right).
\end{aligned}
$$

Again using Lemma 3.1 and substituting $y=x-\frac{x}{u}$ we obtain

$$
\int_0^{x-\frac{x}{\sqrt{n}}}\vartheta_n(x;t)\left|f'_x\left(t\right)\right|dt\leq\frac{K\delta_n^2\left(x\right)}{n}\int_0^{x-\frac{x}{\sqrt{n}}}\frac{\left|f'_x\left(t\right)\right|}{\left(x-t\right)^2}dt\leq\frac{K\delta_n^2\left(x\right)}{nx}\int_1^{\sqrt{n}}\bigvee_{x-\frac{x}{u}}^x\left(f'_x\right)du
$$

$$\leq \frac{K\delta_n^2(x)}{nx}\sum_{k=1}^{[\sqrt{n}]}\bigvee_{x-\frac{x}{k}}^{x}(f'_x).$$

Thus, we can write $P_{n_1}(x)$ as

$$P_{n_1}(x) \leq \frac{x}{\sqrt{n}}\bigvee_{x-\frac{x}{\sqrt{n}}}^{x}(f'_x) + \frac{K\delta_n^2(x)}{nx}\sum_{k=1}^{[\sqrt{n}]}\bigvee_{x-\frac{x}{k}}^{x}(f'_x). \tag{3.10}$$

Next, to estimate $P_{n_2}(x)$, we have

$$\begin{aligned} P_{n_2}(x) &\leq \left|\int_x^{2x}\left(\int_x^t f'_x(u)\,du\right)\mathfrak{q}_{n,\rho}^{\alpha,\lambda}(x;t)\,dt\right| + \left|\int_{2x}^{\infty}\left(\int_x^t f'_x(u)\,du\right)\mathfrak{q}_{n,\rho}^{\alpha,\lambda}(x;t)\,dt\right| \\ &\leq A_n(x) + B_n(x), \end{aligned}$$

where

$$A_n(x) = \left|\int_x^{2x}\left(\int_x^t f'_x(u)\,du\right)\mathfrak{q}_{n,\rho}^{\alpha,\lambda}(x;t)\,dt\right|$$

and

$$B_n(x) = \left|\int_{2x}^{\infty}\left(\int_x^t f'_x(u)\,du\right)\mathfrak{q}_{n,\rho}^{\alpha,\lambda}(x;t)\,dt\right|.$$

Since $1-\vartheta_n(x,t)\leq 1$, by putting $t = x+\frac{x}{u}$ successively, we have

$$\begin{aligned} A_n(x) =& \left|\int_x^{2x} f'_x(u)\,(1-\vartheta_n(x,2x))\,du - \int_x^{2x} f'_x(t)\,(1-\vartheta_n(x,t)\,dt\right| \\ \leq& \frac{K\delta_n^2(x)}{nx^2}\,|f(2x)-f(x)-xf'(x)| + \int_x^{x+\frac{x}{\sqrt{n}}}|f'_x(t)|\,|1-\vartheta_n(x,t)|dt \\ &+ \int_{x+\frac{x}{\sqrt{n}}}^{2x}|f'_x(t)|\,|1-\vartheta_n(x,t)|dt \\ \leq& \frac{K\delta_n^2(x)}{nx^2}\,|f(2x)-f(x)-xf'(x)| + \frac{K\delta_n^2(x)}{n}\int_{x+\frac{x}{\sqrt{n}}}^{2x}\frac{V_x^t(f'_x)}{(t-x)^2}dt + \int_x^{x+\frac{x}{\sqrt{n}}}\bigvee_x^t(f'_x)dt \\ \leq& \frac{K\delta_n^2(x)}{nx^2}\,|f(2x)-f(x)-xf'(x)| + \frac{K\delta_n^2(x)}{n}\sum_{k=1}^{[\sqrt{n}]}\bigvee_x^{x+\frac{x}{\sqrt{k}}}(f'_x) + \frac{x}{\sqrt{n}}\bigvee_x^{x+\frac{x}{\sqrt{n}}}(f'_x). \end{aligned}$$

Further we estimate the value of $B_n(x)$ as follows:

$$B_n(x) = \left|\int_{2x}^{\infty}\left(\int_x^t f'_x(u)du\right)\mathfrak{q}_{n,\rho}^{\alpha,\lambda}(x;t)dt\right| \tag{3.11}$$

$$\leq \text{Ç} \int_{2x}^{\infty} t^{\gamma} \mathfrak{q}_{n,\rho}^{\alpha,\lambda}(x;t)\,dt + |f(x)| \int_{2x}^{\infty} \mathfrak{q}_{n,\rho}^{\alpha,\lambda}(x;t)\,dt + \sqrt{\frac{K\delta_n^2(x)}{n}} f'(x+). \tag{3.12}$$

It is obvious that $t \leq 2(t-x)$ and $x \leq t-x$, when $t \geq 2x$. Now applying Hölder's inequality in the first term of equation (3.11), we get

$$B_n(x) = 2^{\gamma} \text{Ç} \left(\int_0^{\infty} (t-x)^{2\tau} \mathfrak{q}_{n,\rho}^{\alpha,\lambda}(x;t)\,dt \right)^{\frac{\gamma}{2\tau}} + \frac{K\delta_n^2(x)}{nx^2} |f(x)| + \sqrt{\frac{K\delta_n^2(x)}{n}} f'(x+)$$

$$= \wp(\gamma,\tau,x) + \frac{K\delta_n^2(x)}{nx^2} |f(x)| + \sqrt{\frac{K\delta_n^2(x)}{n}} f'(x+). \tag{3.13}$$

Finally, combining equations (3.10)–(3.13) and putting values of $P_{n_1}(x)$ and $P_{n_2}(x)$ in (3.9), we get the required result and the theorem is proved. □

Example 3.1. Let $f(x) = x^4 - 3x^3 + 2x^2 + 1$. We choose parameters $\alpha = \lambda = 2$ and $\rho = 3$. For $n = 10, 50, 100, 200$, we have the following representations.

(a) Figure 1 shows the rate of approximation of the operators $\mathcal{M}_n^{\alpha,\lambda}$ towards the function f. Clearly the proposed operators (1.3) converge to the function f for sufficiently large n.

(b) In Figure 2, the associated absolute error $\Theta_n = |\mathcal{M}_n^{\alpha,\lambda}(f;x) - f(x)|$ is represented graphically for arbitrary values of x in interval $[0,\infty)$. It can be observed that error is monotonically decreasing for increasing n.

(c) An error estimation table is provided in Table 1 which depicts that for higher value of n, the error approaches to zero.

Therefore, it can be concluded that proposed operators (1.3) provide good approximation for n adequately large.

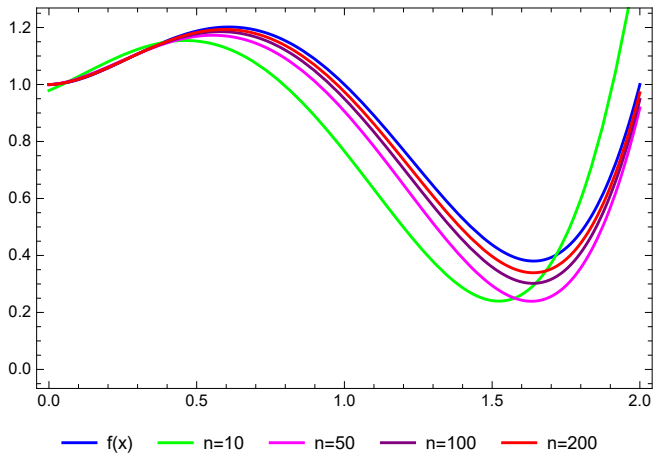


FIGURE 1. Convergence of $\mathcal{M}_n^{\alpha,\lambda}(f;x)$ for the polynomial function $f(x) = x^4 - 3x^3 + 2x^2 + 1$ with parameters $\alpha = \lambda = 2$, $\rho = 3$.

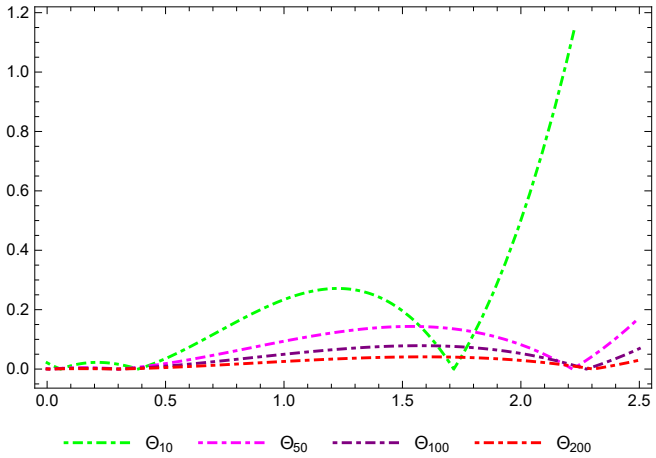


FIGURE 2. Absolute error $\Theta_n = \mathcal{M}_n^{\alpha,\lambda}(f;x) - f(x)|$ of the proposed operators for $f(x) = x^4 - 3x^3 + 2x^2 + 1$ with parameters $\alpha = \lambda = 2$, $\rho = 3$.

TABLE 1. Table for Absolute error $\Theta_n = |\mathcal{M}_n^{\alpha,\lambda}(f;x) - f(x)|$ of the proposed operators $\mathcal{M}_n^{\alpha,\lambda}$.

x	Θ_{10}	Θ_{50}	Θ_{100}	Θ_{200}
0.4	0.0100759	0.00856603	0.0045899	0.00236582
0.8	0.166213	0.0652568	0.0345534	0.0177532
1.2	0.232451	0.120425	0.0647845	0.0335316
1.6	0.0454756	0.123219	0.0698567	0.0369876
2.0	0.921829	0.022785	0.0243437	0.0154083
2.4	2.65087	0.231729	0.0971809	0.0439198
2.8	5.48687	0.691177	0.320143	0.15371
3.2	9.68409	1.40641	0.66997	0.326674
3.6	15.4968	2.42828	1.17209	0.575527
4.0	23.1792	3.80765	1.85192	0.912982

Example 3.2. Figure 3 illustrates the effect of increase in values of parameter ρ on the rate of convergence of proposed operators $\mathcal{M}_n^{\alpha,\lambda}$ for the function $f(x) = 4x(x - 1.1)(x - 1.9)$ while keeping the value of α, λ and n fixed. Here we chose $n = 10$ and $\alpha = \lambda = 2$ to show the impact of the parameter ρ clearly. One can easily deduce from the figure that as we increase the value of ρ the rate of convergence gets relatively faster.

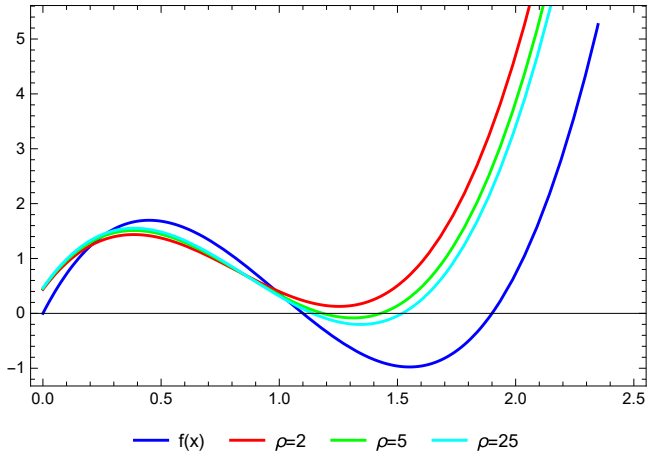


FIGURE 3. Effect of increase in parametric value of ρ for given $n = 10$, $\alpha = \lambda = 2$ on the convergence rate of proposed operators.

Acknowledgements. The authors are grateful to the referee for several valuable comments and suggestions leading to an overall improvement of this paper.

REFERENCES

[1] T. Acar, A. Aral and I. Raşa, *The new forms of Voronovskayas theorem in weighted spaces*, Positivity **20**(1) (2016), 25–40. https://doi.org/10.1007/s11117-015-0338-4

[2] A. M. Acu, T. Acar and V. A. Radu, *Approximation by modified U_n^ρ operators*, Rev. R. Acad. Cienc. Exactas Fís. Nat. (Esp.). Serie A. Matemáticas **113**(3) (2019), 2715–2729. https://doi.org/10.1007/s13398-019-00655-y

[3] A. M. Acu and V. A. Radu, *Approximation by certain operators linking the α-Bernstein and the Genuine α-Bernstein-Durrmeyer operators*, in: N.Deo et al. (Eds.), *Mathematical Analysis I: Approximation Theory*, Springer, Proceedings in Mathematics & Statistics **306**, 2020, 77–88.

[4] I. N. Cangul, H. Ozden and Y. Simsek, *A new approach to q-Genocchi numbers and their interpolation functions*, Nonlinear Analysis: Theory, Methods & Applications **71**(12) (2009), e793–e799. https://doi.org/10.1016/j.na.2008.11.040

[5] N. Deo and S. Kumar, *Durrmeyer variant of Apostol-Genocchi-Baskakov operators*, Quaest. Math. **44**(4) (2020), 1–18. https://doi.org/10.2989/16073606.2020.1834000

[6] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*, Springer, Berlin, 1993.

[7] M. Dhamija, *Durrmeyer modification of Lupas type Baskakov operators based on IPED*, in: N. Deo et al. (Eds.), *Mathematical Analysis I: Approximation Theory*, Springer, Proceedings in Mathematics & Statistics **306**, 2020, 111–120.

[8] M. Dhamija and N. Deo, *Jain-Durrmeyer operators associated with the inverse Pólya–Eggenberger distribution*, Appl. Math. Comput. **286** (2016), 15–22. https://doi.org/10.1016/j.amc.2016.03.015

[9] A. D. Gadjiev, *Theorems of the type of PP Korovkins theorems*, Mat. Zametki. **20**(5) (1976), 781–786.

[10] T. Garg, A. M. Acu and P. N. Agrawal, *Further results concerning some general Durrmeyer type operators*, Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Mat. RACSAM **113**(3) (2019), 2373–2390. https://doi.org/10.1007/s13398-019-00628-1

[11] V. Gupta, *A note on the general family of operators preserving linear functions*, Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Mat. RACSAM **113**(4) (2019), 3717–3725. https://doi.org/10.1007/s13398-019-00727-z

[12] M. Heilmann and I. Raşa, *A nice representation for a link between Baskakov and Szász-Mirakjan-Durrmeyer operators and their Kantorovich variants*, Results Math. **74**(1) (2019), 1–12. https://doi.org/10.1007/s00025-018-0932-4

[13] A. Kajla, N. Ispir, P. N. Agrawal and M. Goyal, *q-Bernstein-Schurer-Durrmeyer type operators for functions of one and two variables*, Appl. Math. Comput. **275** (2016), 372–385. https://doi.org/10.1016/j.amc.2015.11.048

[14] P. P. Korovkin, *Linear Operators and Approximation Theory*, Hindustan Publication Co., Delhi 1960.

[15] Q. M. Luo, *q-Extensions for the Apostol-Genocchi polynomials*, General Mathematics **17**(2) (2009), 113–125.

[16] Q. M. Luo, *Extensions of the Genocchi polynomials and their Fourier expansions and integral representations*, Osaka J. Math. **48** (2011), 291–309.

[17] Q. M. Luo and H. M. Srivastava, *Some generalizations of the Apostol-Genocchi polynomials and the Stirling numbers of the second kind*, Appl. Math. Comput. **217**(12) (2011), 5702–5728. https://doi.org/10.1016/j.amc.2010.12.048

[18] M. A. Özarslan, *Unified Apostol-Bernoulli, Euler and Genocchi polynomials*, Comput. Math. Appl. **62**(6) (2011), 2452–2462. https://doi.org/10.1016/j.camwa.2011.07.031

[19] R. Păltănea, *Modified Szász-Mirakjan operators of integral form*, Carpathian J. Math. **24**(3) (2008), 378–385.

[20] C. Prakash, D. K. Verma and N. Deo, *Approximation by a new sequence of operators involving Apostol-Genocchi polynomials*, Math. Slovaca (to appear).

[21] H. M. Srivastava, B. Kurt and V. Kurt, *Identities and relations involving the modified degenerate hermite-based Apostol-Bernoulli and Apostol-Euler polynomials*, Rev. R. Acad. Cienc. Exactas Fís. Nat. (Esp.). Serie A. Matemáticas **113**(2) (2019), 1299–1313. https://doi.org/10.1007/s13398-018-0549-1

[1]Delhi Technological University
Department of Applied Mathematics
Bawana Road, Delhi-110042, India
Email address: navshaktimishra_phd2k18@dtu.ac.in, navshakti20@gmail.com
Email address: naokantdeo@dce.ac.in

Cite this article
Patel K and Singh SK
Assessing the sustainability of municipal solid waste management using life-cycle analysis.
Proceedings of the Institution of Civil Engineers – Waste and Resource Management,
https://doi.org/10.1680/jwarm.21.00018

Research Article
Paper 2100018
Received 25/05/2021;
Accepted 05/10/2021

ICE Publishing: All rights reserved

Waste and Resource Management

ice Publishing

Assessing the sustainability of municipal solid waste management using life-cycle analysis

K. Patel BTech
MTech student, Department of Environmental Engineering, Delhi Technological University, Delhi, India (corresponding author: kulvendrapatel_2k19ene05@dtu.ac.in)

S. K. Singh BE, MTech, PhD
Professor & Head, Department of Environmental Engineering, Delhi Technological University, Delhi, India (Orcid:0000-0001-8215-8562)

Various municipal solid waste management (MSWM) technologies were evaluated using life-cycle analysis (LCA). The study aimed to compare different MSWM approaches: landfill with composting, anaerobic digestion (AD) combined with landfill, incineration and AD combined with incineration. The Impact 2002 + approach and Simapro 9.1.1.1 software using the Ecoinvent v3.6 database were used to compare environmental impacts. The study showed that landfill with composting had the highest environmental impact due to its high potential of global warming and depletion in ecosystem quality, while other scenarios showed reduced impacts. AD combined with incineration was found to have least environmental impact, mainly due to energy recovery from AD. The goal of the study was to determine more environmentally sustainable alternatives for MSWM using LCA within the system boundary.

Keywords: landfills/sustainability/waste management & disposal

1. Introduction

Rapid industrialisation, urbanisation and exponential population growth has resulted in a significant surge in solid waste generation, which has now become one of the significant environmental challenges. Major concerns related to solid waste management are not only about public health, sustainable management of waste is also a significant concern. Open dumping and unsanitary landfilling of municipal solid waste (MSW) pose a threat to public health and the environment. This problem has grown to alarming proportions and is increasing by the day (Khandelwal *et al.*, 2019). An increasing volume of waste is an environmental problem, but it can satisfy several requirements if considered as a resource (Singh and Gour, 2016). One of the most important techniques for resolving waste management issues is to use integrated waste management systems (IWMS). IWMS provide various recovery, recycling and waste disposal solutions based on waste fractions. Although the definition of municipal solid waste management (MSWM) varies among countries, it generally refers to collecting, transportation, processing, treatment and disposal of solid waste, which includes household waste, yard waste and institutional waste. India's Municipal Solid Wastes (Management and Handling) Rules 2000 define MSW as 'commercial and residential wastes generated in municipal or notified areas in either solid or semi-solid form excluding industrial hazardous wastes but including treated bio-medical wastes' (IL&FS, 2010: p. 3). According to the *Swachhata Sandesh* newsletter produced by the Ministry of Housing and Urban Affairs (MoHUA) of the Government of India (GoI), 147 613 MT of municipal solid waste (MSW) is generated per day up to December 2019 from 84 475 wards (SBM, 2020). In its 2014 report, the Planning Commission predicted that urban India would generate 276 342 tonnes by 2021, 450 132 by 2031, and 1 195 000 by 2050. The output of waste per capita is 450 g/day and has grown at a rate of 1.3% per year (Planning Commission, 2014). Because municipal operation creates the world's most enormous waste volume (Leme *et al.*, 2014), this increasing challenge requires numerous standards and regulations. Examples include: in the fourth five-year plan (1969–1974), the GoI began efforts to improve MSWM facilities by offering grants and loans to state governments to build MSW composting facilities; the Municipal Solid Wastes (Management and Handling) Rules 2000 provide standards and procedures for the collection, segregation, transportation, processing and disposal of MSW for municipal authorities (Talyan *et al.*, 2008). Other policy initiatives informing and guiding MSWM services include the recommendations of the Technical Advisory Group on MSWM (2005) and the Inter-Ministerial Task Force on Integrated Plant Nutrient Management (2005); the Hazardous Waste (Management, Handling & Transboundary Movement) Rules (1989, 2008); the Bio-Medical Waste Management Rules (2016); and the Plastic Waste Rules (2016) (CPHEEO, 2016). Uncontrolled disposal has a detrimental effect on environmental and public health components (Rathi, 2006; Sharholy *et al.*, 2005). Emissions into the air, water and soil can be generated at any MSWM stage from waste collection to final disposal. These emissions have significant consequences on human health and the environment, and, as a result, there is a delay in the approval of any proposed waste treatment and disposal facility, and, therefore, environmental impact assessments (EIAs) are required to reduce the impact on the ecosystem and human health.

To achieve the goals of sustainable waste management (SWM), it is necessary to incorporate various waste management

Downloaded by [Delhi Technological University] on [09/12/21]. Copyright © ICE Publishing, all rights reserved.

technologies strategically. It is also essential to understand different technologies using a comparative analysis of various options, and it can be a method to guide the decision making process (Zaman, 2009). The life-cycle analysis (LCA) approach is reported to be promising to resolve issues (Rebitzer *et al.*, 2004). Acid potentials, exploitation of resources, ecotoxicity (aquatic, terrestrial), greenhouse gas (GHG) emissions and many other impacts from different MSWM scenarios can be correlated and compared with LCA (Laurent *et al.*, 2014). LCA is a method applied to assess environmental impacts associated with a product or a process or a service. It compiles all related products, energy use and emissions released over the entire process. The use of LCA in MSWM serves the following purposes: (*a*) analysing the end phase of a process and the impacts on the environment and (*b*) evaluating waste management services of private and government entities. Numerous analyses have been conducted to examine the use of LCA in the MSWM field. For example, in Delhi, Srivastava and Nema (2011) analysed the impacts of different MSWM systems on the environment. The treatment scheme included a materials recycling facility (MRF), landfilling, composting and incineration. The study found that the environmental effects of MRF are minimal. Another LCA research was conducted in city of Delhi by Bohra *et al.* (2012), which compared 12 separate MSW scenarios. The percentage of waste treated by incineration, anaerobic digestion (AD), composting, refuse-derived fuel and sanitary landfill was changed to create these scenarios. According to the findings, the scenario with the most diversion from sanitary landfill had the lowest global warming potential. Babu *et al.* (2014) conducted an LCA analysis on Bangalore city's MSW, comparing four scenarios: open dumps (S1), landfill without gas recovery (S2), landfill with gas recovery (S3) and bioreactor landfill (S4), and concluded that the S4 alternative was the best among them. Besides these, several LCA applications in MSWM have been focused on its use as a decision support tool in selecting the best MSW management strategy. It has been applied in several countries, including Spain (Bovea and Powell, 2006; Bueno *et al.*, 2015; Fernández-Nava *et al.*, 2014), Switzerland (Boesch *et al.*, 2014; Rossi *et al.*, 2015) and Denmark (Andersen *et al.*, 2010; Manfredi and Christensen, 2009). It will be challenging to choose the right treatment and recycling methods for solid waste from an environmental viewpoint. Consistent analysis methods of advantages and disadvantages related to the wide range of treatment or disposal methods available are needed to establish an improved waste management system (Song *et al.*, 2013).

2. Methodology

2.1 Study area and MSW composition

Due to huge investments to improve the city's infrastructure, Delhi is one of India's fastest-growing cities. The city's population density had grown from 9340 persons/km^2 in 2001 to 11320 persons/km^2 in 2011 (MHA, 2011). Five municipal authorities manage the city's solid waste collection. The three corporations – North, East and South – account for 96% of the region's total area. The city collects 10 050 t/day of MSW, as per the Municipal Corporation of Delhi. The composition of MSW produced is 50–55% organic, 11% paper, 13.5% plastics and packaging, 6% metal and glass and the rest inert materials (DSLSA, 2017). According to the composition, the major portion of the waste produced is organic, making biological treatment techniques such as composting and AD the most suitable.

2.2 Life-cycle assessment

The International Organization for Standardization (ISO) standards 14040 and 14044 characterise the LCA methodology: the first describes the principles and structure (ISO, 2006a), while the second outlines the requirements and criteria (ISO, 2006b) for conducting an LCA.

LCA comprises four-stage models: defining the goal and scope of the study, a life-cycle inventory (LCI) based on energy and mass quantification, life-cycle impact assessment (LCIA) aimed at determining the importance of a system's possible environmental impacts, and an assessment of the findings to help conclude (Rana *et al.*, 2019). It is becoming more common in the decision making and policy preparation process. LCA has been integral to every monitoring/protection policy for the environment (Bras-Klapwijk, 1999; Seidel, 2016). Public policy-making is listed by ISO as one of the potential direct applications of LCA (ISO 2006). ISO states that LCA can assist in decision-making in industry, governmental or non-governmental organisations (ISO 2006), which would presume to include public policy development. Examples of application of LCA in policy development includes the California Environmental Protection Agency's Department of Toxic Substances Control's evaluation of alternative hazardous waste management systems. LCA, suggested by Frankl and Rubik (2000), providing knowledge about renewable waste transportation and recycling pathways, waste-to-energy proposals and potential for product development from waste may be instrumental to SWM. The assessment of treatment methods and various techniques for particular waste fractions for optimised SWM is currently being used in many countries. LCA examines the potential environmental impacts of the waste life cycle (from generation to disposal).

2.3 Goal and scope of the study

The scope of this study is limited to the use of secondary data to compile the LCI for various SWM options. Using the LCA method, the study analyses the effects of various potential waste management options for Delhi. It analyses the treatment and disposal methods concerning the environmental issues related to waste management; human health, climate change, ecosystem quality (EQ) and resource depletion, to propose a suitable method with the least environmental impacts.

Downloaded by [Delhi Technological University] on [09/12/21]. Copyright © ICE Publishing, all rights reserved.

The scenarios include the existing scenario of Delhi (S1) and three proposed scenarios for Delhi (S2, S3 and S4). Figure 1 shows the boundary system of the study. Within the system boundary, all mass and energy inputs were included; all outputs, such as emission to the atmosphere, electricity and compost generation from MSWM operations, and avoided products (high-density polyethylene, low-density polyethylene, packaging glass, polyethylene terephthalate (PET)) were considered. The waste collection scheme for all the scenarios within the study area was considered the same.

2.4 Functional unit

The functional unit was defined as per tonne of MSW treated. It is assumed that for all waste management options, similar waste characteristics were considered. It was also assumed that the electricity generated from the MSWM method would replace the electricity produced from a fossil fuel thermal power plant. The total distance travelled by the collection trucks and the tonnes transported for S1 and S2 are assumed as 7.4 tkm (5 tkm from the collection centre to the waste treatment facility and 2.4 tkm from the treatment facility to the landfill), and for S3 and S4, it is assumed as 5 tkm. The unit process for the transport was selected from Simapro 9.1.1.1 database 'Municipal waste collection service by 21 metric ton lorry {GLO}|market for| APOS, S'.

2.5 Scenarios

2.5.1 Scenario S1 (baseline scenario)

This scenario emphasises the recovery of compost from the biodegradable fraction of solid waste. Currently, the solid wastes are transported to the centre after sorting the recyclable products (PET, packaging glass and other plastics, contributing 3% per ton of waste) (Rajaeifar *et al.*, 2015; Talyan *et al.*, 2008). From the total MSW generated, 17% is then sent for compost production, and the wastes which are rejected (contributes 80% per ton of waste) are sent to the landfill where no treatment facility is available for biogas and leachates. Transportation, sorting, landfilling and composting are the main phases, as shown in Figure 2.

2.5.2 Scenario S2 (future scenario)

This scenario emphasises the production of biogas and bio-fertilisers from the biodegradable fraction through an AD process. Initially, MSW is transported to the facility for sorting, where PET, glass and other plastics, are separated (contributes 3% per ton of waste). The organic MSW is introduced to the AD plant (contributes 48% per ton of waste) (Rajaeifar *et al.*, 2015; Talyan *et al.*, 2008), and the waste that is rejected goes to the landfill site (contributes 49% per ton of waste). The produced biogas in the AD tank is used for heat and electricity generation. This scenario includes waste transportation to the treatment facility, sorting, landfill and AD.

2.5.3 Scenario S3 (future scenario)

In this proposed scenario, the MSW is subjected to an incinerator after separating the recyclable materials. The organic and rejected fractions of MSW, which contribute 97% per ton of waste, are incinerated (Bohra *et al.*, 2012). Transportation of waste to the treatment facility, sorting and finally waste incineration are the main phases of the scenario, as shown in Figure 2.

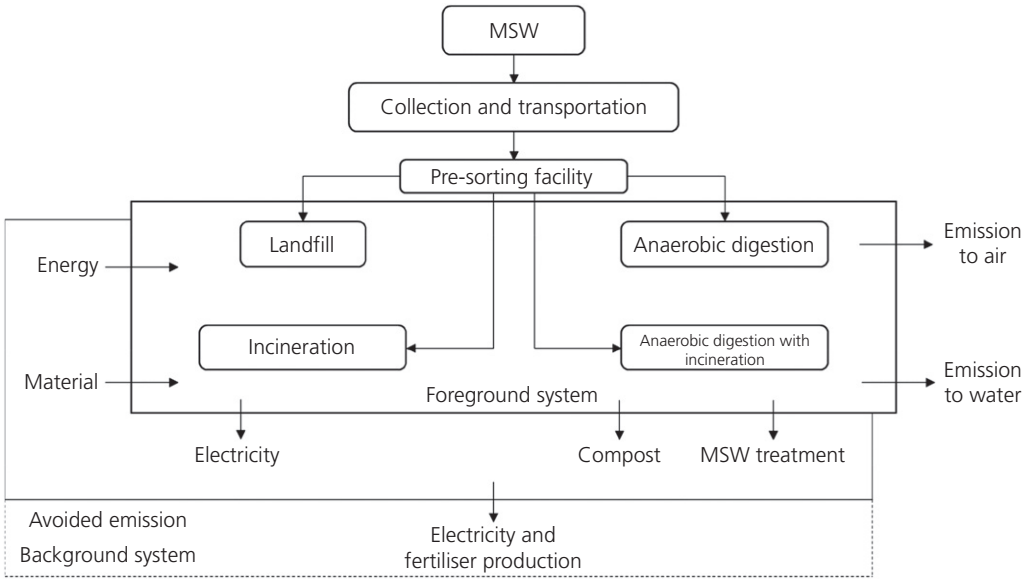


Figure 1. System boundary of MSWM system

Downloaded by [Delhi Technological University] on [09/12/21]. Copyright © ICE Publishing, all rights reserved.

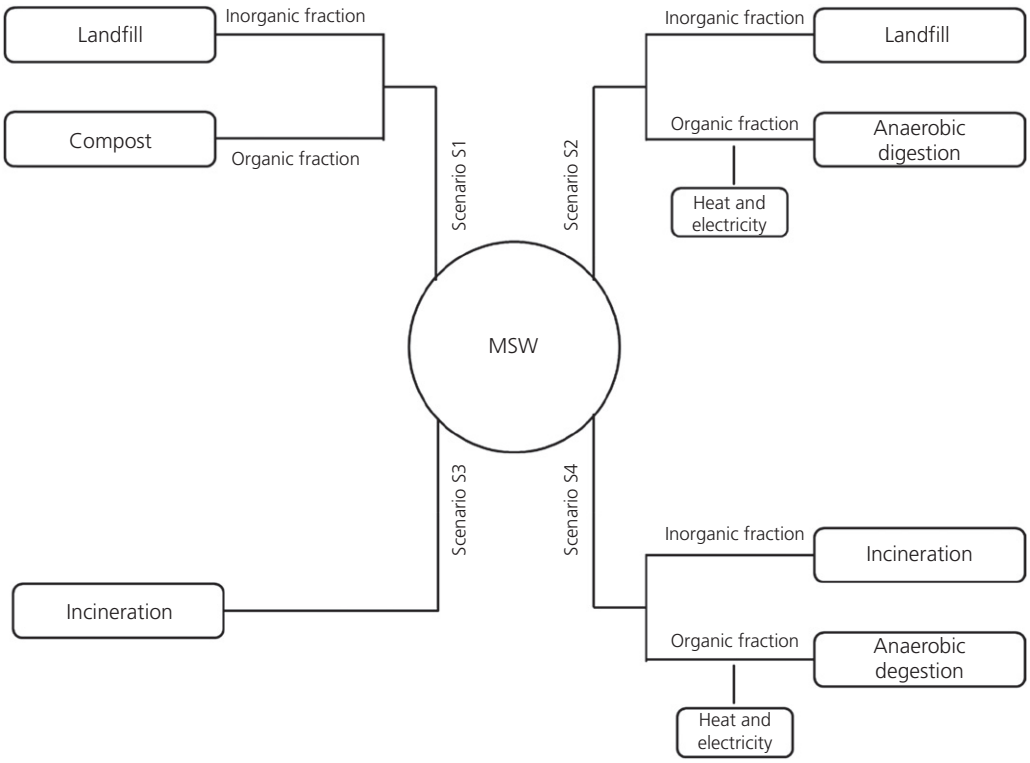


Figure 2. Flow of MSW fractions in considered scenarios

2.5.4 Scenario S4 (future scenario)

In this proposed scenario, MSW is subjected to AD combined with incineration. After recyclables were separated, contributing 3% per ton of waste, the organic fraction is subjected to AD process (contributes 48% per ton of waste), and the heat and electricity are produced. The waste rejected (contributes 49% per ton of waste) is then subjected to an incinerator. This scenario includes waste transportation, sorting, AD and incineration (Table 1).

Table 1. Description of scenarios used in the study

Scenario	Description
1 – Landfill with composting	3% recyclables + 17% to compost + 80% to landfill
2 – AD with landfill	3% recyclables + 48% to AD + 49% to landfill
3 – Incineration	3% recyclables + 97% to incineration
4 – AD with incineration	3% recyclables + 48% to AD + 49% to incineration

2.6 Life-cycle inventory

This is the LCA step, which includes collecting inputs and outputs from the system for the entire life cycle of a process. Raw material and energy are the system inputs, and materials such as compost, emissions to air, water and soil are the system outputs. For each phase of the system boundary, data for input/output, including the raw material, products and emissions, are obtained and recorded. The inventory included in this analysis was taken from the Municipal Corporation Office, Simapro 9.1.1.1 database and previously published literature (Bohra *et al.*, 2012; Rajaeifar *et al.*, 2015). The system boundary (Figure 1) was divided into two systems – namely, the foreground system consisting of emissions from treatment processes and the background system consisting of energy and fertiliser production. The inventory data of the various scenarios including the emissions to air and water is shown in Table 2.

2.6.1 Composting

Composting is an organic matter decomposition process in a regulated state for the production of a final product called humus or compost. The method of composting is not a solution to all waste management problems; however, it can serve as a valuable component in waste management. As a soil amendment, the compost thus produced may be used for land application. In Delhi, two centralised composting plants, at Okhla and Bawana, together process about 1200 t/day of

Downloaded by [Delhi Technological University] on [09/12/21]. Copyright © ICE Publishing, all rights reserved.

Table 2. Inventory data (per ton of MSW)

	Scenarios			
Parameters	**S1**	**S2**	**S3**	**S4**
Inventory for background data				
Input from technosphere				
Electricity: kWh	33.5	10.9	74.2	46.9
Diesel: kg	2.94	2.69	3.31	3
Truck: tkm	7.4	7.4	5	5
Sodium hydroxide: kg	—	1	—	1
Lime: kg	—	—	14	8
Avoided products				
Fertiliser as nitrogen: kg	3.24	0.35	—	0.35
Fertiliser as phosphorus pentoxide: kg	1.73	0.18	—	0.18
High-density polyethylene: kg	3.5	3.5	3.5	3.5
Low-density polyethylene: kg	4.5	4.5	4.5	4.5
Packaging glass: kg	3.5	3.5	3.5	3.5
PET (bottle grade) : kg	5	5	5	5
Electricity: kWh	—	201.6	331.3	520.1
Heat: MJ	—	4.12	—	4.12
Inventory for foreground data				
Emission to air				
Carbon dioxide: g	9200	8400	10300	9070
Sulfur dioxide: g	23	44.8	169	205
Methane: g	114 930.34	0.35	0.42	0.373
Methane, biogenic: g	—	716.28	—	716.3
Dinitrogen monoxide: g	0.35	0.31	0.4	0.35
Ammonia: g	8000	3.95	0.066	0.058
Non-methane volatile organic compound: g	8.38	38.7	9.43	8.27
Nitrogen oxides: g	171	1320	147	1330
Nitrogen dioxide: g	—	—	665	639
Carbon monoxide: g	18.5	141	187	178
Formic acid: g	5			
Acetic acid: g	5			
Carbon monoxide, biogenic: g	—	598.75	—	599
Particulates > 2.5 and <10 μm: g	—	0.45	—	0.45
Particulates > 10 μm: g	—	1	—	1
Particulates < 2.5 μm: g	13.2	12.4	14.9	13.3
Volatile organic compound: g	11.7	—	—	—
Hydrogen sulfide: g	646	—	—	—
Hydrocarbons: g	—	563.3	—	563.3
Formaldehyde: g	—	47	—	47
Total organic carbon: g	—	—	33.2	32
Wood (dust) : g	—	—	33.2	32
Lead: g	—	—	1.6	1.6
Hydrogen chloride: g	—	—	33.2	31.9
Hydrogen fluoride: g	—	—	3.32	3.2
Emission to water: g				
Chemical oxygen demand: g	1652.33	12.34	—	—
Mercury: mg	0.32	0.12	—	—
Magnesium: g	11.88	—	—	—
Calcium: g	93	—	—	—
Sodium: g	91.20	—	—	—
Biological oxygen demand: kg	770	—	—	—
Potassium: g	55.2	—	—	—
Iron: g	20.016	—	—	—

biodegradable waste (DSLSA, 2017). The existing facility uses aerobic compost plant with an open windrow system (Mandal *et al*., 2014). The composting plants are not operating at maximum capacity due to a lack of adequate waste component segregation. The compost generated by the composting plant is sold in nearby markets, reducing the need for fertilisers.

Downloaded by [Delhi Technological University] on [09/12/21]. Copyright © ICE Publishing, all rights reserved.

2.6.2 Landfill

One of the major methods adopted worldwide to manage MSW is landfill. Although it is considered a cost-effective MSW disposal method, improper management causes major environmental problems, especially in developing nations like India. Certain gases are produced due to the decomposition of wastes in landfills, including methane, carbon dioxide (CO_2), nitrogen and carbon monoxide. Three sites in Delhi, namely, Gazipur, Bhalswa and Okhla, are currently operational. Around 8000 t/day of waste is collected and sent to these sites. The MSW is placed in 2–5 m layers and is covered at the end of the day. Currently, with minimal compaction and no daily cover, the waste is discarded in an unorganised manner, without any treatment facility for landfill gas (LFG) (gas produced in landfills under anaerobic conditions) and leachate.

2.6.3 Anaerobic digestion

It is a process in which the organic matter is disintegrated into biogas and bio-fertilisers without oxygen. Due to the high biodegradable portion in Delhi's MSW, AD is one of the preferred waste treatment options. It produces energy from the methane gas collected and also produces digestate. However, proper sorting and segregation and higher energy inputs are needed for this treatment facility. The anaerobic method generates biogas and digestate. Biogas is made up of methane (55–70%) and carbon dioxide, and is emitted after 2.3 weeks. Based on its nutrient content, digestate can be used as an organic fertiliser (Zaman, 2009).

2.6.4 Incineration

The combustion temperature of the incineration plant exceeds 850°C. Carbon dioxide is, however, released into the atmosphere in this process and non-combustible materials are formed. It is assumed that the electricity generated from the incineration plant is 331.3 kWh/t of waste. The emissions into the air are carbon dioxide, sulfur dioxide, ammonia, lead and hydrogen chloride. Indian MSW is not much recommended in MSW incineration technology due to its high moisture content of around 40–60% and low calorific value, which is approximately 800–1100 kcal/kg. It has been reported that a 3.75 MW power generated from 300 t/day of MSW, installed in Timarpur, Delhi, has not succeeded due to MSW's low net heat value (Pujara *et al.*, 2019).

2.7 LCIA method

The objective of LCIA is to consider and assess a product or a process or service's environmental impacts. LCIA consists of several steps; selecting the relevant impact categories, assigning the elementary flows, characterisation, normalisation, grouping and weighting (Nieuwlaar, 2013). There are several impact methods for determining the environmental impacts. Of these, the Impact 2002 + method was used in this study. This method consists of four other methods: Impact 2002, Eco-indicator 99, CML (Guinée *et al.*, 2002) and that of the Intergovernmental Panel on Climate Change (IPCC). The considered method includes 15 mid-point level categories which are grouped into four damage impact score or damage categories: human health, EQ, climate change and resources. Mid-points are links in an impact category's cause-and-effect chain prior to the end points, from which indicators may be obtained to represent the relative value of emissions such as eutrophication potential caused by the release of phosphate, ammonia; acidification potential caused by the release of substances such as oxides of sulfur and oxides of nitrogen (SO_x, NO_x); human toxicity potential (includes carcinogenic and non-carcinogenic potential) caused by pollutant emissions such as SO_x, NO_x, heavy metals and particulate matter (PM); global warming potential caused by the release of GHGs such as carbon dioxide, nitrous oxide, methane and chlorofluorocarbon. End-point indicators are estimated to represent changes in stressors at an end point in a cause–effect chain, which can have a significant impact on society's perception of the final effect. Table 3 shows the impact categories of Impact 2002 + method.

Table 3. Environmental impact categories of Impact 2002 + method

Impact category	Damage category	Unit
Carcinogens: kg C_2H_3Cl eq Non-carcinogens: kg C_2H_3Cl eq Respiratory inorganics: kg $PM_{2.5}$ eq Respiratory organics: kg C_2H_4 eq Ionising radiation: Bq C-14 eq Ozone layer depletion: kg CFC-11 eq	Human health	Daly
Aquatic ecotoxicity: kg TEG water Aquatic eutrophication: kg PO_4-P limited Terrestrial ecotoxicity: kg TEG soil Aquatic acidification: kg SO_2 eq Terrestrial acid/nutrient: kg SO_2 eq Land occupation: m^2 organic arable land	Ecosystem quality	PDF m^2 year
Global warming: kg CO_2 eq	Climate change	kg CO_2 eq
Non-renewable energy: MJ primary	Resources	MJ primary

C_2H_3Cl, vinyl chloride; C_2H_4, ethylene; CFC, chlorofluorocarbon; Daly, disability adjusted life years; PDF, potentially disappeared fraction; $PM_{2.5}$, particulates < 2.5 μm; PO_4-P, phosphate; SO_2, sulfur dioxide; TEG, triethylene glycol

3. Results and discussion

In this study a comparative analysis of LCA of various MSWM scenarios was performed. The data obtained from the analysis of damage categories are as follows.

3.1 Human health

The human health is a combination of the mid-point environmental impact categories shown in Table 3. In disability adjusted

Downloaded by [Delhi Technological University] on [09/12/21]. Copyright © ICE Publishing, all rights reserved.

life years (Daly), the human health effect is reflected, taking into account the years lost to early mortality and reflecting the reduced quality of life due to illness over the years. Scenario S1, landfill combined with composting, has the highest impact on human health, 7.28×10.4 Daly/t waste. After analysing the mid-point score making up the damage category, the impact was primarily due to the respiratory inorganics (RI), which is 1.05 kg $PM_{2.5}$ eq/t waste, followed by respiratory organics (RO) 0.609 kg C_2H_4 eq/t waste. These results are similar with those of Bohra *et al.* (2012), who described the worst scenario as landfilling combined with composting. The RI mid-point category was introduced as the key element influencing the human health damage category by Fernández-Nava *et al.* (2014). The landfilling situation was considered to be the worst in the human health damage category in their research study. In this study, the sorting process seemed to have helped minimise the RI effect in all situations due to the recovery of PET, glass and other plastics (Figure 3).

3.2 Ecosystem quality

EQ is a combination of the mid-point impact category as tabulated in Table 3. EQ is an aspect of protection which has an impact on the natural environment. EQ is expressed as potentially disappeared fraction (PDF) of species over a certain area in square metres over a certain year (PDF m^2 year). PDF is measured using global and real species values rather than local values. As shown in Figure 4 all scenarios have negative impacts on EQ due to the avoided products, but when one compares the scenarios, the landfill with composting method (S1) has the least impact on EQ. 18 338.756 PDF m^2 year/t of waste, due to avoided emissions resulting from the recovery of avoided products (fertilisers, PET, packaging glass and other

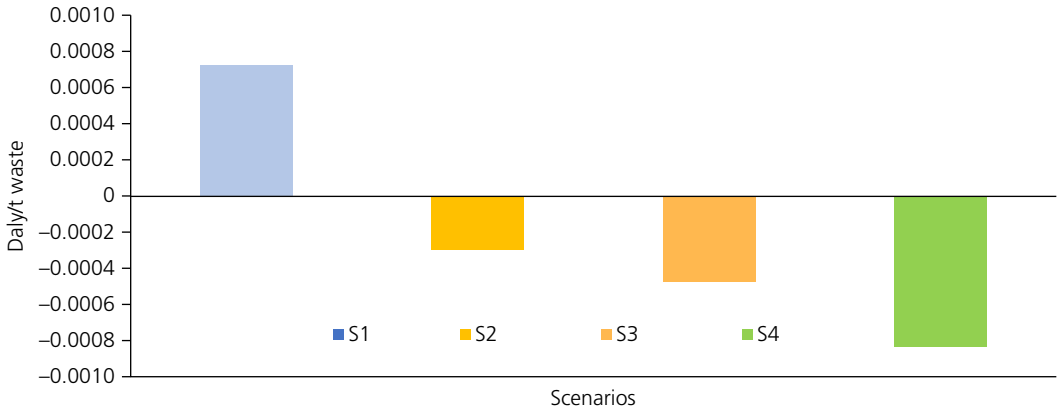


Figure 3. Human health end-point score of all scenarios

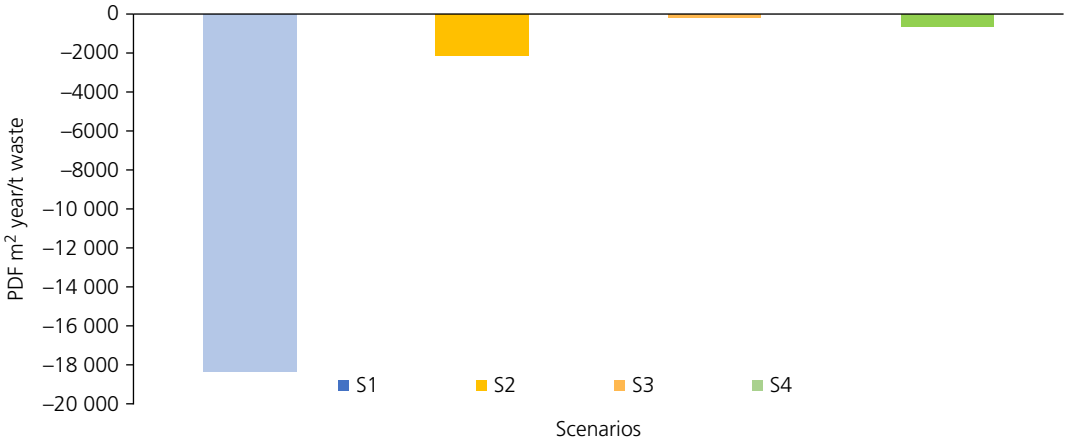


Figure 4. Ecosystem quality end-point score of all scenarios

Downloaded by [Delhi Technological University] on [09/12/21]. Copyright © ICE Publishing, all rights reserved.

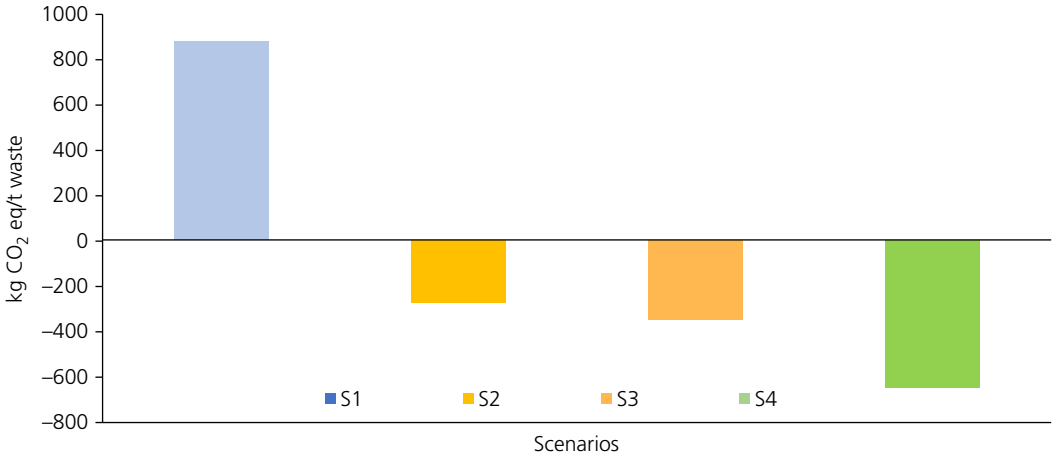


Figure 5. Climate change end-point score of all scenarios

plastics), the sorting process substantially reduces the impacts on this damage category.

3.3 Climate change

The climate change end-point category is similar to the global warming category. If it is considered as a damage category, the effect of climate change is still expressed in kg CO_2eq. As shown in Figure 5, S1 (landfill with composting) has the worst climate change impact, 873.29363 kg CO_2eq/t waste. The AD with incineration (S4) has a negative value of 649.01141 kg CO_2eq/t waste due to the avoided products, production of biogas from the AD plant. It can be concluded that a recovery of methane will significantly reduce the burden exerted by waste disposal on climate change. In addition to the significant effect of methane's recovery on landfill sites, other considerations could also play a role in the climate change damage category, such as variations in MSW characteristics, emissions from infrastructure and landfill sites, electricity or heat recovery and biogas production periods. Similar results were published by Mendes *et al.* (2004) and Hong *et al.* (2010): they found that landfilling had the highest impact on the climate change damage category in Sao Paulo in Brazil, and China, respectively. Khandelwal *et al.* (2019) found similar results for Nagpur, India. They found that the base scenario landfill with composting has the most climate change impact due to high carbon dioxide and methane emissions and no LFG control system.

3.4 Resources

The 'Resources' damage category is the combination of the 'non-renewable energy consumption' and 'mineral extraction' mid-point categories. The resources damage category is expressed in 'MJ'. Figure 6 shows that all the analysed scenarios show negative values for the resources category, indicating that the environment benefited from this damage category in the considered scenarios. S4, however, has the most eco-friendly index with 9127.7094 MJ primary/t of waste among the scenarios analysed, and S1 was the lowest with 487.69359 MJ primary/t of waste. Table 4 shows the damage assessment values of all scenarios. In both scenarios (S2 and S4), AD improved the eco-friendly index; this is due to the generation of biogas from the AD plant to produce electricity. Moreover, S4 was more favourable in the resources damage category than S2 due to the incineration rather than landfilling of rejected MSW. Similar findings were obtained by Rajaeifar *et al.* (2015), who found that in the resources damage category, all of the scenarios analysed had negative values, indicating that all of the scenarios benefited the environment in this category. AD, however, had the most eco-friendly index in this damage category. Fernández-Nava *et al.* (2014) obtained similar results, with the incineration being identified as the main contributor in this category.

3.5 Normalisation

Normalisation allows abstract impact rates to be converted into relative product inputs to the reference situation for each category of impact. The fact is that the results of normalisation are seen in the same unit. It is easier for each impact score to compare the impacts of the various impact categories. Numerical scores allow for the comparison of product options with climate change, ozone degradation, acidification, eutrophication, toxicity, depletion of fossil fuels and other types of environmental effects (Norris, 2001). By dividing the impact per unit of emission by the cumulative impact of all substances of the specific type, per person per year, the damage factor reported in Ecoinvent is normalised (for Europe). Figure 7

Downloaded by [Delhi Technological University] on [09/12/21]. Copyright © ICE Publishing, all rights reserved.

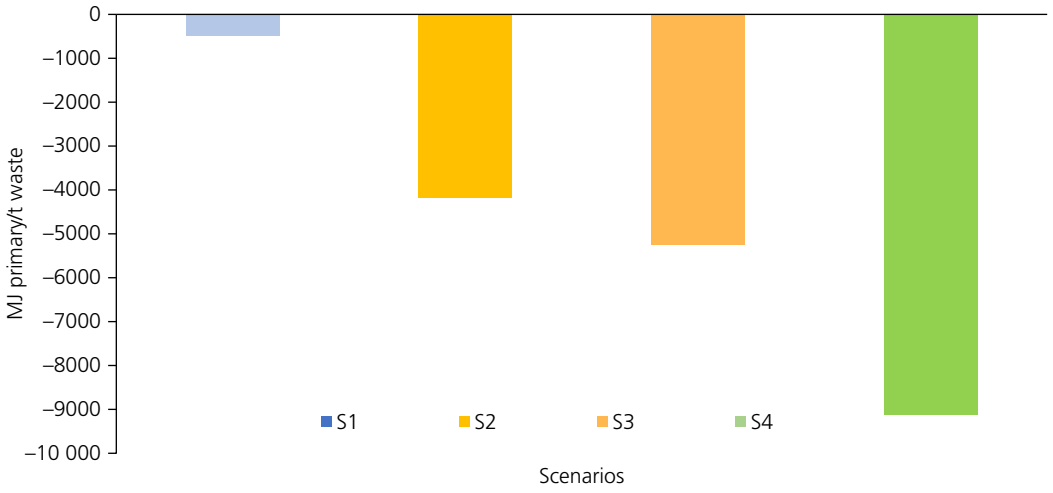


Figure 6. Resources end-point score of all scenarios

Table 4. Damage assessment values of all scenarios – end-point scores

Damage category	S1	S2	S3	S4
Human health: Daly	0.000728331	–0.000297049	–0.000473307	–0.00083031
Ecosystem quality: PDF m^2 year	–18338.756	–2115.5686	–166.32923	–2287.8901
Climate change: kg CO_2 eq	873.29363	–273.17227	–348.32458	–649.01141
Resources: MJ primary	–487.69359	–4170.8092	–5254.0488	–9127.7094

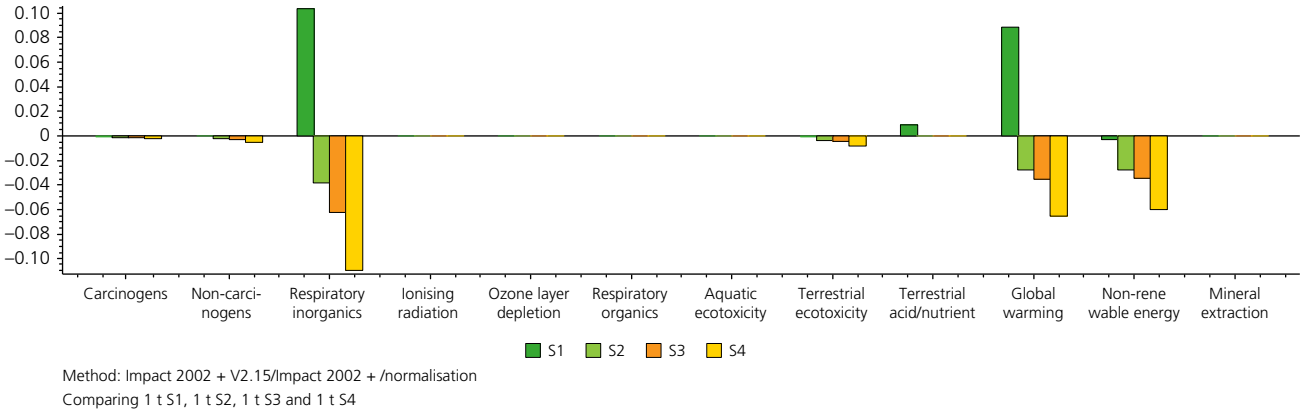


Figure 7. Normalised values of mid-point categories in all the scenarios

shows the normalised values of mid-point categories of considered scenarios (Table 5).

4. Conclusion

This analysis highlights LCA's role in Indian MSWM, providing numerous scenarios for composting, AD, incineration and waste disposal. In this analysis, waste management alternatives were investigated primarily from an environmental perspective; it should be accompanied by other decision making methods that take into account the economic and social aspects of solid waste management. The best choice for MSWM is to segregate waste in combination with energy and biogas. In MSWM, LCA should be sufficient to accomplish solid waste

Downloaded by [Delhi Technological University] on [09/12/21]. Copyright © ICE Publishing, all rights reserved.

Table 5. Normalised scores of all the scenarios

Impact category	S1	S2	S3	S4
Carcinogens	−0.00088	−0.00129	−0.00145	−0.00205
Non-carcinogens	−0.00011	−0.00239	−0.00314	−0.00553
Respiratory inorganics	0.103499	−0.03818	−0.0621	−0.10942
Ionising radiation	-4.79×10^{-6}	-2.05×10^{-5}	-2.57×10^{-5}	-4.71×10^{-5}
Ozone layer depletion	4.80×10^{-8}	-6.33×10^{-9}	-1.04×10^{-7}	-5.00×10^{-7}
Respiratory organics	0.000183	-6.29×10^{-7}	-1.46×10^{-5}	-2.05×10^{-5}
Aquatic ecotoxicity	-1.89×10^{-5}	-9.50×10^{-5}	−0.00012	−0.00022
Terrestrial ecotoxicity	−0.00092	−0.00358	−0.00428	−0.00788
Terrestrial acid/nutrient	0.009104	0.000226	-8.13×10^{-5}	4.12×10^{-5}
Global warming	0.088203	−0.02759	−0.03518	−0.06555
Non-renewable energy	−0.00319	−0.02743	−0.03456	−0.06004
Mineral extraction	-2.02×10^{-5}	-1.30×10^{-5}	-1.33×10^{-5}	-1.95×10^{-5}

management and meet environmental goals to minimise the associated impacts. However, formal and informal incentives and training are highly important to conduct source segregation, which minimises losses of material and electricity. The study shows that while it has the least impact on ecosystem quality, S1, which is the current scenario being implemented, has the highest environmental impact. Based on the damage categories, the findings obtained indicate that the method of AD with incineration has the least environmental impacts. This is primarily due to the production of energy and heat that is assumed will replace the generation of electricity elsewhere. In India, the landfill combined with the composting method has detrimental environmental impacts; therefore, the method adopted should be more eco-friendly. The study recommends that S4 AD with incineration would be more eco-friendly. In this study, generic LCI data from several countries have been used in the absence of databases relating to waste management methods in India. However, adequate inventory databases should be developed and maintained to provide more reliable findings and potential LCA studies. The findings of the LCA analysis can be used to aid decision making in the preparation and optimisation of the IWMS system, which may be useful in developing future waste management strategies.

Declaration of competing interest

The authors declare no conflicts of interest concerning the research, authorship, and publication of this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

REFERENCES

Andersen J, Boldrin A, Samuelsson J, Christensen T and Scheutz C (2010) Quantification of greenhouse gas emissions from windrow composting of garden waste. *Journal of Environmental Quality* **39(2)**: 713–724, https://doi.org/10.2134/jeq2009.0329.

Babu GL, Lakshmikanthan P and Santhosh LG (2014) Life cycle analysis of municipal solid waste (MSW) land disposal options in Bangalore City. *International Conference on Sustainable Infrastructure 2014, Long Beach, CA, USA*, https://doi.org/10.1061/9780784478745.075.

Boesch M, Vadenbo C, Saner D, Huter C and Hellweg S (2014) An LCA model for waste incineration enhanced with new technologies for metal recovery and application to the case of Switzerland. *Waste Management* **34(2)**: 378–389, https://doi.org/10.1016/j.wasman.2013.10.019.

Bohra A, Nema A and Ahluwalia P (2012) Global warming potential of waste management options: case study of Delhi. *International Journal of Environmental Technology and Management* **15(3)**: 346–362, https://doi.org/10.1504/IJETM.2012.049233.

Bovea M and Powell J (2006) Alternative scenarios to meet the demands of sustainable waste management. *Journal of Environmental Management* **79(2)**: 115–132, https://doi.org/10.1016/j.jenvman.2005.06.005.

Bras-Klapwijk RM (1999) *Adjusting Life Cycle Assessment Methodology for Use in Public Policy Discourse*. Doctoral dissertation, Delft University of Technology, Delft, the Netherlands.

Bueno G, Latasa I and Lozano P (2015) Comparative LCA of two approaches with different emphasis on energy or material recovery for a municipal solid waste management system in Gipuzkoa. *Renewable and Sustainable Energy Reviews* **51**: 449–459, https://doi.org/10.1016/j.rser.2015.06.021.

CPHEEO (Central Public Health and Environmental Engineering Organisation) (2016) *Municipal Solid Waste Management Manual Part II: The Manual*. Ministry of Urban Development, New Delhi, India.

DSLSA (Delhi State Legal Services Authority) (2017) *Recommendations for Long Term Action Plan for Solid Waste Management in Delhi*. DLSA, New Delhi, India.

Fernández-Nava Y, Del Río J, Rodríguez-Iglesias J, Castrillón L and Marañón E (2014) Life cycle assessment of different municipal solid waste management options: a case study of Asturias (Spain). *Journal of Cleaner Production* **81**: 178–189, https://doi.org/10.1016/j.jclepro.2014.06.008.

Frankl P and Rubik F (2000) *Life Cycle Assessment in Industry and Business: Adoption Patterns, Applications and Implications*, 1st edn. Springer-Verlag, Berlin, Germany.

Guinée J, Gorrée M, Heijungs R *et al.* (2002) Handbook on life cycle assessment: operational guide to the ISO standards. In *The Second Dutch LCA-Guide* (Guinée, J. (ed.)), 1st edn. Springer,

Downloaded by [Delhi Technological University] on [09/12/21]. Copyright © ICE Publishing, all rights reserved.

Dordrecht, the Netherlands, pp. 311–313, https://doi.org/10.1007/0.306.48055.7.

Hong J, Li X and Zhaojie C (2010) Life cycle assessment of four municipal solid waste management scenarios in China. *Waste Management* **30(11)**: 2362–2369, https://doi.org/10.1016/j.wasman.2010.03.038.

IL&FS (2010) *Technical EIA Guidance Manual for Common Municipal Solid Waste Management Facilities*. The Ministry of Environment and Forests, Government of India, New Delhi, India.

ISO (International Organization for Standardization) (2006a) *ISO 14040:2006 Environmental Management – Life Cycle Assessment – Principles and Framework*. ISO, Geneva, Switzerland. See http://www.iso.org (accessed 12/11/2020).

ISO (2006b) *ISO 14044:2006 Environmental Management – Life Cycle Assessment – Requirements and Guidelines*. ISO, Geneva, Switzerland. See http://www.iso.org (accessed 12/11/2020).

Khandelwal H, Dhar H, Thalla A and Kumar S (2019) Application of life cycle assessment in municipal solid waste management: a worldwide critical review. *Journal of Cleaner Production* **209**: 630–654, https://doi.org/10.1016/j.jclepro.2018.10.233.

Laurent A, Bakas I, Clavreul J *et al.* (2014) Review of LCA studies of solid waste management systems – part I: lessons learned and perspectives. *Waste Management* **34(3)**: 573–588, https://doi.org/10.1016/j.wasman.2013.10.045.

Leme M, Rocha M, Lora E *et al.* (2014) Techno-economic analysis and environmental impact assessment of energy recovery from municipal solid waste (MSW) in Brazil. *Resources, Conservation and Recycling* **87**: 8–20, https://doi.org/10.1016/j.resconrec.2014.03.003.

Mandal P, Chaturvedi M, Bassin J, Vaidya A and Gupta R (2014) Qualitative assessment of municipal solid waste compost by indexing method. *International Journal of Recycling Organic Waste in Agriculture* **3(4)**: 133–139, https://doi.org/10.1007/s40093-014-0075-x.

Manfredi S and Christensen T (2009) Environmental assessment of solid waste landfilling technologies by means of LCA-modeling. *Waste Management* **29(1)**: 32–43, https://doi.org/10.1016/j.wasman.2008.02.021.

Mendes M, Aramaki T and Hanaki K (2004) Comparison of the environmental impact of incineration and landfilling in São Paulo city as determined by LCA. *Resources, Conservation and Recycling* **41(1)**: 47–63, https://doi.org/10.1016/j.resconrec.2003.08.003.

MHA (Ministry of Home Affairs) (2011) *Census 2011*. Office of the Registrar General & Census Commissioner, India, Ministry of Home Affairs, Government of India, New Delhi, India. See https://censusindia.gov.in/ (accessed 10/11/2021).

Nieuwlaar E (2013) Life cycle assessment and energy systems. *Reference Module in Earth Systems and Environmental Sciences*, https://doi.org/10.1016/B978.0.12.409548.9.01334.8.

Norris GA (2001) The requirement for congruence in normalisation. *International Journal of Life Cycle Assessment* **6(2)**: article 85, https://doi.org/10.1007/BF02977843.

Planning Commission (2014) *Report of the Task Force on Waste to Energy (Volume I)*. Planning Commission, New Delhi, India. See http://swachhbharaturban.gov.in/writereaddata/Task_force_report_on_WTE.pdf (accessed 13/03/2021).

Pujara Y, Pathak P, Sharma A and Govani J (2019) Review on Indian municipal solid waste management practices for. *Journal of Environmental Management* **248**: 109238, https://doi.org/10.1016/j.jenvman.2019.07.009.

Rajaeifar M, Tabatabaei M and Ghanavati H (2015) Data supporting the comparative life cycle assessment of different municipal solid waste management scenarios. *Data in Brief* **3**: 189–194, https://doi.org/10.1016/j.dib.2015.02.020.

Rana R, Ganguly R and Gupta AK (2019) Life-cycle assessment of municipal solid-waste management strategies in Tricity region of India. *Journal of Material Cycles and Waste Management* **21(3)**: 606–623, https://doi.org/10.1007/s10163-018-00822-0.

Rathi S (2006) Alternative approaches for better municipal solid waste management in Mumbai, India. *Waste Management* **26(10)**: 1192–1200, https://doi.org/10.1016/j.wasman.2005.09.006.

Rebitzer G, Ekvall T, Frischknecht R *et al.* (2004) Life cycle assessment: part 1: framework, goal and scope definition, inventory analysis, and applications. *Environment International* **30(5)**: 701–720, https://doi.org/10.1016/j.envint.2003.11.005.

Rossi V, Cleeve-Edwards N, Lundquist L and Schenker U (2015) Life cycle assessment of end-of-life options for two biodegradable packaging materials: sound application of the European waste hierarchy. *Journal of Cleaner Production* **86**: 132–145, https://doi.org/10.1016/j.jclepro.2014.08.049.

SBM (Swachh Bharat Mission (Urban)) (2020) Editorial. *Swachhata Sandesh* **3(3)**: 3. See http://swachhbharaturban.gov.in/Auth/dsdocumentsfile.aspx?DOCTYPE=922&DOCID=701 (accessed 10/11/2021).

Seidel C (2016) The application of life cycle assessment to public policy development. *The International Journal of Life Cycle Assessment* **21(3)**: 337–348, https://doi.org/10.1007/s11367.015.1024.2.

Sharholy M, Ahmad K, Mahmood G and Trivedi R (2005) Analysis of municipal solid waste management systems in Delhi: a review. *Proceedings of the Second International Congress of Chemistry and Environment, Indore, India*, pp. 773–777.

Singh S and Gour A (2016) Status & impact of solid waste management in India. *India International Science Festival 2016, New Delhi, India*.

Song Q, Wang Z and Li J (2013) Environmental performance of municipal solid waste strategies based on LCA method: a case study of Macau. *Journal of Cleaner Production* **57**: 92–100, https://doi.org/10.1016/j.jclepro.2013.04.042.

Srivastava A and Nema A (2011) Life cycle assessment of integrated solid waste management system of Delhi. In *Towards Life Cycle Sustainability Management* (Finkbeiner M (ed.)). Springer, Dordrecht, the Netherlands, pp. 267–276.

Talyan V, Dahiya R and Sreekrishnan T (2008) State of municipal solid waste management in Delhi, the capital of India. *Waste Management* **28(7)**: 1276–1287.

Zaman AU (2009) Life cycle environmental assessment of municipal solid waste to energy technologies. *International Journal of Environmental Research* **3(3)**: 155–163.

How can you contribute?

To discuss this paper, please email up to 500 words to the editor at journals@ice.org.uk. Your contribution will be forwarded to the author(s) for a reply and, if considered appropriate by the editorial board, it will be published as discussion in a future issue of the journal.

Proceedings journals rely entirely on contributions from the civil engineering profession (and allied disciplines). Information about how to submit your paper online is available at www.icevirtuallibrary.com/page/authors, where you will also find detailed author guidelines.

Downloaded by [Delhi Technological University] on [09/12/21]. Copyright © ICE Publishing, all rights reserved.

2021 Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 28-29, 2021

Automatic Image Colorization with Convolutional Neural Networks

Mohammad Amir Qureshi
Electronics and Communication department
Delhi Technological University
Delhi, India
mohammadamirqureshi_2k17ec101@dtu.ac.in

Mayank Gautam
Electronics and Communication department
Delhi Technological University
Delhi, India
mayankgautam_2k17ec99@dtu.ac.in

Kaushal Pal
Electronics and Communication department
Delhi Technological University
Delhi, India
kaushalpal_2k17ec80@dtu.ac.in

N. Jayanthi
Electronics and Communication department)
Delhi Technological University
Delhi, India
njayanthi@dce.ac.in

2021 Asian Conference on Innovation in Technology (ASIANCON) | 978-1-7281-8402-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ASIANCON51346.2021.9544799

***Abstract*—Image colorization is the procedure of adding RGB color space value to the original greyscale image. Greyscale Image carries only the intensity information at each pixel value. This intensity mostly varies from 0 to 256 which is just enough to observe a gradual change in intensity. Image colorization is a very crucial process and has many benefits of its own for example in medical science, the entertainment industry, and the study of scientific and astrological images etc. We have developed a CNN model in which we designed our own CNN layers that takes a greyscale image as input and returns a colored image as output. The model proposed here does not require any human interaction in the process of colorization of the image once it is fed to the neural network.**

Keywords—convolutional neural network, deep learning, image recognition, image data generator.

I. INTRODUCTION

Grayscale images are those images which only contains brightness information but no color information. They contain all shades of grey which ranges from pure white to pure black. They are also called achromatic images. RGB meaning Red, Green and Blue. They are three fundamental colors of light which are used for creating different colors. Every light consists of these three colors.

Image colorization can be described as the process of assigning colors to the pixels of a black and white image. This issue is ill-posed in the sense that, without prior information regarding the image, there is often more than one possible colorization. Therefore, the colors of an object cannot be recognized by simply looking at its grayscale component. For example, all other things equal, a red balloon would most likely look the same as a green balloon in a greyscale image. This is why automatic colorization is a difficult task.

For this purpose CNN is used, it is a category of artificial neural network that is trained to identify and manage patterns. CNN's purpose is to identify patterns is what makes it so useful for image processing. CNNs are being used in a broad array of applications, namely pattern recognition, sound recognition, image recognition, and video analysis [15].

TensorFlow is a machine learning framework that makes acquiring data, training models, solving decisions, and refining accurate trend easier. It is a library for anyone using to perform numerical analysis and significant machine learning over the internet [4]. And Keras is python base deep learning framework which is high level API of TensorFlow. In Keras, building model is very simple as it's only consists of making the layers one over the other and connecting the graphs [3].

Image Data Generator is a class in Keras which gives a very fast and effortless method to supplement the image. At each epoch, the Image Data Generator class guarantees that the model generates new elements of the images. However, it only returns the modified images and does not include them in the existing collection [15].

II. RELATED WORK

The main objective of image colorization is to give colors to greyscale (black and white) images so that it gives meaning to the image. This area of study is very important because it can lighten up the interest or gives feeling to the very old images.

Many works had been done previously in the field of image colorization by many researchers and they had done this process with their very own and different methods and algorithms. Some of them are mentioned below:

Jeff Hwang and You Zhou developed a convolutional neural network model which automatically colorizes black and white images fed to it. Their model is based on many ImageNet-trained layers from VGG16 architecture [2].

Zezhou Cheng, Qingxiong Yang, and Bin sheng also presented a fully automatic colorization method using a neural network ensemble. In this method, DAISY and a new lingual feature are acting as output and will be given in the form of input in the neural network [9].

Yu Chen, Yeyun Luo, Youdong Ding, and Bing Yu also presented the novel neural network model which combines the two convolutional neural networks and uses multi-scale convolution kernels to get reliable spatial consistency [8].

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:36:29 UTC from IEEE Xplore. Restrictions apply.

Brian Sam Thomas, Bhaskar Dixit, Rajat Dogra, and Aditi Raut also presented a model on convolutional neural network (CNN)-Long short-term memory (LSTM) that takes the greyscale image as input and gives the corresponding colored image with some improvements to the loss function [7].

Jiancheng An*, Kpeyiton Koffi Gagnon, Qingnan Shi, Hongwei Xie, and Rui Cao also developed a model which uses VGG-16 CNN architecture which depends upon the loss of cross-entropy to produce reasonable colorization[12].

III. METHADOLOGY

We have designed our model in such a way that it can colorize different types of images efficiently and for this purpose, we have experimented and took images from different sources. Images from "people" categories are taken from the Unsplash website [6] and then from the "Open Country" category of MIT CVCL Urban and natural scene category dataset [5].

We have used the TensorFlow environment to build our model [4]. Initially, we split the data for which it is 95% for training and 5% for testing, and then import and added images to the environment. Along with that, we used Keras library [3] and Sequential to create the model by adding an initial input layer of the dimension (256, 256, 1) and (3,3) kernel size for convolution.

A. Model Building

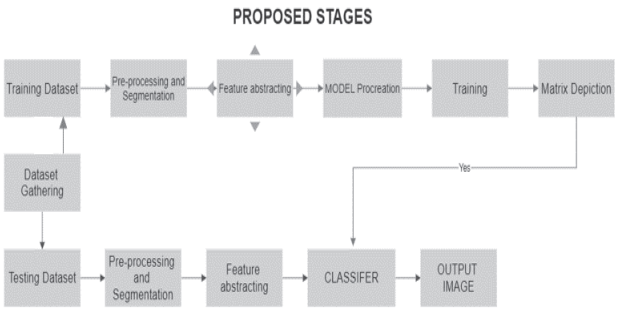


Fig. 1. Proposed CNN Model

The model is designed with convolutional neural networks which are widely used and essential for image recognition. We have not used any prebuilt architecture in our model. A flowchart of the proposed model is shown in figure 1[1]. Then hidden layers are added for the CNN and compiling of the model is done [1]. Eleven Hidden layers and one input layer with filters having 2 - 512 range are used in the model and almost all layers use the relu activation function except for the last layer which uses tanh as an activation function. Table 1 contains all the CNN layers-

TABLE I. CNN LAYERS

Layer	Filters	Activation	Padding	Strides
Input Layer	64	relu	same	-
Hidden Layer 1	64	relu	same	2
Hidden Layer 2	128	relu	same	-
Hidden Layer 3	128	relu	same	2
Hidden Layer 4	256	relu	same	-
Hidden Layer 5	256	relu	same	2
Hidden Layer 6	512	relu	same	-
Hidden Layer 7	256	relu	same	-
Hidden Layer 8	128	relu	same	-
Hidden Layer 9	64	relu	same	-
Hidden Layer 10	32	relu	same	-
Hidden Layer 11	2	tanh	same	-

We have used the relu activation function for most of our layers because it is a good approximator and it can approximate any function with a combination of different layers [1]. It is defined as-

$$f(x) = \max(0, x)$$

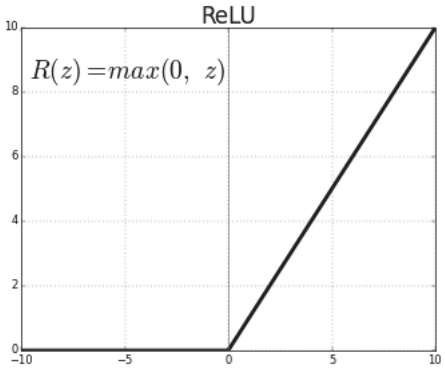


Fig. 2. relu Activation function

The rectified linear activation feature identifies and overcomes the vanishing gradient issue, enabling models to perform better. For designing Multilayer perceptron and convolutional neural networks, rectified linear activation is the default activation function.

To build the model we have used the sequential model with keras framework which allowed us to build the model layer by layer with different properties [3]. After that, once the model was built, we worked on training the model. Here we have used 500 epochs, 30 steps per epoch and measured accuracy and loss at every epoch cycle while training the model.

Then we have used Image Data Generator for augmentation which helps us to augment the images while the model is still in training [19].

We have revised our data, for making it readable to the model, and then fed the training images to the CNN model. The training data is fitted on our training dataset. We then finally test our model on the testing dataset.

The trained model is given to a classifier in the neural network to carry out the colorization of new inputs. The Classifier Identifies the input categories based on the data from the provided information.

After that we have also calculated values such as mean square error (MSE) [1], Root Mean squared error (RSME), cosine similarity, etc. for benchmarking the model.

B. Formulae Used

1) RMSE (Root Mean Square Error):

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

2) MSLE (Mean Squared Logarithmic Error):

$$\frac{1}{n}\sum_{i=1}^{n}[\log(x_i + 1) - \log(y_i + 1)]^2$$

where x= predicted value, y= actual value.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:36:29 UTC from IEEE Xplore. Restrictions apply.

3) MAE (Mean Absolute Error):

$$\frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

where y= predicted value, x= actual value.

4) MSE (Mean Squared Error) [1]

$$\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Finally, we plotted the model accuracy and model loss and manually tested our model by inputting greyscale images to the model and verified the colorized images at the output.

IV. Dataset Used

We have used two datasets on our model. Table 2 shows the different datasets with training and testing images.

TABLE II. NUMBER OF TRAINING AND TESTING IMAGES IN DATASETS

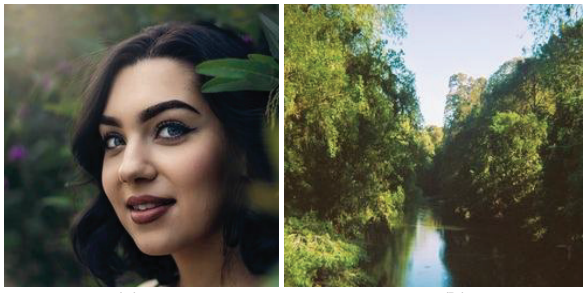
Datasets	Training	Testing
Unsplash	840	22
MIT CVCL	800	22

All the images that are used for testing and training purpose are in the "JPG" format and the dimensions are 256x256x3 for training dataset and 265x256x1 for testing dataset. We have taken images from the "people" category on Unsplash website[6] which contains indefinite number of images. We have made a dataset from the website to test the model.

Then, to more diversify the capabilities of our model we have taken images from "Open Country" category of MIT CVCL Urban and Natural Scene Categories dataset which contains thousands of images divided into eight different categories [5].

A. Training Dataset

The training dataset is taken from MIT CVCL Open Country dataset [5] and Unsplash website [6]. Then we have downloaded the JPG type images for training of our model. Then we made separate dataset for MIT CVCL Open Country test set [5] and Unsplash website [6]. Different images for training and testing is taken from these datasets. For training dataset we took 840 images from Uplsash Website which contains color images of human subjects with 256x256x3 dimension and 800 images from MIT CVCL Open Country test set which contain various color images of different objects with dimension of 256x256x3. Some of the images of training dataset are given in figure 3.



(a) (b)

Fig. 3. Training Dataset: (a) Unsplash Website (b) MIT CVCL

B. Testing Dataset

The testing dataset is also taken from MIT CVCL Open Country dataset [5] and Unsplash website [6]. For both of these datasets we have downloaded 22 images each, of type grayscale with 256x256x1 image dimension and JPG file format. Then we can finally evaluate the performance and efficiency of our model and find its accuracy. Some of the images of testing dataset are as follows are given in figure 4.



(a) (b)

Fig. 4. Testing Dataset: (a) Unsplash (b) MIT CVCL

V. Observation and Results

During the training process we have calculated the accuracy and loss of the model with each epoch cycle and saw improvement as the number of epochs passed. And finally we got the accuracy graph as shown in figure 5.

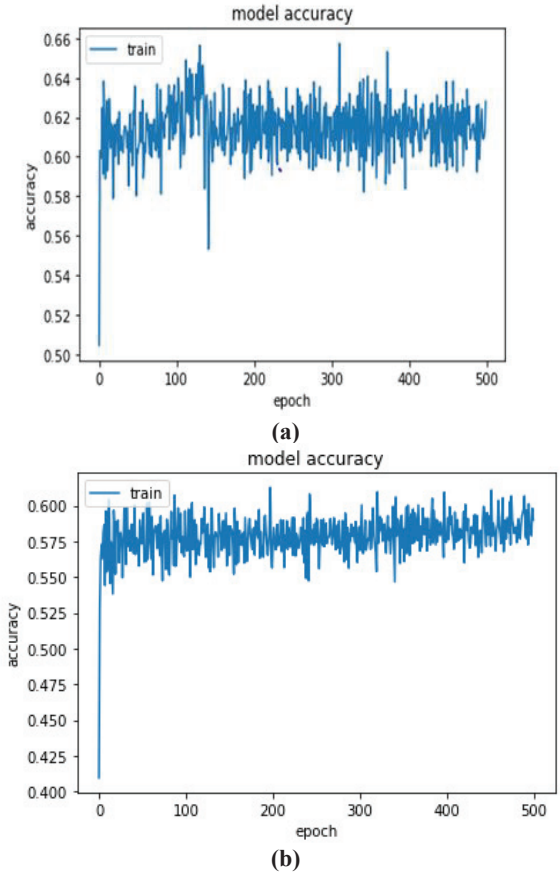


Fig. 5. Accuracy of model (a)MIT CVCL (b) Unsplash Website

We have measured the accuracy and loss at every epoch cycle, along with that we have also calculated Mean squared error, root mean squared error, mean squared logarithmic

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:36:29 UTC from IEEE Xplore. Restrictions apply.

error, and mean absolute error for better benchmarking of our model as shown in table 3.

TABLE III. MODEL ACCURACY AND LOSSES ON DIFFERENT DATASETS

Dataset	Accuracy	Loss/Error			
		Mean Squared Error	Root Mean Squared Error	Mean Squared Logarithmic Error	Mean Absolute Error
MIT CVCL	62.81%	0.0151	0.1228	0.0049	0.0859
Unsplash	61.79%	0.0110	0.1049	0.0054	0.0707

Also we have compared the accuracy of our model with the work done by "Jeff Hwang and You Zhou" [2] who worked on the same dataset of MIT CVCL[6]. We have used that same dataset and got better accuracy as shown in table 4.

TABLE IV. ACCURACY COMPARISON

Model	Accuracy (MIT CVCL Dataset)
Proposed Model	62.81%
Hwang and Zhou Model [2]	60.78%

An image from the final output is shown in figure 6.



Fig. 6. An image from colorized images

VI. CONCLUSION

We have developed a CNN model for automatic colorization of greyscale images using convolutional neural network. We have built our own model with 11 hidden layers that uses backpropagation to generate colors for greyscale images. To make our model more robust we have used two different datasets for testing and training of our model. We were able to achieve an accuracy of 62.81% of the same dataset previously used by "Jeff Hwang and You Zhou"[2]. In the near future we also tend to implement the same deep learning algorithm for colorization with modern CNN architectures like GoogLeNet and ShuffleNet.

REFERENCES

[1] S. Titus and J. R. N.M., "Fast Colorization of Grayscale Images by Convolutional Neural Network, "International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET), 2018.

[2] Jeff Hwang and You Zhou, "Image Colorization with Deep Convolutional Neural Networks".

[3] The keras website, [Online], Available: https://keras.io/.

[4] The TensorFlow website, [Online], Available: https://www.tensorflow.org/.

[5] The MIT CVCL Dataset[Online]: http://places2.csail.mit.edu/download.html

[6] The Unsplash website [Online]: https://unsplash.com/

[7] B. S. Thomas, R. Dogra, B. Dixit and A. Raut, "Automatic Image and Video Colorization using deep learning," International Conference on Smart City and Emerging Technology (ICSCET), 2018

[8] Y. Chen, Y. Luo, Y. Ding, and B. Yu, "Automatic Colorization of Images from Chinese Black and White Films Based on CNN," International Conference on Audio, Language and Image Processing (ICALIP), 2018.

[9] Z. Cheng, Q. Yang and B. Sheng, "Colorization Using Neural Network Ensemble," in IEEE Transactions on Image Processing, Nov. 2017.

[10] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks", In International conference on artificial intelligence and statistics, 2010.

[11] S. Narang, S. Bansal and Deepali, "Still Image Colorization using CNN," International Journal of Engineering Science and Computing, May 2019.

[12] J. An*, K. Koffi Gagnon, Q. Shi, H. Xie and Rui Cao, "Image Colorization with Convolutional Neural Networks", 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2019.

[13] J. Li, H. Xiao, D. Tan, M. Zhang and Y. Liu*, "Image Colorization Based on Texture by Using of CNN," IEEE 4th International Conference on Image, Vision and Computing (ICIVC), 2019.

[14] A. Krizhevsky, I. Sutskever, G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks," Neural Information Processing Systems, 2012.

[15] Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," International Interdisciplinary PhD Workshop (IIPhDW), 2018.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:36:29 UTC from IEEE Xplore. Restrictions apply.

Banking Reforms and Competition: Performance of Indian Banking Industry

AAYAM
Vol. 11 No. 1 Page 14-25
January-June, 2021 © AKGIM
AKGIM Ghaziabad
www.akgim.edu.in

Renu Gupta*
K. V. Bhanu Murthy**

ABSTRACT

Performance is the end result of the activities undertaken or behavior exhibited by the firm in the face of the dynamics of the environment in which it is placed. Furthermore, the performance of any market or industry can also be known at macro level provided that element of competition is also studied specifically that is generally eluded while evaluating a firm'sperformance alone.

This paper aims at analyzing and comparing the performance of the four banking segments understudy, i.e., SBI group, nationalized banks, old private banks and new private banks in terms of operating profits, net profits and spread for the period ranging between 1995-96 to 2009-10 making use of structure-conduct-performance approach.Results depicted by preliminary data analysis and panel regression analysis highlight that new private banks are remarkably ahead as compared to incumbents in terms of three profitability variables analyzed.

Key Words: Banking Reforms, Competition, S-C-P Paradigm, Performance of Indian Banks, Operating Profits, Net Profits and Spread.

I INTRODUCTION

At micro level, the health of any entity or firm can be determined both in quantitative and qualitative terms by judging its performance. Performance is the end result of the activities undertaken or behavior exhibited by the firm in the face of the dynamics of the environment in which it is placed. On the similar lines, the performance of any market or industry constituting group of firms can be known at macro level with the only difference that element of competition is also needed to be studied specifically at macro level that is generally eluded while evaluating the performance of a firm alone.

The concern of an economist always lies in determining the enhancement in economic welfare and consumer surplus as consequent to technical efficiency and all ocative efficiency of the industry as a whole. In simple words, an economist is mostly interested with the performance of firms to know whether or not operations of firms lead to economic welfare and how well firms satisfy consumer requirements. Being productively efficient, are they avoiding wasteful use of available factors of production? Also, are they being allocatively efficient in producing the 'right' goods and in the 'right' quantities? These issues may be resolved by considering aspects of performance such as the relationship between prices and costs, and the level of profits earned. According to Burgess (1988), performance is a judgement about the results of market behavior. As per S-C-P paradigm, basic conditions impact market structure, market structure impacts conduct and it is the conduct or behavior of the market or industry that influences performance finally.

Hence, performance is the final stage in understanding competition. Competition encompasses basic conditions, market structure, conduct and impact of all of these on performance. After Analyzing structure of Indian banking industry (Murthy, Gupta and Deb, 2016) and conduct of Indian banking industry (Murthy, Gupta and Deb, 2015);and proving that banking reforms have actually infused competition in Indian banking by way of deregulation and opening up. The present paper aims at analyzing and comparing the performance of the four banking segments under analysis, that is, SBI group, nationalized banks, old private banks and new private banks in the terms of the three performance variables namely operating profits, net profits and spread for the period between 1995-96 to 2009-10 in the light of S-C-P framework.

In this backdrop, the present paper has been segregated into nine Sections. Section II discusses the conceptual framework in banking. The literature review is given in Section III. Section IV highlights the variables selected for the purpose of the study. Methodology is expounded in Section V and Section VI presents the hypotheses. Preliminary data analysis and panel regression analysis have been made in Section VII and Section VIII respectively. Finally, conclusion is given in Section IX.

II. CONCEPTUAL FRAMEWORK

This paper uses structure-conduct-performance paradigm that was pioneered by Mason and Bain (1939) which states that there are some basic conditions which are given; these basic conditions impact market structure; market structure influences conduct or behaviour of industry; and conduct influences the industry's performance. Hence, present paper is

*Associate Professor, Départements of Commerce SGGSCC, University of Delhi.
**Professor, Delhi Technological University, Former Dean, Faculty of Commerce and Business and Head, Department of Commerce, University of Delhi.

2021 Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 28-29, 2021

BGR to HSV based Text Extraction from Manuscripts Using Slidebars

Mayank Singh
M.tech (Signal Processing and Digital Design)
Delhi Technological University
Delhi, India
mayanksingh9634@gmail.com

Prof. S. Indu
Electronics and Communication Engineering
Delhi Technological University
Delhi, India
s.indu@dce.ac.in

2021 Asian Conference on Innovation in Technology (ASIANCON) | 978-1-7281-8402-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ASIANCON51346.2021.9544674

***Abstract*-Text extraction is a very important step in language processing and pattern recognition. The method proposed here first converts the images to the HSV color space and then their Hue(H), Value(V), and Saturation(S) is varied manually with the help of the slide-bars introduced. The experimental results show the algorithm used to extract the text of a palm leaf manuscript performs fairly well.**

Keywords – New Gamma Correction Method, Gamma Correction Method, Mean Square Error, Power to Signal Noise Ratio.

I. INTRODUCTION

Text extraction analysis is the process of extracting named entities from unstructured text such as: real-world images, some unrecognizable text, manuscripts etc. It is composed of several steps: Identification, localisation, orientation, derivation, improvisation and identification. The three initial steps are usually assumed to be one, but they are quite different from each other. The detection detects text in every single image. Localization checks the whereabouts of the text present in the image and creates bounding boxes enclosing the text. The location is determined and is stored in the memory to decrease the pre-processing time. In the extraction step the image is remodelled into a binary image after which the image is divided into foreground and background where the text is preferred to be on the foreground after which the improvement is done to yield text with a better resolution. Optical character recognition (OCR)[1] or the Natural Language Processing(NLP)[2] technique helps in recognition, which consequently transforms an analog (printed) text to a digital file made of ASCII characters in lieu of analog file (pixels).

The efficiency of a specific algorithm is determined by how well it detects the text present in the image. This property is exploited in this paper and the resultant image is a clear evidence of the effectiveness of the proposed method. The algorithms reviewed here are not designed to draw out text from the manuscripts as the text is embedded in manuscripts which is more difficult to extract than the normal images. The difficulty in text extraction relies on different aspects like height, length, depth of the text and the background complexity.

Existing methods tried to alleviate this problem. An algorithm was proposed by Zhong et al.[3] based on RGB histogram, which was calculated using a finite number of classes. Their distinct peaks were drawn out again and again. All derived portions were then categorised into “non-text” or “text”. Zhou et al.[4] combined RGB color distance with spatial color proximity in the image space and proposed a graduated clustering algorithm. Some interesting results(partially) were yielded by the algorithm based on images obtained from compound internet streamers. A continuous approach was proposed by Sin et al.[5] where each region was transfigured into a single dimension sequence. If the Fourier spectrum’s auto-correlation function contained a relevant amount of peaks the region was affirmed as a text region.

Amid the recent research in text extraction, Sumathi et al. (2012) presented a detailed description of methods already in existence. In 2014 a novel algorithm was proposed by them for text extraction in the image. This technique allowed to omit the image background on the basis of Gamma Correction Method (GCM)[6]. Yet, it is time consuming for text extraction due to tedious calculations of various tasks in various GCM blocks which is disadvantageous for real-time purposes.

Another recent algorithm was proposed by Mohamed et al.[7], an effective Gamma Correction Method acceleration referred as New Gamma Correction Method (NGCM) which is the derivative of GCM. NGCM reduces the execution time of GCM while ensuring the same reliability. The gamma values in various portions of the image is predicted and then the corresponding pixels are transformed relative to the predicted gamma value. This method does not incorporate the depth of the text that’s why it cannot be used on manuscript images

The algorithm proposed here is used for text extraction in palm leaf manuscripts. Text extraction here is performed excluding the text recognition step. The working process of the algorithm introduced here is much simpler than all of the previous algorithm as here the image is transfigured to HSV color space[8]. Consequently, it’s saturation, hue and value estimates are varied manually just as the output the user needs.

The succeeding paper is compiled as follows, GCM algorithm is briefly described in section 2. Section 3 introduces the proposed new gamma correction method. Experimental results are displayed in section 4. The conclusions are drawn in section 5.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:43:33 UTC from IEEE Xplore. Restrictions apply.

II. PROPOSED METHODOLOGY

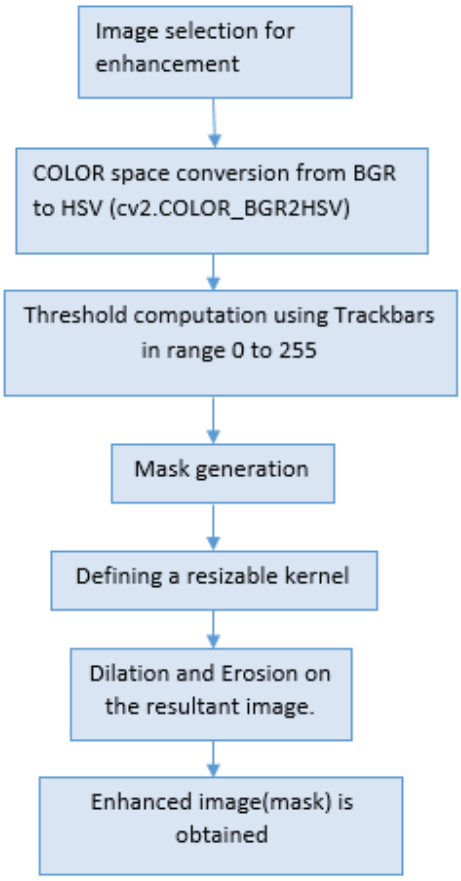


Fig. 1. Flowchart of the proposed method

The flow-chart for the proposed methodology is shown here:

- An image is chosen from the numerous images present in the dataset collected. The image is chosen such that some of the text is discernible and some is not which is retrieved through image enhancement done here.
- The color space is being changed from the initial BGR to HSV[9].

Color space transformation is described as conversion of a color representation from one base to other. It is usually seen in context of an image transformation being represented in one color space to the other color space. The objective is to construct the transformed image as identical as possible to the original image.

A. Color Space

A color space[10] is another name for a pre-defined arrangement of colors. "COLOR SPACE" is a useful conceptual tool for transcribing the color capacities for a specific file or device. The color, when reproduced on a different device, the color space indicates the degree of compromise for color saturation, shadow/highlight detail.

1) BGR color space

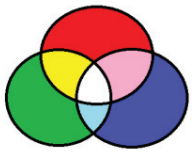


Fig. 2. Visual representation of BGR color space

The default color space used by OpenCV is RGB but the values in it is stored as BGR format [11] where the order of the colors is reversed. Significance of 'Red' and 'Blue' will be reversed.

2) HSV color space

HSV is also known as hue, saturation value.

In HSV color space is the color information is stored in a cylindrical format of RGB color. This color space depicts the color as it is visible to the human eye.

- After the image is converted to the HSV color space, track-bars are defined for Hue, Saturation and Value. Two track-bars are defined for each of them correspondingly of minimum and maximum value, for ex: "hue_min", "hue_max", "sat_min", "sat_max", "val_min", "val_max".
- After defining the Track-bars the values of the H, S, V are varied manually, and a corresponding mask is generated.
- A resizable kernel is defined which means that we can change the degree of dilation and erosion that we want and can experiment with different kernels.

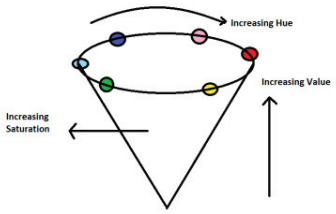


Fig. 3. Figure 1: Visual representation of HSV color space

3) Kernel

A small numbers' matrix utilised for image convolutions is defined as a kernel[12]. Different dimensions of the matrix of the kernel containing different combinations of numbers results in different outputs under convolution.

For example: A 3 × 3 kernel is shown below

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:43:33 UTC from IEEE Xplore. Restrictions apply.

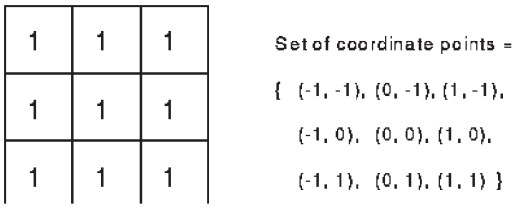


Fig. 4. An example of kernel matrix

A 'kernel' is also known as a 'structuring element' which is an indistinguishable object employed in 'mathematical morphology'. However, the structuring element also has a fixed origin.

- The kernel defined in the previous step is now used to perform dilation and erosion by using the function "cv2.morphologyEx(mask, cv2.MORPH_GRADIENT, kernel)"
- cv2.morphologyEx: This function is used to perform morphological transformations on the image. These are some simple morphological operations [13]with their basis as the image shape usually performed on the binary image. This function takes two inputs in the form of the original image and the structuring element or kernel. It's two primary operations are dilation and erosion with their derivatives like opening, closing etc.

4) Erosion

The principal idea of erosion is to erode the boundaries of the foreground object (The foreground should be kept white). This is achieved by the kernel sliding over the boundary of the object and the pixels under it are observed if all the pixels are '1' then it is kept as it is otherwise it is eroded which results in the boundary pixels getting discarded according to the size of the kernel.

5) Dilation

- It is a dual of erosion, here if even one pixel under the kernel is '1' then all the pixels are considered as '1' and thus it expands the white region in the image.
- Now after all of the above steps are done the final mask is obtained which is the final result.

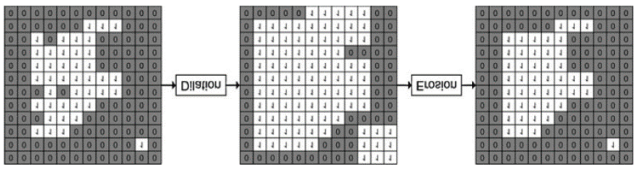


Fig. 5. Pictorial representation of Dilation and Erosion

B. BGR(RGB) to HSV Color space conversion

The translation of each pixel of BGR (Blue, Green, Red) to HSV (Hue, Saturation and Value) is presented.

- Hue: Responsible for detecting the pieces of the image which appear to be similar the perceived colors i.e. Blue, Green, Red or a combination of them.
- Saturation: The "colourfulness of an impulse relative to its own brightness".
- Value (or Brightness): Responsible for determining where the image will be lighter or darker.

Calculations have been in the following manner[9]:

$$R' = \frac{R}{255};\ G' = \frac{G}{255};\ B' = \frac{B}{255} \tag{1}$$

$$C_{max} = MAX(R', G', B') \tag{2}$$

$$C_{min} = MIN(R', G', B') \tag{3}$$

$$\Delta = C_{max} - C_{min} \tag{4}$$

$$H = \begin{Bmatrix} 60^o \times \left(\frac{G' - B'}{\Delta} mod\ 6\right), C_{max} = R' \\ 60^o \times \left(\frac{B' - R'}{\Delta} + 2\right), C_{max} = G' \\ 60^o \times \left(\frac{R' - G'}{\Delta} + 4\right), C_{max} = B' \end{Bmatrix} \tag{5}$$

$$S = \begin{Bmatrix} 0, & C_{max} = 0 \\ \frac{\Delta}{C_{max}}, & C_{max} \neq 0 \end{Bmatrix} \tag{6}$$

$$V = C_{max} \tag{7}$$

III. RESULTS

The dataset contains images of the manuscripts written on the palm leaves as showed in figure 6. Since these manuscripts are ancient, most of the leaves have deteriorated due to environmental effects. Contemporarily if researchers try to interpret the text written on these manuscripts, then it would be difficult due to degradation. Therefore, we apply text extraction methods to recover the writings.

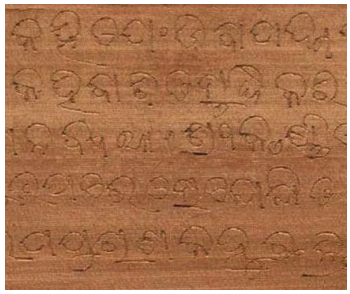


Fig. 6. A small part of an image from the dataset

The figure 7, 8 and 9 shown below depicts the text extracted from the manuscript image using BGR to HSV based text extraction incorporating slidebars at different thresholds.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:43:33 UTC from IEEE Xplore. Restrictions apply.

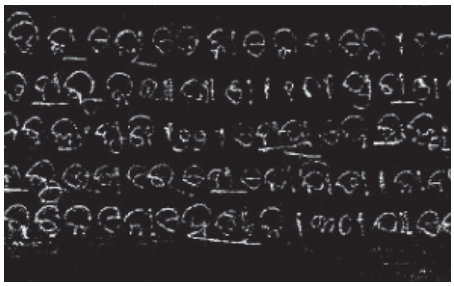


Fig. 7. A small part of the output image after going through the method proposed(threshold = 244)

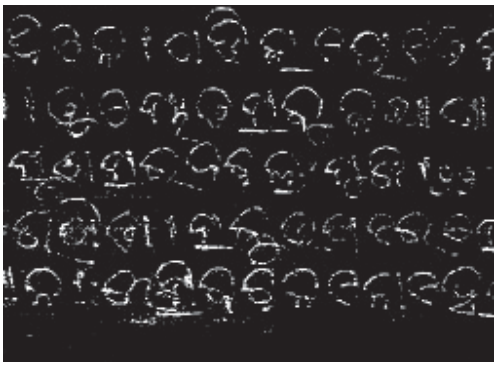


Fig. 8. A small part of the output image after going through the method proposed(threshold = 200)

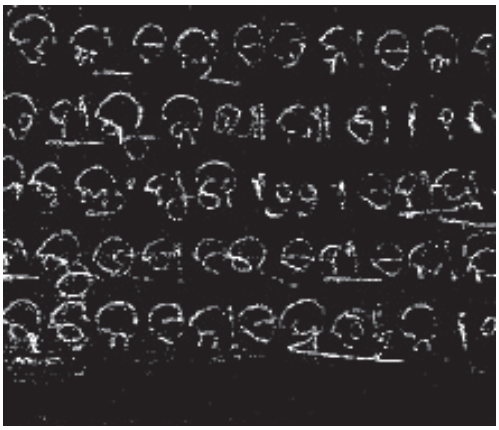


Fig. 9. A small part of the output image after going through the method proposed(threshold = 150)

TABLE I. ENTROPY(EN) OF THE SAME IMAGE CORRESPONDING TO DIFFERENT THRESHOLDS(TH)

Hue(0 - 179)		Saturation(0 - 255)		Val(0 - 255)		th	En
min	max	min	max	min	max		
Thresholded Image						-	7.39 (OrigImg)
0	29	0	255	123	154	15	4.76
0	18	115	255	48	145	50	3.83
0	20	109	219	56	149	75	4.22
0	21	86	240	66	149	100	4.21
0	27	100	217	69	153	150	4.64
0	89	0	255	0	147	200	4.09
0	19	0	254	52	152	244	4.57

TABLE II. PSNR OF THE SAME IMAGE WITH DIFFERENT THRESHOLDS

img 1	img 2 (threshold)	PSNR(img 1 – img 2)
Orig Img	15	27.84
Orig Img	50	27.85
Orig Img	75	27.85
Orig Img	100	27.85
Orig Img	150	27.85
Orig Img	200	27.84
Orig Img	244	27.85

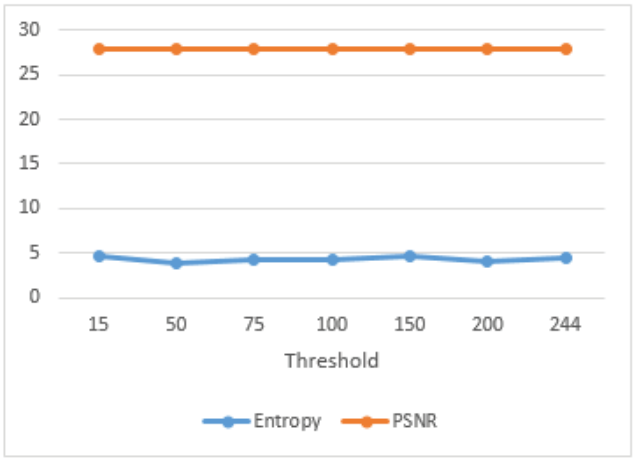


Fig. 10. Resultant image's Entropy vs PSNR

Entropy[14] is also known as average information of an image. It is employed to compute the randomness or disorder in the image and hence can be said to be a measure of uncertainty. It is used to estimate the how similar the original image is to the edited image. It can also be said to be a criterion for assessment for the ciphering algorithms, more entropy means better ciphering.

As it can be seen in Table 1, Entropy of the thresholded images is less than the entropy of the original image which indicates that there will be some information loss or it can also be said that all of the characters in the manuscripts will not be recognized and will be left out for which we took our image on different thresholds and then calculated every image's entropy after which the image with the highest entropy is accepted as result even if it's a very less difference in entropy it will e]mean at-least one more recognized character.

IV. DISCUSSION

This section highlights the efficiency of the method proposed. The method is compared with the existing NGCM method in order to depict the similarity between the enhanced image results. The proposed technique is simpler as compared to NGCM with almost the same or even better output in some cases.

A. Comparative Work

TABLE III. COMPARISON BETWEEN THE TWO METHODS BASED ON PSNR

Image	Entropy
Original	6.603
NGCM method	1.096
Proposed method	1.534

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:43:33 UTC from IEEE Xplore. Restrictions apply.

TABLE IV. COMPARISON BETWEEN THE TWO METHODS BASED ON ENTROPY

Image	PSNR
Original Image, NGCM Image	28.153
Original Image, Proposed method Image	28.031

As shown in Table 3 and Table 4, the entropy provided by the method proposed is more than NGCM so it means that the images encoded after being edited the method proposed will be safer than the NGCM one's or it can also be said that the image produced by our method has more information than the NGCM.

MSE[15] is the most preferred measurement metric of image quality. The values tending towards zero are considered to be better. It describes the Root-Mean-Square Deviation (RMSD) or Root-Mean-Square Error (RMSE)[15]. It is frequently termed as variance's standard deviation. The term Mean Squared Deviation (MSD) is also used for MSE. The process of measuring an unobserved quantity of image is done using estimator. MSD gives an average value for the square of the errors.

PSNR[15] is the ratio of the maximum signal power to the distorting noise power. It impacts representation quality (in decibels). It is frequently evaluated using logarithm of decibel as the signals' dynamic range is quite broad. The range lies within the maximum and the minimum achievable values which are varied in accordance with their quality. PSNR is a technique for quality assessment. It measures the reconstruction quality of leaky image compression codecs. The actual data is the signal (assumed) while the noise acts as the error occurrence due to distortion or compression.

For video and image compression quality degradation, the value of PSNR for 16-bit data lies between 60 to 80 dB and for 8-bit data representation, 30 to 50 dB The PSNR is calculated using the equation shown below:

$$PSNR = 10\, log_{10}\,(peakval^2)/MSE$$

As shown here PSNR of Original Image with the NGCM image and PSNR of original image with the method proposed does have a significant difference, so the method proposed here can be used to do almost everything that NGCM does and more.

NGCM algorithm was mainly developed to extract text from the real time images as in for the video stream or the webcam. Therefore, the same images have been used as a parameter for comparison between BGR to HSV based text extraction and NGCM. It is evident visually that the proposed method produces similar results to NGCM. Additionally, the BGR to HSV based text extraction gives four images stacked with each other. Images are stacked in the fashion of the figure given below.

Image 1	Image 2
Image 3	Image 4

Fig. 11. Mannerism of stacked imaged in the method

Every stacked image represents a different result described as follows:

- Image 1 displays the original mask image generated using the cv2.inRange() function after being converted to BGR color space.
- Image 2 displays the bitwise AND operation of the image with itself and the original mask generated.
- Image 3 displays the morphological transformed image (Dilated and eroded image)
- Image 4 displays the addition of two images which are the binary thresholded image and the Otsu thresholded image.

The best image among the four can be then selected as a result of text extraction.



Fig. 12. a) The original Image, b) Image output with NGCM, c) Image output with BGR to HSV text extraction using slidebars

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:43:33 UTC from IEEE Xplore. Restrictions apply.

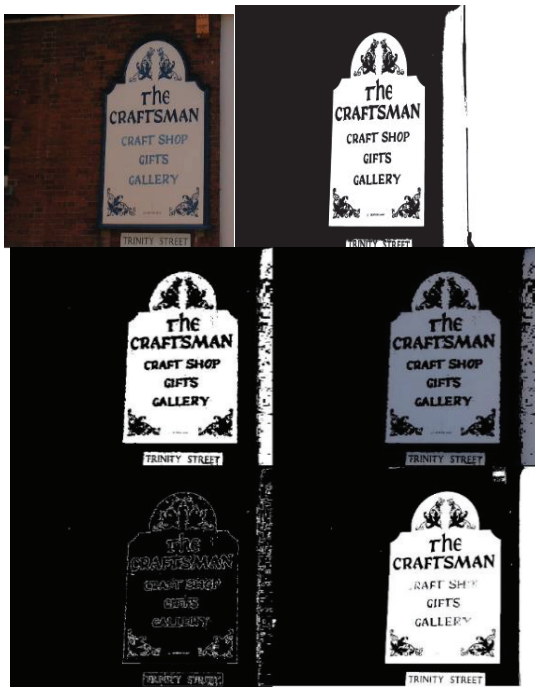


Fig. 13. a) The original Image, b) Image output with NGCM, c) Image output with BGR to HSV text extraction using slidebars



Fig. 14. a) The original Image, b) Image output with NGCM, c) Image output with BGR to HSV text extraction using slidebars

V. CONCLUSION

In this work, text extraction from the manuscripts is done using BGR to HSV based text extraction. The manuscript dataset was inherently difficult to decipher. This was due to same color of leaves and the text inscribed and damage over the years. The proposed method makes the text of the resultant image legible. This helps in easy interpretation of the text.

VI. FUTURE SCOPE

This can be further extended to recognise the text extracted and thus gather relevant information from ancient scripts. The output image can be enhanced in quality by efficient noise removal.

This algorithm can be automated using Deep learning or any other relevant method.

Algorithm can be further modified such that variation of hue in the image does not cause any changes to the image because it can be seen in the results that hue does not have a very significant impact on the results

REFERENCES

[1] R. Mithe, S. Indalkar, and N. Divekar, “Optical Character Recognition,” no. 1, pp. 72–75, 2013.

[2] G. G. Chowdhury, “Natural language Processing,” pp. 51–89.

[3] Y. U. Zhong, K. Karu, and A. K. Jain, “LOCATING TEXT IN COMPLEX COLOR IMAGES,” vol. 28, no. 10, 1995.

[4] D. Lopresti, “Locating and Recognizing Text in WWW Images,” vol. 206, pp. 177–206, 2000.

[5] B. Sin, S. Kim, and B. Cho, “Locating characters in scene images using frequency features,” pp. 489–492, 2002.

[6] G. G. Devi, “AUTOMATIC TEXT EXTRACTION FROM COMPLEX COLORED IMAGES USING GAMMA CORRECTION METHOD,” vol. 10, no. 4, pp. 705–715, 2014, doi: 10.3844/jcssp.2014.705.715.

[7] M. Amin, B. Atitallah, and A. Ben Atitallah, “New Gamma Correction Method for real time image text extraction,” pp. 2–7, 2019.

[8] D. Li, Ed., “HSV Color Space,” in *Encyclopedia of Microfluidics and Nanofluidics*, Boston, MA: Springer US, 2008, p. 793.

[9] G. Saravanan, G. Yamuna, and S. Nandhini, “Real Time Implementation of RGB to HSV / HSI / HSL and Its Reverse Color Space Models,” pp. 462–466, 2016.

[10] R. G. Kuehni, “Color Space and Its Divisions,” vol. 26, no. 3, 2001.

[11] I. Conference and D. Hutchison, *Image and*. 2014.

[12] V. Gupta, N. Sambyal, A. Sharma, and P. Kumar, “Restoration of artwork using deep neural networks,” *Evol. Syst.*, no. 0123456789, 2019, doi: 10.1007/s12530-019-09303-7.

[13] R. M. Haralick and S. R. Sternberg, “Image Analysis Morphology,” no. 4, pp. 532–550, 1987.

[14] C. H. Thum, “Measurement of the entropy of an image with application to image focusing,” *Opt. Acta (Lond).*, vol. 31, no. 2, pp. 203–211, 1984, doi: 10.1080/713821475.

[15] U. Sara, M. Akter, and M. S. Uddin, “Image Quality Assessment through FSIM, SSIM, MSE and PSNR—A Comparative Study,” *J. Comput. Commun.*, vol. 07, no. 03, pp. 8–18, 2019, doi: 10.4236/jcc.2019.73002.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:43:33 UTC from IEEE Xplore. Restrictions apply.

arXiv:2110.13068v1 [math.CV] 25 Oct 2021

BOHR PHENOMENON FOR K-QUASICONFORMAL HARMONIC MAPPINGS AND LOGARITHMIC POWER SERIES

KAMALJEET GANGANIA

Abstract. In this article, we establish the Bohr inequalities for the sense-preserving K-quasiconformal harmonic mappings defined in the unit disk $\mathbb{D}$ involving classes of Ma-Minda starlike and convex univalent functions, usually denoted by $\mathcal{S}^*(\psi)$ and $\mathcal{C}(\psi)$ respectively, and for $\log(f(z)/z)$ where f belongs to the Ma-Minda classes or satisfies certain differential subordination. We also estimate Logarithmic coefficient's bounds for the functions in $\mathcal{C}(\psi)$ for the case $\psi(\mathbb{D})$ be convex.

2010 AMS Subject Classification. 30B10, 30C45, 30C50, 30C80, 30C62, 31A05

Keywords and Phrases. Bohr radius, Radius problems, Harmonic mappings, Starlike and Convex functions, Logarithmic coefficients.

1. Introduction

Let $\mathcal{H}$ be the class of complex valued harmonic functions h (which satisfy the Laplacian equation $\Delta h = 4h_{z\bar{z}} = 0$) defined on the unit disk $\mathbb{D} := \{z \in \mathbb{C} : |z| < 1\}$, then we can write $h = f + \bar{g}$, where f and g are analytic and satisfies $h(0) = g(0)$. We say that h is sense-preserving in $\mathbb{D}$ whenever the Jacobian $J_h := |f'|^2 - |g'|^2 > 0$. A sense-preserving homeomorphism defined in $\mathbb{D}$ which is also harmonic is called K-*quasiconformal*, $K \in [1, \infty)$ if the (second complex) dilatation $w_h := g'/f'$ satisfies $|w_h(z)| \leq k$, $k = (K-1)/(K+1) \in [0, 1)$.

In 1914, H. Bohr [7] proved a power series inequality which is also known as the *classical Bohr inequality*:

Theorem 1.1 (Bohr's Theorem, [7])**.** *Let $f(z) = \sum_{m=0}^{\infty} a_m z^m$ be an analytic function in $\mathbb{D}$ and $|g(z)| < 1$ for all $z \in \mathbb{D}$, then*

$$\sum_{m=0}^{\infty} |a_m||z|^m \leq 1, \quad \textit{for} \quad |z| \leq \frac{1}{3}. \tag{1.1}$$

K. Gangania thanks to University Grant Commission, New-Delhi, India for providing Junior Research Fellowship under UGC-Ref. No.:1051/(CSIR-UGC NET JUNE 2017).

Since then several inequalities of similar kind are being studied in various context, which are usually named as *Bohr inequalities or Bohr type inequalities*. The readers are referred to see [2, 3, 6, 7, 9, 10, 12, 14, 15, 16, 19] and the references therein.

For instance if in the classical Bohr inequality in Theorem 1.1, we try to replace the initial coefficients $a_m, (m = 0, 1)$ by $|f(z)|$ and $|f'(z)|$, and further z by some suitable choice of functions $\omega(z)$ such that $|\omega(z)| < 1$. Or replace the Taylor coefficients a_m completely by the higher order derivatives of f. Then the combinations obtained lead us to so called Bohr-type inequalities. We now mention a few such combinations: Suppose that $f(z) = \sum_{m=0}^{\infty} a_m z^m$ be analytic in $\mathbb{D}$ and $a = |a_0|$ and $||f_0||_r^2 = \sum_{m=1}^{\infty} |a_{2m}| r^{2m}$, where $f_0(z) = f(z) - a_0$.

(i) $|f(z)|^n + \sum_{m=1}^{\infty} |a_m| r^m$, $n = 0$ or 1
(ii) $|f(z)| + |f'(z)||z| + \sum_{m=2}^{\infty} |a_m| r^m$
(iii) $|f(z)| + \sum_{m=N}^{\infty} \left|\frac{f^{(m)}(z)}{m!}\right| r^m$
(iv) $|f(\omega(z))| + \sum_{m=1}^{\infty} |a_m| r^m + \frac{1+ar}{(1+a)(1-r)} ||f_0||_r^2$

For some important work in this direction, we refer to see [21, 25].

Recall that f is subordinate to ϕ, written as $f \prec \phi$, if $f(z) = \phi(w(z))$, where w is a Schwarz function. Further if ϕ is univalent, then $f \prec \phi$ if and only if $f(\mathbb{D}) \subseteq \phi(\mathbb{D})$ and $f(0) = \phi(0)$. Note that the concept of subordination for analytic functions can be adopted for harmonic functions without any change, see [30]. Now let us consider the class $\mathcal{A}$ which consists of analytic function with power series of the form $f(z) = z + \sum_{m=2}^{\infty} a_m z^m$, and it's important subclass of univalent functions denoted by $\mathcal{S}$. Further, the classes of Ma-Minda starlike and convex function [13], respectively are defined as:

$$\mathcal{S}^*(\psi) := \left\{ f \in \mathcal{A} : \frac{z f'(z)}{f(z)} \prec \psi(z) \right\}$$

and

$$\mathcal{C}(\psi) := \left\{ f \in \mathcal{A} : 1 + \frac{z f''(z)}{f'(z)} \prec \psi(z) \right\},$$

where ψ is analytic and univalent with $\Re\psi(z) > 0$, $\psi'(0) > 0$, $\psi(0) = 1$ and $\psi(\mathbb{D})$ is symmetric about real axis. Note that $\psi \in \mathcal{P}$, the class of normalized Carathéodory functions. Also when $\psi(z) = (1 + z)/(1 - z)$, $\mathcal{S}^*(\psi)$ and $\mathcal{C}(\psi)$ reduces to the standard classes $\mathcal{S}^*$ and $\mathcal{C}$ of univalent starlike and convex functions.

In view of Muhanna [14], the class $S(f)$ of functions g subordinate to f has Bohr phenomenon if for any $g(z) = \sum_{m=0}^{\infty} b_m z^m \in S(f)$, there exist an $r_0 \in (0, 1]$ such that

$$\sum_{m=1}^{\infty} |b_m| r^m \leq d(f(0), \partial f(\mathbb{D})) \tag{1.2}$$

holds for $|z| = r \leq r_0$, where $d(f(0), \partial f(\mathbb{D}))$ denotes the Euclidean distance between $f(0)$ and the boundary of domain $f(\mathbb{D})$. When h in $S(h)$ is an

harmonic mapping, several Bohr inequalities have been investigated with certain assumptions on the analytic part f. Bhowmik and Das [5] assumed that h is a sense-preserving K-quasiconformal harmonic mapping, where the analytic part f is univalent of convex univalent function, and their result states as follows:

Theorem 1.2. [5, Theorem 1] *suppose that* $h(z) = f(z)+\overline{g(z)} = \sum_{m=0}^{\infty} a_m z^m + \overline{\sum_{m=1}^{\infty} b_m z^m}$ *be a sense-preserving* K*-quasiconformal harmonic mapping defined in* $\mathbb{D}$ *such that* f *is univalent and* $h_1(z) = f_1(z)+\overline{g_1(z)} = \sum_{m=0}^{\infty} c_m z^m + \overline{\sum_{m=1}^{\infty} d_m z^m} \in S(h)$. *Then*

$$\sum_{m=1}^{\infty}(|c_m| + |d_m|)r^m \leq d(f(0), \partial f(\mathbb{D}))$$

holds for $|z| = r \leq (5K + 1 - \sqrt{8K(3K + 1)})/(K + 1)$. *This result is sharp for the function* $p(z) = z/(1-z)^2 + k\overline{z/(1-z)^2}$, *where* $k = (K-1)/(K+1)$. *Moreover, if we take* f *to be convex univalent then the result holds for* $r \leq r_0 = (K + 1)/(5K + 1)$ *with sharpness for the function* $q(z) = z/(1 + z) + k\overline{z/(1 - z)}$.

Previously, several improved classical Bohr type inequality were discussed in [4] for the sense-preserving K-quasiconformal harmonic mappings such that f satisfy $|f(z)| < 1$ and it's applications to the corresponding analytic functions by taking $k = 0$ were shown. We state here one among them:

Theorem 1.3. [4, Theorem 2.9] *Let* $h(z) = f(z) + \overline{g(z)} = \sum_{m=0}^{\infty} a_m z^m + \overline{\sum_{m=1}^{\infty} b_m z^m}$ *be a sense-preserving* K*-quasiconformal harmonic mapping defined in* $\mathbb{D}$ *such that* $|f(z)| < 1$ *and* $0 \leq a = |a_0| < 1$. *Then the following inequality holds*

$$|f(z)| + \sum_{m=1}^{\infty}(|a_m| + |b_m|)r^m \leq 1,$$

for all $a \geq \alpha_k$ *and* $|z| = r \leq r_{a,k}$ *(the radius is sharp), where*

$$\alpha_k = \frac{\sqrt{k^2 + 12k12} - (2k + 3)}{k + 1} \quad \text{and} \quad r_{a,k} = \frac{B_{a,k} - (k + 2)(1 + a)}{2a^2(k + 1) + 2ak},$$

where $B_{a,k} = \sqrt{a^2(k^2 + 8k + 8) + 2a(k^2 + 6k + 4) + (k + 2)^2}$.

Motivated by the above results and noticing the role of sharp coefficient bounds for functions in a given class to establish such inequalities, see [3, 19, 20, 22, 23], and observing that sharp coefficients bounds are not available in general for the class $\mathcal{S}^*(\psi)$ and $\mathcal{C}(\psi)$. It is promising to discuss the case where analytic part f of a sense-preserving K-quasiconformal harmonic mapping belongs to the general class $\mathcal{S}^*(\psi)$ and $\mathcal{C}(\psi)$. More precisely,

Definition 1.4. Let $f \in \mathcal{S}^*(\psi)$ (or $\mathcal{C}(\psi)$), and $h(z) = f(z) + \overline{g(z)} = z + \sum_{m=2}^{\infty} a_m z^m + \overline{\sum_{m=1}^{\infty} b_m z^m}$ be a sense-preserving K-quasiconformal harmonic

mapping defined in $\mathbb{D}$. Then

$$S_\psi(h) := \{h_1(z) : h_1(z) \prec h(z)\},$$

where $h_1(z) = f_1(z) + \overline{g_1(z)} = z + \sum_{m=2}^{\infty} c_m z^m + \overline{\sum_{m=1}^{\infty} d_m z^m}$.

Similarly, in the literature of coefficient's problems, much of the attention has been on the inequalities related to the *logarithmic coefficients*, see [26], due to it's important role in settling the Bieberbach conjecture, we refer to see the recent articles [1, 28, 29] and their references, which are defined by the following logarithmic power series

$$\log\left(\frac{f(z)}{z}\right) = 2\sum_{m=1}^{\infty} \gamma_m z^m, \tag{1.3}$$

where $f \in \mathcal{S}$. Recently, Adegani et. al [1, Theorem 1, Sec 2] settled the problem of sharp bounds for γ_m for the functions in the class $\mathcal{S}^*(\psi)$, but obtained only three initial logarithmic coefficient's bounds for the class $\mathcal{C}(\psi)$ in [1, Theorem 2, Sec 2]. Therefore, we consider the problem of Bohr inequality similar to (1.1) for $\log(f(z)/z)$, where f either belongs to the classes $\mathcal{S}^*(\psi)$ and $\mathcal{C}(\psi)$ or satisfies certain standard differential subordination[27]. The reason why we consider the Bohr inequality of classical type for the series (1.3) instead of (1.2) follows with the observations discussed by Bhowmik and Das: The quantity $d(g, \partial\Omega)$, where Ω is the image of $\mathbb{D}$ under $g(z) := \log(f(z)/z)$, can be arbitrarily small positive number for $f \in \mathcal{S}$ which is observed with the help of univalent polynomials $f_n(z) = z + (z^2/n)$ for each $n \geq 2$ such that the image of $\log(f_n(z)/z)$ does not include the point $\log(1 + (1/n))$.

While proving our result related to Bohr inequality for the logarithmic power series, we surprisingly get estimates for the logarithmic coefficient's bounds for the class $\mathcal{C}(\psi)$ for the case $\psi(\mathbb{D})$ being convex, see Theorem 3.4. *Note that till date obtaining sharp bounds even for a particular class of convex functions is an open problem.*

2. Bohr phenomenon for K-quasiconformal mappings

For convenience, let us consider the functions $f_n \in \mathcal{S}^*(\psi)$, $n \in \{0, 1, 2, \cdots\}$ defined in the unit disk $\mathbb{D}$ as

$$\frac{z f_n'(z)}{f_n(z)} = \psi(z^{n+1}). \tag{2.1}$$

In case when the coefficients of f_n in its power series expansion are positive, we denote f_n by $\hat{f_n}$. Since sharp bounds for the Taylor coefficients in general for the functions in the class $\mathcal{S}^*(\psi)$ are yet not known. Therefore, we need the following result, obtained in [19, Lemma 2.1, Sec 2].

Lemma 2.1. *let $f(z) = \sum_{n=0}^{\infty} a_n z^n$ and $g(z) = \sum_{k=0}^{\infty} b_k z^k$ be analytic in $\mathbb{D}$ and $g \prec f$, then*

$$\sum_{k=N}^{\infty} |b_k| r^k \leq \sum_{n=N}^{\infty} |a_n| r^n$$

for $|z| = r \leq \frac{1}{3}$ *and* $N \in \mathbb{N}$.

which was further used to prove [19, Corollary 2.1, Sec 2], and there replacing $g(z)$ by $g(z)/M$ for $M > 0$ we easily get:

Lemma 2.2. *Let the analytic functions* f, g *and* h *satisfies* $g(z) = M\phi(z)f(\omega(z))$ *in* $\mathbb{D}$, *where* ω *is the Schwarz function. Assume* $|\phi(z)| \leq \tau$ *for* $|z| < \tau \leq 1$. *Then*

$$\sum_{k=N}^{\infty} |b_k| r^k \leq \tau M \sum_{n=N}^{\infty} |a_n| r^n, \quad \text{for} \quad 0 \leq |z| = r \leq \frac{\tau}{3},$$

where $M > 0$ *and* $N \in \mathbb{N}$.

The Lemma 2.2 generalizes the result of Alkhaleefah et al. [4, Theorem 2.1]. For other inequalities of this kind, see [17].

Now the following result is a generalization of Theorem 1.2 in the sense that f be a Ma-Minda univalent starlike functions, while it's convex analogue is presented in Theorem 2.8.

Theorem 2.1. *Let* $h(z) = f(z) + \overline{g(z)} = z + \sum_{m=2}^{\infty} a_m z^m + \overline{\sum_{m=1}^{\infty} b_m z^m}$ *be a sense-preserving* K*-quasiconformal harmonic mapping defined in* $\mathbb{D}$ *such that* $f \in \mathcal{S}^*(\psi)$ *and* $h_1(z) = f_1(z) + \overline{g_1(z)} = z + \sum_{m=2}^{\infty} c_m z^m + \overline{\sum_{m=1}^{\infty} d_m z^m} \in S_\psi(h)$. *Then*

$$\sum_{m=1}^{\infty} (|c_m| + |d_m|) r^m \leq d(f(0), \partial f(\mathbb{D})) \tag{2.2}$$

holds for $|z| \leq r^* = \min\{r_0, \frac{1}{3}\}$, *where* r_0 *is the unique root in* $(0, 1)$ *of*

$$\frac{2K}{K+1} \hat{f}_0(r) = -f_0(-1).$$

The result is sharp for the function $p(z) = f_0(z) + k\overline{f_0(z)}$, *where* $\hat{f}_0 = f_0$ *as defined in* (2.1) *when* $r^* = r_0$.

Proof. Clearly, $f'(z) \neq 0$ as $f \in \mathcal{S}^*(\psi)$. From the dilatation $w_h(z) = g'/f'$, we have $|w_f| \leq k = (K-1)/(K+1) < 1$. We consider the case w_h being non constant, and the case $w_f(z) = ck$, $|c| = 1$ can be handled on similar lines. Thus, by Maximum-Modulus principle there exist $\phi : \mathbb{D} \to \mathbb{D}$ such that

$$g'(z) = k\phi(z) f'(z). \tag{2.3}$$

Now in Lemma 2.2 setting $\tau = 1 = N$ and $w(z) = z$ with $p(z) = \sum_{n=0}^{\infty} p_n z^n$ and $q(z) = \sum_{n=0}^{\infty} q_n z^n$ such that $p(z) = M\phi(z)q(z)$ for some $M > 0$, $\phi : \mathbb{D} \to \mathbb{D}$, we have for $|z| = r \leq 1/3$

$$\sum_{n=0}^{\infty} |q_n| r^n \leq M \sum_{n=0}^{\infty} |p_n| r^n,$$

which on applying to (2.3) and Lemma 2.1 with $N = 1$ on $f(z)/z \prec f_0(z)/z$, and then integrating from 0 to r finally yield

$$\sum_{m=1}^{\infty} |b_m| r^m \leq k \sum_{m=1}^{\infty} |a_m| r^m \leq k \hat{f}_0(r) \tag{2.4}$$

for $|z| = r \leq 1/3$. Thus from (2.4), we have for $|z| = r \leq 1/3$

$$\sum_{m=1}^{\infty} |a_m| r^m + \sum_{m=1}^{\infty} |b_m| r^m \leq (1+k) \sum_{m=1}^{\infty} |a_m| r^m. \tag{2.5}$$

Now let us consider the function

$$T(r) := (1+k)\hat{f}_0(r) + f_0(-1), \quad 0 \leq r \leq 1.$$

Then it is a continuous function of r and $T'(r) \geq 0$. Also $T(0) = f_0(-1) < 0$ and $T(1) > 0$. Therefore, $T(r)$ has a unique root, say r_0 in $(0, 1)$. Hence by (2.5)

$$\sum_{m=1}^{\infty} |a_m| r^m + \sum_{m=1}^{\infty} |b_m| r^m \leq d(f(0), \partial f(\mathbb{D}))$$

holds for $|z| = r \leq \min\{r_0, 1/3\}$, where r_0 is the root of $T(r)$. Now it remains to prove that

$$\sum_{m=1}^{\infty} (|c_m| + |d_m|) r^m \leq \sum_{m=1}^{\infty} |a_m| r^m + \sum_{m=1}^{\infty} |b_m| r^m$$

is true for $|z| \leq r^*$ which in fact holds, since $f_1 \prec f$ and $g_1 \prec g$ [30, P. 164, Sec 2] with the application of Lemma 2.1 yields $\sum_{m=1}^{\infty} |c_m| r^m \leq \sum_{m=1}^{\infty} |a_m| r^m$ and $\sum_{m=1}^{\infty} |d_m| r^m \leq \sum_{m=1}^{\infty} |b_m| r^m$ for $r \leq 1/3$. Sharpness of the result follows for the function $p(z)$ as defined in the hypothesis by direct computation for $|z| = r \leq r^*$ such that equality holds in (2.2). □

In the next result, we establish the Bohr-Rogosinski phenomenon for the the class $S_\psi(h)$ following the proof of Theorem 2.1. Note that if we take $N \to 1$ and $n \to \infty$ in the result below, we trace back Theorem 2.1.

Corollary 2.2. *Let $h(z) = f(z) + \overline{g(z)} = z + \sum_{m=2}^{\infty} a_m z^m + \overline{\sum_{m=1}^{\infty} b_m z^m}$ be a sense-preserving K-quasiconformal harmonic mapping defined in $\mathbb{D}$ such that $f \in \mathcal{S}^*(\psi)$ and $h_1(z) = f_1(z) + \overline{g_1(z)} = z + \sum_{m=2}^{\infty} c_m z^m + \overline{\sum_{m=1}^{\infty} d_m z^m} \in S_\psi(h)$. Then*

$$|f(z^n)| + \sum_{m=N}^{\infty} (|c_m| + |d_m|) r^m \leq d(f(0), \partial f(\mathbb{D})) \tag{2.6}$$

holds for $|z| \leq r^* = \min\{r_0, \frac{1}{3}\}$, where $n \in \mathbb{N}$ and r_0 is the unique root in $(0, 1)$ of

$$(K+1)(\hat{f}_0(r^n) + f_0(-1)) + 2K(\hat{f}_0(r) - S_N(\hat{f}_0)) = 0,$$

where $S_N(f) = \sum_{m=1}^{N-1} a_m r^m$. The result is sharp for the function $p(z) = f_0(z) + k\overline{f_0(z)}$, where $\hat{f}_0 = f_0$ as defined in (2.1) when $r^* = r_0$.

Proof. Since $|f(z^n)| \leq \hat{f}_0(r^n)$ for $|z| = r$. Therefore, applying Lemma 2.1 on (2.3), and Lemma 2.1 on $f(z)/z \prec f_0(z)/z$, and then integrating from 0 to r, we deduce that

$$|f(z^n)| + \sum_{m=N}^{\infty} (|c_m| + |d_m|) r^m$$

$$\leq |\hat{f}_0(r^n)| + (1+k)\sum_{m=N}^{\infty}|a_m|r^m$$
$$\leq -f_0(-1) \leq d(f(0), \partial f(\mathbb{D}))$$

holds for $|z| = r \leq \{r_0, 1/3\}$, where $r_0 \in (0,1)$ is the unique positive root of the equation

$$T_N(r) := \hat{f}_0(r^n) + f_0(-1) + (1+k)\sum_{m=N}^{\infty}|a_m|r^m = 0.$$

Existence of the root r_0 follows, since $T_N(r)$ is continuous function of r and $T_N'(r) \geq 0$ with $T_N(0) = f_0(-1)$ and $T_N(1) > 0$. □

Corollary 2.3. [23, Theorem 5.1][20, Theorem 3.1] *Let $f \in \mathcal{S}^*(\psi)$ and $f_1(z) = \sum_{m=1}^{\infty} c_m z^m \in S_\psi(f)$. Then*

$$\sum_{m=1}^{\infty}|c_m|r^m \leq d(f(0), \partial f(\mathbb{D}))$$

holds for $|z| \leq r^* = \min\{r_0, \frac{1}{3}\}$, where r_0 is the unique root in $(0,1)$ of

$$\hat{f}_0(r) = -f_0(-1).$$

The result is sharp for the function $\hat{f}_0 = f_0$ as defined in (2.1) when $r^* = r_0$.

Corollary 2.4. *Let $f \in \mathcal{S}^*$ in Theorem 2.1. Then the inequality (2.2) holds for $|z| = r \leq r_0 < 1/3$, where*

$$r_0 = (5K + 1 - \sqrt{8K(3K+1)})/(K+1).$$

The radius is sharp.

Corollary 2.5. *Let $\psi(z) = \frac{1+Dz}{1+Ez}$ in Theorem 2.1, where $-1 \leq E < D \leq 1$. Then the inequality (2.2) holds for $|z| = r \leq \min\{r_0, 1/3\}$, where r_0 is the unique root of the equation*

$$\frac{2K}{K+1}(r + \sum_{m=2}^{\infty}\prod_{t=0}^{m-2}\frac{|E - D + Et|}{t+1}r^m) - (1-E)^{\frac{D-E}{E}} = 0.$$

Further assume that $f_0 = \hat{f}_0$ as defined in (2.1), and

(i) If $E \neq 0$ and $3(1-E)^{\frac{D-E}{E}} \leq (1+\frac{E}{3})^{\frac{D-E}{E}}$;
(ii) If $E = 0$ and $D \geq \frac{3}{4}\log 3$.

Then the radius r_0 is sharp.

Corollary 2.6. *Let $\psi(z) = \frac{1+(1-2\alpha)z}{1-z}$ in Theorem 2.1, where $0 \leq \alpha \leq 1/2$. Then the inequality (2.2) holds for $|z| = r \leq r_0 < 1/3$, where r_0 is the unique root of the equation*

$$K2^{2(1-\alpha)+1}r - (K+1)(1-r)^{2(1-\alpha)} = 0.$$

The radius is sharp.

Remark 2.1. Substituting $\alpha = 0$ in Corollary 2.6, we retrace Corollary 2.4. One should also note that if we take $K = 1$ and $\alpha = 0$ in Corollary 2.4, then the radius $r_0 = 3 - 2\sqrt{2}$, which is equal to the Bohr radius for the class $\mathcal{S}^*$. Moreover, thinking in the direction of inclusion relationship for a class $\mathcal{M}$ of starlike functions such that $\mathcal{M} \subset \mathcal{S}^*(\alpha)$ for some α, then we observe that Corollary 2.6 provides a lower bound for the radius r_0 in Theorem 2.1 for the case of $\mathcal{M}$ in place of $\mathcal{S}^*(\psi)$ in general.

Now recall that a function $f \in \mathcal{S}$ which has the property that for every circular arc Γ contained in $\mathbb{D}$ with center $\xi \in \mathbb{D}$, the image arc $f(\Gamma)$ is a starlike arc with respect to $f(\xi)$, is called Goodman uniformly starlike ($\equiv \mathcal{UST}$). Similary the class of uniformly convex function ($\equiv \mathcal{UCV}$) is also defined. Motivated by this class recently Darus introduced the class:

Let $f \in \mathcal{A}$. Then $f \in k - \mathcal{UCST}(\alpha)$ if and only if

$$\Re\left\{\frac{(zf'(z))'}{f(z)}\right\} > k\left|(1-\alpha)\frac{zf'(z)}{f(z)} + \alpha\frac{(zf'(z))'}{f(z)} - 1\right|,$$

$k \geq 0$, $0 \leq \alpha \leq 1$. Now following the Remark 2.1 and the inclusion relation given in [8, Theorem 2.1], Corollary 2.6 gives:

Example 2.7. *Let us replace $\mathcal{S}^*(\psi)$ by $k - \mathcal{UCST}(\alpha)$ in Theorem 2.1. Then the inequality (2.2) holds for $|z| = r \leq r_0 < 1/3$, where r_0 is the unique root of the equation*

$$\frac{2Kr}{(1-r)^{2(1-\delta)}} - \frac{K+1}{4^{(1-\delta)}} = 0,$$

where

$$\delta = \frac{(2\gamma - \beta) + \sqrt{(2\gamma-\beta)^2 + 8\beta}}{4}$$

and $0 \leq \delta < 1$, $\beta = \frac{1+\alpha k}{1+k}$ and $\gamma = \frac{1}{1+k}$ and further satisfy

$$(1-\alpha)k^2 - (1+\alpha)k - 2 \geq 0.$$

In view of the Remark 2.1, note that $k - \mathcal{UCST}(\alpha) \subset \mathcal{S}^*(\delta)$, and therefore the radius r_0 can be further improve.

Now we conclude this section with the following result which is convex analogue of Theorem 2.1:

Theorem 2.8. *Let $h(z) = f(z) + \overline{g(z)} = z + \sum_{m=2}^{\infty} a_m z^m + \overline{\sum_{m=1}^{\infty} b_m z^m}$ be a sense-preserving K-quasiconformal harmonic mapping defined in $\mathbb{D}$ such that $f \in \mathcal{C}(\psi)$ and $h_1(z) = f_1(z) + \overline{g_1(z)} = z + \sum_{m=2}^{\infty} c_m z^m + \overline{\sum_{m=1}^{\infty} d_m z^m} \in S_\psi(h)$. Then*

$$\sum_{m=1}^{\infty}(|c_m| + |d_m|)r^m \leq d(f(0), \partial f(\mathbb{D}))$$

holds for $|z| \leq r^* = \min\{r_0, \frac{1}{3}\}$, where r_0 is the unique root in $(0,1)$ of

$$\frac{2K}{K+1}\hat{f}_0(r) = -f_0(-1).$$

The result is sharp for the function $p(z) = f_0(z)+k\overline{f_0(z)}$ *when* $r^* = r_0$*, where* $\hat{f}_0 = f_0$ *is the solution of*

$$1 + \frac{zf''(z)}{f'(z)} = \psi(z).$$

Proof. It follows from Theorem 2.1 by suitably using Alexander's relation between starlike and convex functions. □

Corollary 2.9. *Let* $f \in \mathcal{C}$ *in Theorem 2.8. Then the inequality* (2.2) *holds for* $|z| = r \leq r_0 < 1/3$*, where*

$$r_0 = (K + 1)/(5K + 1).$$

The radius is sharp.

The above corollary was obatined by Liu and Ponnusamy [24, Theorem 1], (also see [11, Theorems 1.1 and 1.3]), where the analytic part f was not normalized.

3. Bohr phenomenon for Logarithmic power series

Let us recall that logarithmic coefficients γ_m of the functions $f \in \mathcal{S}^*(\psi)$ are defined as

$$\log\left(\frac{f(z)}{z}\right) = 2\sum_{m=1}^{\infty} \gamma_m z^m. \tag{3.1}$$

In the following theorem, we generalize and provide a simple proof of the result [5, Theorem 4, Sec 3].

Theorem 3.1. *Let* $\psi(z) = 1+B_1z+B_2z^2+\cdots$ *and* $f(z) = z+\sum_{m=2}^{\infty} a_m z^m \in \mathcal{S}^*(\psi)$*, where* $\psi(\mathbb{D})$ *is convex. Then the logarithmic coefficient's of* f *satisfies*

$$2\sum_{m=1}^{\infty} |\gamma_m| r^m \leq 1, \quad \textit{for} \quad |z| = r \leq 1 - \frac{1}{\exp\left(\frac{1}{|B_1|}\right)}.$$

The result is sharp for the case when the function f *defined as* $zf'(z)/f(z) = \psi(z)$ *is an extremal for the logarithmic coefficient's bounds.*

Proof. Recall the result [1, Theorem 1, Sec 2] which says: If $f \in \mathcal{S}^*(\psi)$, then γ_m as defined in (3.1) satisfies

$$|\gamma_m| \leq \frac{|B_1|}{2m}, \tag{3.2}$$

whenever $\psi(\mathbb{D})$ is convex. Hence

$$2\sum_{m=1}^{\infty} |\gamma_m| r^m \leq 2\sum_{m=1}^{\infty} \frac{|B_1|}{2m} r^m = |B_1| \sum_{m=1}^{\infty} \frac{1}{m} r^m.$$

Since the sum of the power series $\sum_{m=1}^{\infty}\frac{x^m}{m} = -\log(1-x)$ for $x \in (-1,1]$. Therefore, we have

$$2\sum_{m=1}^{\infty}|\gamma_m|r^m \leq |B_1|\log\left(\frac{1}{1-r}\right) \leq 1,$$

which holds whenever $r \leq 1 - 1/(\exp\left(\frac{1}{|B_1|}\right))$. □

Now we apply the above theorem to some well-known and recently introduced classes:

Corollary 3.2. *Let $g \in \mathcal{A}$ and $f \in \mathcal{S}^*(\psi)$, where $\psi(\mathbb{D})$ is convex. Then*

$$2\sum_{m=1}^{\infty}|\gamma_m|r^m \leq 1, \quad \textit{for} \quad |z| = r \leq r_0 = 1 - \frac{1}{\exp\left(\frac{1}{|B_1|}\right)},$$

follows for each one of the following cases:

(*i*) $B_1 = D - E$ *when* $\psi(z) = \frac{1+Dz}{1+Ez}$, *where* $-1 \leq E < D \leq 1$.
(*ii*) $B_1 = 2(1-\alpha)$ *when* $\psi(z) = \frac{1+(1-2\alpha)z}{1-z}$, *where* $0 \leq \alpha < 1$.
(*iii*) $B_1 = 2\eta$ *when* $\psi(z) = \left(\frac{1+z}{1-z}\right)^{\eta}$, *where* $0 < \eta \leq 1$.
(*iv*) $B_1 = (5\sqrt{2}-6)/(2\sqrt{2})$ *when* $\psi(z) = \sqrt{2} - (\sqrt{2}-1)\sqrt{\frac{1-z}{1+2(\sqrt{2}-1)z}}$.
(*v*) $B_1 = b^{\frac{1}{a}}/a$ *when* $\psi(z) = (b(1+z))^{1/a}$, *where* $a \geq 1$ *and* $b \geq 1/2$.
(*vi*) $B_1 = 1-\alpha$ *when* $\psi(z) = \alpha + (1-\alpha)e^z$, *where* $0 \leq \alpha < 1$.
(*vii*) $B_1 = (1-\alpha)/2$ *when* $\alpha + (1-\alpha)\sqrt{1+z}$, *where* $0 \leq \alpha < 1$.
(*vii*) $B_1 = 1/2$ *when* $\psi(z) = \frac{2}{1+e^{-z}}$.

The result is sharp for the case when the function f defined as $zf'(z)/f(z) = \psi(z)$ is an extremal for the logarithmic coefficient's bounds.

Remark 3.1. In Theorem 3.1, let $f(z) = z + \sum_{m=2}^{\infty} a_m z^m \in \mathcal{S}^*(\psi)$, where $\psi(\mathbb{D})$ is starlike with respect to 1 but convex. Then the logarithmic coefficients of f satisfies, see [1, Theorem 1, Sec 2],

$$|\gamma_m| \leq \frac{|B_1|}{2}, \quad n \in \mathbb{N},$$

and following the Proof of Theorem 3.1, we deduce that

$$2\sum_{m=1}^{\infty}|\gamma_m|r^m \leq 1 \quad \text{for} \quad |z| = r \leq (1+|B_1|)^{-1}.$$

Note that all sharp logarithmic coefficient's bounds for the functions in the class $\mathcal{C}(\psi)$ of normalized convex function are not available till now, see [1, Theorem 1, Sec 2]. But using the technique of differential subordination, we overcome this difficulty to achieve the Bohr radius for logarithmic power series for the Ma-Minda classes of convex functions in the following result.

Theorem 3.3. *Let ϕ be convex in $\mathbb{D}$ and suppose ψ be the convex solution of the differential equation*

$$\psi(z) + \frac{z\psi'(z)}{\psi(z)} = \phi(z). \tag{3.3}$$

If $f \in \mathcal{C}(\phi)$ with $\phi(z) = 1+B_1z+B_2z^2+\cdots$. Then the logarithmic coefficients of f satisfies

$$2\sum_{m=1}^{\infty} |\gamma_m| r^m \leq 1, \tag{3.4}$$

for

$$|z| = r \leq 1 - \frac{1}{\exp\left(\frac{2}{|B_1|}\right)},$$

The result is sharp for the case when the function f defined as $1+zf''(z)/f'(z) = \phi(z)$ is an extremal for the logarithmic coefficient's bounds.

Proof. Let us define $p(z) := zf'(z)/f(z)$. Since $f \in \mathcal{C}(\phi)$, therefore we have $1 + zf''(z)/f'(z) \prec \phi(z)$, which can be equivalently written as

$$p(z) + \frac{zp'(z)}{p(z)} = 1 + \frac{zf''(z)}{f'(z)} \prec \phi(z). \tag{3.5}$$

Since $\Re\phi(z) > 0$ and ϕ is convex in $\mathbb{D}$, therefore using [27, Theorem 3.2d, p. 86] the solution ψ of the differential equation (3.3) is analytic in $\mathbb{D}$ with $\Re\psi(z) > 0$ and has the following integral form given by

$$\psi(z) := h(z)\left(\int_0^z \frac{h(t)}{t}dt\right)^{-1} = 1 + \frac{B_1}{2}z + \frac{{B_1}^2 + 4B_2}{12}z^2 + \cdots, \tag{3.6}$$

where

$$h(z) = z\exp\int_0^z \frac{\phi(t)-1}{t}dt.$$

Since $\Re\psi(z) > 0$ and p satisfies the subordination (3.5), therefore using [27, Lemma 3.2e, p. 89] we conclude that ψ is univalent and $p \prec \psi$, where ψ is the best dominant. Thus we have obtained that $f \in \mathcal{C}(\phi)$ implies $zf'(z)/f(z) \prec \psi(z)$ and ψ is the best dominant, which is a univalent Carathéodory function. Now applying Theorem 3.1, the desired radius is achieved. □

Remark 3.2. If $f \in \mathcal{C}$. Then the logarithmic coefficients of f satisfies

$$2\sum_{m=1}^{\infty} |\gamma_m| r^m \leq 1,$$

for

$$|z| = r \leq 1 - 1/e.$$

and the radius is sharp for the function $f(z) = z/(1-z)$. The result was obtained by Bhowmik and Das in [5, Remark on Page no 741].

Now we estimate the Logarithmic coefficient's bounds for the functions in the class $\mathcal{C}(\psi)$ for the case $\psi(\mathbb{D})$ being convex, which generalize the result [1, Theorem 2, Sec 2] where sharp two initial Logarithmic coefficient's bounds and estimate for the third bound (γ_m, for $m = 1, 2$ and 3) were discussed.

Theorem 3.4. *(Convex analogue of* [1, Theorem 1, Sec 2]*) Let ϕ be convex in $\mathbb{D}$ and suppose ψ be the solution of the differential equation*

$$\psi(z) + \frac{z\psi'(z)}{\psi(z)} = \phi(z).$$

If $f \in \mathcal{C}(\phi)$ with $\phi(z) = 1+B_1z+B_2z^2+\cdots$. Then the logarithmic coefficients of f satisfies the inequalities:

(1) If ψ is convex. Then

$$|\gamma_m| \leq \frac{|B_1|}{4m}, \quad m \in \mathbb{N}, \tag{3.7}$$

$$\sum_{m=1}^{n} |\gamma_m|^2 \leq \frac{1}{4}\sum_{m=1}^{n} \frac{|c_m|^2}{m^2}, \quad n \in \mathbb{N},$$

and

$$\sum_{m=1}^{\infty} |\gamma_m|^2 \leq \frac{1}{4}\sum_{m=1}^{\infty} \frac{|c_m|^2}{m^2} \tag{3.8}$$

where c_m are the Taylor series coefficients of ψ (see Eq. (3.6)*).*

(2) If ψ is starlike with respect to 1*. Then*

$$|\gamma_m| \leq \frac{|B_1|}{4}, \quad m \in \mathbb{D}. \tag{3.9}$$

The inequalities (3.7)*,* (3.8) *and* (3.9) *are sharp for the case when the function f defined as $zf'(z)/f(z) = \psi(z)$ serves as an extremal for the logarithmic coefficient's bounds.*

Proof. Following the Proof of Theorem 3.3, and then applying the result [1, Theorem 1, Sec 2] completes the proof. □

In our next result, we show the application to the Janowski class, which covers many well-known classes. Here $\mathcal{C}[D, E] := \mathcal{C}((1 + Dz)/(1 + Ez)$.

Let us first consider the confluent and Gaussian hypergeometric functions, respectively as follows:

$$q(z) = \begin{cases} {}_2F_1\left(1 - \frac{D}{E}, 1, 2; \frac{Ez}{1+Ez}\right), & \text{if } E \neq 0; \\ {}_1F_1\left(1, 2; -Dz\right), & \text{if } E = 0. \end{cases} \tag{3.10}$$

We note that for $1 + D/E \geq 0$ and $-1 \leq E < 0$,

$$\min_{|z|=r} \Re\psi(z) = \psi(-r) = \frac{1}{q(-r)} > 0, \text{ where } \psi(z) = 1/q(z).$$

Corollary 3.5. *Let f belongs to $\mathcal{C}[D, E]$, where $-1 \leq E < D \leq 1$, and in addition q as defined in* (3.10) *be convex. Then the logarithmic coefficients of f satisfies* (3.4) *for*

$$|z| = r \leq \begin{cases} 1 - \frac{1}{\exp\left(\frac{2}{D-E}\right)}, & \text{if } E, D \neq 0; \\ \\ 1 - \frac{1}{\exp\left(\frac{2}{|E|}\right)}, & \text{if } E \neq 0, D = 0, \end{cases}$$

Proof. In Theorem 3.3, put $\phi(z) = (1+Dz)/(1+Ez)$. Then we have $\psi(z) := 1/q(z)$, where

$$q(z) = \begin{cases} \int_0^1 \left(\frac{1+Etz}{1+Ez}\right)^{\frac{D-E}{E}} dt, & \text{if } E \neq 0; \\ \int_0^1 e^{D(t-1)z} dt, & \text{if } E = 0, \end{cases}$$

which with some computation can be written explicitly as

$$\psi(z) = \begin{cases} \frac{Dz(1+Ez)^{\frac{D}{E}-1}}{-1+(1+Ez)^{\frac{D}{E}}} = 1 + \frac{D-E}{2}z + \frac{D^2-6DE+5E^2}{12}z^2 + \cdots, & \text{if } E, D \neq 0; \\ \frac{Ez}{(1+Ez)\log(1+Ez)} = 1 - \frac{E}{2}z + \frac{5E^2}{12}z^2 - \cdots, & \text{if } E \neq 0, D = 0; \\ \frac{Dze^{Dz}}{e^{Dz}-1}, & \text{if } E = 0, \end{cases}$$

and satisfies the differential Eq. (3.3). Hence, the result follows from Theorem 3.1. □

Now we have the result for the class of convex functions of order α using Corollary 3.5:

Corollary 3.6. *Let f belongs to $\mathcal{C}[1-2\alpha, -1]$, where $0 \leq \alpha < 1$. Then the logarithmic coefficients of f satisfies* (3.4) *for*

$$|z| = r \leq 1 - \exp(1/(\alpha - 1)), \text{ when } \alpha \neq 1/2$$

and

$$|z| = r \leq 1 - 1/\exp(2), \text{ when } \alpha = 1/2.$$

Corollary 3.7. *Let f belongs to $\mathcal{C}[D, 0]$. Then the logarithmic coefficients of f satisfies* (3.4) *for*

$$|z| = r \leq 1 - \exp(-\tfrac{2}{D}),$$

Proof. From the proof of Corollary 3.5, we obtain that

$$\psi(z) = Dze^{Dz}/(e^{Dz} - 1) = 1 + \frac{D}{2}z + \frac{D^2}{2}z^2 - \cdots,$$

when $\phi(z) = 1 + Dz$. Now with a little computation, we find that the function $l(z) = ze^z/(e^z - 1)$ is convex univalent in $\mathbb{D}$. Therefore, the function $\psi(z) = l(Dz)$ is also convex in $\mathbb{D}$ for each fixed $0 < D \leq 1$. □

Theorem 3.8. *Let ϕ be convex in $\mathbb{D}$ with $\Re\phi(z) > 0$, and suppose $f \in \mathcal{A}$ satisfies the differential subordination*

$$\frac{zf'(z)}{f(z)} + z\left(\frac{zf'(z)}{f(z)}\right)' \prec \phi(z) = 1 + B_1 z + B_2 z^2 + \cdots. \tag{3.11}$$

Then the logarithmic coefficients of f satisfies (3.4) *for*

$$|z| = r \leq 1 - \frac{1}{\exp\left(\frac{2}{|B_1|}\right)}.$$

The result is sharp for the case when the function f defined as $zf'(z)/f(z) = \psi(z)$ is an extremal for the logarithmic coefficient's bounds.

Proof. Let $p(z) = zf'(z)/f(z)$. Then the subordination (3.11) can equivalently be written as: $p(z) + zp'(z) \prec \phi(z)$. A simple calculation show that the analytic function

$$\psi(z) := (1/z)\int_0^z \phi(t)dt = 1 + \frac{B_1}{2}z + \frac{B_2}{3}z^2 + \cdots$$

satisfies

$$\psi(z) + z\psi'(z) = \phi(z).$$

Now from the Hallenbeck and Ruscheweyh result [27, Theorem 3.1b, p. 71], we have $p \prec \psi$, where ψ is the best dominant and also convex. Further, since $\Re\phi(z) > 0$, using the integral operator [27, Theorem 4.2a, p. 202] preserving functions with positive real part, we see that ψ is a Caratheódory function. Thus we have

$$\frac{zf'(z)}{f(z)} \prec \psi(z).$$

Now applying Theorem 3.1, the result follows. □

Corollary 3.9. *Suppose $f \in \mathcal{A}$ satisfies the differential subordination*

$$\frac{zf'(z)}{f(z)} + z\left(\frac{zf'(z)}{f(z)}\right)' \prec \frac{1+z}{1-z}.$$

Then the logarithmic coefficients of f satisfies (3.4) *for $|z| = r \leq (e-1)/e \approx 0.632121$,*

Corollary 3.10. *Suppose $f \in \mathcal{A}$ satisfies the differential subordination*

$$\frac{zf'(z)}{f(z)} + z\left(\frac{zf'(z)}{f(z)}\right)' \prec e^z.$$

Then the logarithmic coefficients of f satisfies (3.4) *for $|z| = r \leq (e^2 - 1)/e^2 \approx 0.864665$,*

Theorem 3.11. *Let ϕ be convex in $\mathbb{D}$ with $\Re\phi(z) > 0$, and suppose $f \in \mathcal{A}$ satisfies the differential subordination*

$$\frac{zf'(z)}{f(z)}\left(\frac{zf'(z)}{f(z)} + 2z\left(\frac{zf'(z)}{f(z)}\right)'\right) \prec \phi(z) = 1 + B_1 z + B_2 z^2 + \cdots, B_1 \neq 0. \tag{3.12}$$

Then the logarithmic coefficients of f satisfies (3.4) *for*

$$|z| = r \leq 1 - \frac{1}{\exp(\frac{4}{|B_1|})}.$$

The result is sharp for the case when the function f defined as $zf'(z)/f(z) = \sqrt{\psi(z)}$ is an extremal for the logarithmic coefficient's bounds.

Proof. Let $p(z) = zf'(z)/f(z)$. Then the subordination (3.12) can be equivalently written as:

$$p^2(z) + 2zp(z)p'(z) \prec \phi(z),$$

which using the change of variable $P(z) = p^2(z)$ becomes

$$P(z) + zP'(z) \prec \phi(z).$$

Now proceeding as in Theorem 3.8, we see that $p(z) \prec \sqrt{\psi(z)}$ and $\sqrt{\psi(z)}$ is the best dominant. Further, since $\Re\phi(z) > 0$, using [27, Theorem 4.2a, p. 202], we see that ψ is a Carathéodory function. Thus we have

$$\frac{zf'(z)}{f(z)} \prec \sqrt{\psi(z)} = 1 + \frac{B_1}{4}z + \left(\frac{B_2}{6} - \frac{B_1^2}{32}\right)z^2 + \cdots,$$

where $\psi(z) = \frac{1}{z}\int_0^z \phi(t)dt$. Now if we let $q(z) = 1 + z\psi''(z)/\psi'(z)$. Then

$$q(0) = 1 \quad \text{and} \quad \psi'(z)(q(z) + 1) = \phi'(z),$$

which further by logarithmic differentiation gives

$$\Psi(r, s; z) := q(z) + \frac{zq'(z)}{q(z) + 1} = 1 + \frac{z\phi''(z)}{\phi'(z)}. \tag{3.13}$$

Since the function Ψ satisfy the admissible conditions [27, 2.3-11, p. 35], using [27, Theorem 2.3i (i), p. 35] in (3.13), we see that ψ is convex. Let

$$h(z) = (\psi(z) - 1)/\psi'(0) \in \mathcal{A}.$$

Now it is easy to conclude that $h \in \mathcal{S}^*$, hence univalent. Finally, note that

$$|\arg\sqrt{\psi(z)}| = \frac{1}{2}|\arg\psi(z)| \leq \frac{\pi}{4},$$

which implies $\Re\sqrt{\psi(z)} > 0$. Thus we see that $\sqrt{\psi(z)}$ is a Carathéodory univalent convex function. Now applying Theorem 3.1, the result follows. □

Corollary 3.12. *Suppose $f \in \mathcal{A}$ satisfies the differential subordination*

$$\frac{zf'(z)}{f(z)}\left(\frac{zf'(z)}{f(z)} + 2z\left(\frac{zf'(z)}{f(z)}\right)'\right) \prec \frac{1 + (2\alpha - 1)z}{1 + z},\ \alpha \in [0, 1).$$

Then the logarithmic coefficients of f satisfies (3.4) *for $|z| = r \leq 1 - \exp\left(\frac{2}{\alpha - 1}\right)$.*

Corollary 3.13. *Suppose $f \in \mathcal{A}$ satisfies the differential subordination*

$$\frac{zf'(z)}{f(z)}\left(\frac{zf'(z)}{f(z)} + 2z\left(\frac{zf'(z)}{f(z)}\right)'\right) \prec 1 + \alpha z, \quad \alpha \in (0, 1].$$

Then the logarithmic coefficients of f satisfies (3.4) *for $|z| = r \leq 1 - 1/\exp(4/\alpha)$.*

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] E. A. Adegani, N. E. Cho, M. Jafari, Logarithmic Coefficients For Univalent Functions Defined by Subordination, Mathematics 2019, 7, 408

[2] R. M. Ali, R. W. Barnard and A. Yu. Solynin, A note on Bohr's phenomenon for power series, J. Math. Anal. Appl. **449** (2017), no. 1, 154–167.

[3] R. M. Ali, N. K. Jain and V. Ravichandran, Bohr radius for classes of analytic functions, Results Math. **74** (2019), no. 4, Paper No. 179, 13 pp.

[4] S. A. Alkhaleefah, I. R. Kayumov and S. Ponnusamy, On the Bohr inequality with a fixed zero coefficient, Proc. Amer. Math. Soc. **147** (2019), no. 12, 5263–5274. arXiv:1903.12646v1

[5] B. Bhowmik and N. Das, Bohr phenomenon for locally univalent functions and logarithmic power series, Comput. Methods Funct. Theory **19** (2019), no. 4, 729–745.

[6] H. P. Boas and D. Khavinson, Bohr's power series theorem in several variables, Proc. Amer. Math. Soc. **125** (1997), no. 10, 2975–2979.

[7] H. Bohr, A Theorem Concerning Power Series, Proc. London Math. Soc. (2) **13** (1914), 1–5.

[8] M. Darus, S. Hussain, M. Raja, J. Sokół, On a subclass of starlike functions, Results Math. **73** (2018), no. 1, Paper No. 22, 12 pp.

[9] I. R. Kayumov and S. Ponnusamy, Bohr inequality for odd analytic functions, Comput. Methods Funct. Theory **17** (2017), no. 4, 679–688.

[10] I. R. Kayumov and S. Ponnusamy, Improved version of Bohr's inequality, C. R. Math. Acad. Sci. Paris **356** (2018), no. 3, 272–277.

[11] I. R. Kayumov, S. Ponnusamy and N. Shakirov, Bohr radius for locally univalent harmonic mappings, Math. Nachr. 291(2018), 1757–1768.

[12] M.-S. Liu, S. Ponnusamy and J. Wang, Bohr's phenomenon for the classes of quasi-subordination and K-quasiregular harmonic mappings, Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Mat. RACSAM **114** (2020), no. 3, Paper No. 115, 15 pp.

[13] W. C. Ma and D. Minda, A unified treatment of some special classes of univalent functions, in *Proceedings of the Conference on Complex Analysis (Tianjin, 1992)*, 157–169, Conf. Proc. Lecture Notes Anal., I, Int. Press, Cambridge, MA.

[14] Y. A. Muhanna, Bohr's phenomenon in subordination and bounded harmonic classes, Complex Var. Elliptic Equ. **55** (2010), no. 11, 1071–107

[15] Y. A. Muhanna, R. M. Ali., Z. C. Ng and S. F. M. Hansi, Bohr radius for subordinating families of analytic functions and bounded harmonic mappings, J. Math. Anal. Appl. **420** (2014), no. 1, 124–136.

[16] Y. Abu Muhanna, R. M. Ali and S. Ponnusamy, On the Bohr inequality, in *Progress in approximation theory and applicable complex analysis*, 269–300, Springer Optim. Appl., 117, Springer, Cham.

[17] Y. Abu-Muhanna, R. M. Ali and S. K. Lee, The Bohr operator on analytic functions and sections, J. Math. Anal. Appl. **496** (2021), no. 2, Paper No. 124837, 11 pp.

[18] V. I. Paulsen, G. Popescu and D. Singh, On Bohr's inequality, Proc. London Math. Soc. (3) **85** (2002), no. 2, 493–512.

[19] K. Gangania, and S. S. Kumar, Bohr-Rogosinski phenomenon for $\mathcal{S}^*(\psi)$ and $\mathcal{C}(\psi)$. Mediterranean Journal of Mathematics (To Appear). https://arxiv.org/abs/2105.08684v1.

[20] H. Hamada, Bohr phenomenon for analytic functions subordinate to starlike or convex functions, J. Math. Anal. Appl., **499** (2021), no. 1,, Art. 125019.

[21] Y. Huang, M. S. Liu and S. Ponnusamy, Refined bohr-type inequalities with area measure for bounded analytic functions, Analysis and Mathematical Physics, **10** (2020), no. 4, 1–21.

[22] S. S. Kumar and G. Kamaljeet, On Geometrical Properties of Certain Analytic functions, Iran. J. Sci. Technol. Trans. A Sci. (2020), https://doi.org/10.1007/s40995-021-01116-1

[23] S. S. Kumar and G. Kamaljeet, On Certain Generalizations of $\mathcal{S}^*(\psi)$, (2020), arXiv:2007.06069v1
[24] Z. Liu and S. Ponnusamy, Bohr radius for subordination and K-quasiconformal harmonic mappings, Bull. Malays. Math. Sci. Soc. **42** (2019), no. 5, 2151–2168.
[25] M. S. Liu, Y.-M. Shang and J.-F. Xu, Bohr-type inequalities of analytic functions, J. Inequal. Appl. **2018**, Paper No. 345, 13 pp.
[26] I. M. Milin, Comments on the proof of the conjecture on logarithmic coefficients, in *The Bieberbach conjecture (West Lafayette, Ind., 1985)*, 109–112, Math. Surveys Monogr., 21, Amer. Math. Soc., Providence, RI.
[27] S. S. Miller and P. T. Mocanu, *Differential subordinations*, Monographs and Textbooks in Pure and Applied Mathematics, 225, Marcel Dekker, Inc., New York, 2000.
[28] S. Ponnusamy and T. Sugawa, Sharp inequalities for logarithmic coefficients and their applications, Bull. Sci. Math. **166** (2021), Paper No. 102931, 23 pp.
[29] O. Roth, A sharp inequality for the logarithmic coefficients of univalent functions, Proc. Amer. Math. Soc. **135** (2007), no. 7, 2051–2054.
[30] L. E. Schaubroeck, Subordination of planar harmonic functions, Complex Variables Theory Appl. **41** (2000), no. 2, 163–178.

Department of Applied Mathematics, Delhi Technological University, Delhi–110042, India

Email address: gangania.m1991@gmail.com

Booth-Encoded Karatsuba: A Novel Hardware-Efficient Multiplier

Riya JAIN[1], *Khushbu PAHWA*[2], *Neeta PANDEY*[1]

[1]Department of Electronics and Communication Engineering, Delhi Technological University, Bawana Road, Shahbad Daulatpur Village, Rohini, 110042 Delhi, India
[2]Department of Electrical Engineering, Delhi Technological University, Bawana Road, Shahbad Daulatpur Village, Rohini, 110042 Delhi, India

rj.riyajain26@gmail.com, khushbu16win@gmail.com, n66pandey@rediffmail.com

DOI: 10.15598/aeee.v19i3.4199

Article history: Received Mar 07, 2021; Revised Jun 17, 2021; Accepted Jul 14, 2021; Published Sep 30, 2021.
This is an open access article under the BY-CC license.

Abstract. *There is a recent boom being witnessed in emerging areas like IoMT (Internet of Medical Things), Artificial Intelligence for healthcare, and disaster management. These novel research frontiers are critical in terms of hardware and cannot afford to compromise accuracy or reliability. Multiplier, being one of the most heavily used components, becomes crucial in these applications. If optimized, multipliers can impact the overall performance of the system. Thus, in this paper, an attempt has been made to determine the potential of accurate multipliers while meeting minimal hardware requirements. In this paper, we propose a novel Booth-Encoded Karatsuba multiplier and provide its comparison with a Booth-Encoded Wallace tree multiplier. These architectures have been developed using two types of Booth encoding: Radix-4 and Radix-8 for 16-bit, 32-bit and 64-bit multiplications. The algorithm is designed to be parameterizable to different bit widths, thereby offering higher flexibility. The proposed mul- tiplier offers advantage of enhanced performance with significant reduction in hardware while negligibly trad- ing off the Power Delay Product (PDP). It has been observed that the performance of the proposed architecture increases with increasing multiplier size due to significant reduction in hardware and slight increase in PDP. All the architectures have been implemented in Verilog HDL using Xilinx Vivado Design Suite.*

Keywords

Accurate, Booth-encoding, Karatsuba, Wallace.

1. Introduction

Approximate multipliers [1], [2], [3], [4], [5], [6] and [7] have garnered a lot of attention from the researchers in the recent past. This is primarily due to the need to develop architectures for arithmetic computing that can reduce delay, power and hardware utilization, while relaxing the constraint on accuracy. However, there has been an upsurge in requirement to develop accurate hardware-efficient circuits for healthcare applications, resulting in emergence of novel technologies like Wireless Body Area Networks (WBAN) [8], and biomedical circuits for Brain Machine Interfaces [9]. Finite field arithmetic is another research direction that requires optimized hardware architectures, especially for coding theory and public-key cryptosystems [10], [11] and [12]. Complex field operations like exponentiation and inversion can be attained by iterative multiplication. Thus, it is necessary to design efficient multiplier. Being the core components of these systems, slight improvement in multiplication unit can offer significant performance enhancement of overall system.

Generally, multiplication operation can be bisected in two stages: (1) partial-products generation, and (2) addition of generated partial products. Research is currently underway to optimize the multiplication operation by optimizing either of the two stages. Booth encoding is being used as an effective technique to reduce the number of partial products generated in the multiplication process [13], [14] and [15] which in turn results in substantial reduction in area. Furthermore, for optimizing the second stage, Wallace tree multipliers are used as they are faster due to logarithmic depth and utilize lesser hardware by using fewest adders pos-

sible. Therefore, Booth and Wallace tree multipliers can be modified and hybridized together to obtain an optimized multiplier, namely Booth-Encoded Wallace tree multiplier. This hybrid architecture has thus been leveraged by several researchers [16], [17], [18], [19] and [20] to attain higher speed and area optimization. Hence, we have chosen the same over other architectures for comparison with our proposed multiplier.

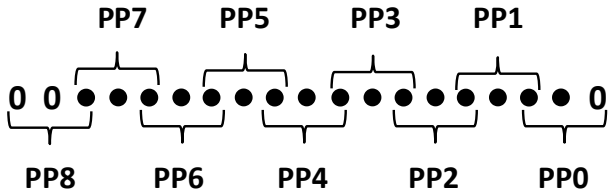


Fig. 1: Partial Products generation for 16-bit operand using Radix-4.

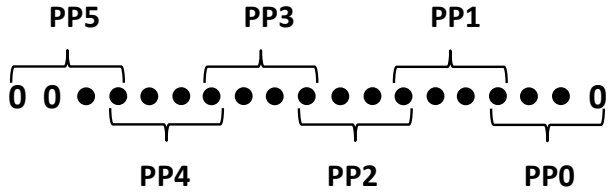


Fig. 2: Partial Products generation for 16-bit operand using Radix-8.

In Wallace tree multiplication, further optimization can be obtained by dividing an $N \times N$ multiplication in 4 small $\frac{N}{2} \times \frac{N}{2}$ multiplications which are processed in parallel [21]. The outputs of these 4 smaller size multipliers are then arranged suitably and added together using a ripple carry adder or carry look-ahead adder to obtain the final result. This helps in reducing the delay as 4 multiplications are done in parallel. However, the Karatsuba multiplier [7] uses 3 small multiplier for an $N \times N$ multiplication: two $\frac{N}{2} \times \frac{N}{2}$ multipliers and one $(\frac{N}{2}+1) \times (\frac{N}{2}+1)$ multiplier. Thus, Karatsuba multiplier requires one less multiplier compared to Wallace tree multiplier which helps in reducing the hardware drastically while causing meagre increase in delay and power due to complex structure. One less multiplier in Karatsuba algorithm is compensated by using a series of addition and subtraction operation. Therefore, the reduction in hardware becomes more significant as the size of multiplier increases. This is owed to the fact that hardware utilization increases exponentially with increase in multiplier size; however, the trend is linear for an adder. Therefore, in this paper, we have utilized the properties of Karatsuba algorithm and proposed a novel Booth-Encoded Karatsuba multiplier which is a hybrid of Booth and Karatsuba multiplier. The proposed multiplier uses one less multiplier compared to Booth-Encoded Wallace tree multiplier while individual multipliers use Booth encoding to reduce the number of partial products which is then employed with Wallace tree addition scheme.

The rest of the manuscript has been structured as follows: Sec. 2. puts forth the background of our research, focusing on Booth's algorithm, Wallace tree multiplier and how they can be hybridized to obtain an optimized Booth-Encoded Wallace tree multiplier. Section 3. elaborates on the working and implementation of proposed Booth-Encoded Karatsuba multiplier which is followed by the results in Sec. 4. Finally, the paper has been concluded in Sec. 5. based on all the findings from Sec. 4.

2. Background

2.1. Modified Booth's Algorithm

The modified Booth's algorithm [22] and [23] is a slightly modified and improved version of the Booth's algorithm. In the modified Booth algorithm, parallel encoding is performed as opposed to the serial encoding used in the originally proposed Booth algorithm. For Radix-R, encoding is obtained in following steps [24]:

- LSB of Multiplier (MR) is padded with a single bit 0.
- MR with padded 0 at LSB is partitioned into overlapping groups of x-bits where $x = 1 + \log_2 R$.
- Each block of x-bits is encoded to generate a single partial product. Encoding is done based on the Booth encoding table which differs for different Radix.
- Each partial product is then shifted s-bits to the left of its preceding neighbor, where $s = x - 1$.

For a N-bit multiplier, total of P partial products are generated, where:

$$P = \text{floor}\left(\frac{N+s}{s}\right). \tag{1}$$

All the partial products, after shifting and getting arranged suitably, are then added using appropriate addition scheme.

In this paper, Radix-4 and Radix-8 modified Booth's algorithms are being utilized. For Radix-4, partitioning of the MR is done using overlapping groups of 3-bits as shown in Fig. 1 and each block of 3-bits is encoded to generate a single partial product as per the

Tab. 1. For instance, PP0 in Fig. 1 is the the product of the encoding obtained from $\{B_{i+1},B_i,B_{i-1}\}$ and MD, where B_{i-1} is the 1'b0 appended to the right of the LSB, B_i is the 1^{st} bit (LSB), and B_i is the 2^{nd} bit. Finally, each partial product generated is shifted 2-bits left to its precedent before adding them. For an N-bit operand, the number of partial products generated is equal to floor($\frac{N+2}{2}$). Similarly, the generation of the partial products for Radix-8 is obtained by partitioning the multiplier into overlapping groups of 4-bits as shown in Fig. 2, and then each block of 4-bits is encoded as per the Tab. 2.

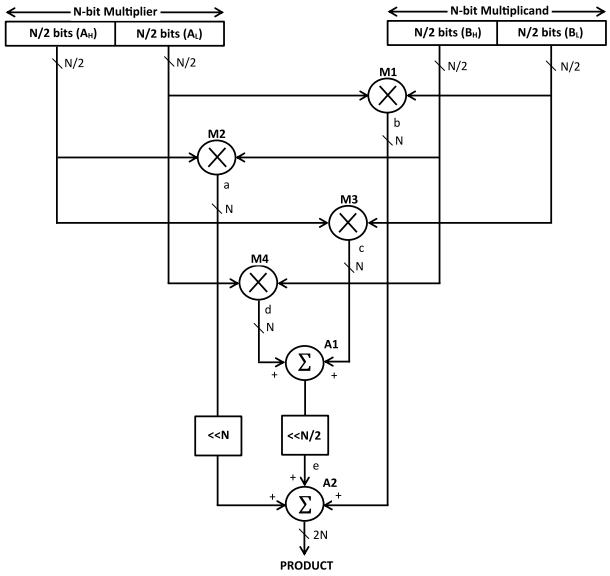


Fig. 3: Generalized Wallace Tree Architecture.[7]

Finally, before addition, all partial product is then shifted 3-bits left to its precedent. For an N-bit operand, the number of partial products generated are floor($\frac{N+3}{3}$). One factor to be noted here is that the MSB of the MR needs to be padded with at least an extra 0 bit. This is done to make the final sum a positive multiple of the Multiplicand (MD). This accounts for the appearance of 1 in the floor($\frac{N}{2}+1$) and floor($\frac{N}{3}+1$) expressions.

Thus, by means of using Booth encoding, the number of partial products generated can be reduced which in turn helps in reduction of hardware usage as well as delay.

Tab. 1: Radix-4 Booth Encoding.

B_{i+1}	B_i	B_{i-1}	Encoding
0	0	0	0× MD
0	0	1	1× MD
0	1	0	1× MD
0	1	1	2× MD
1	0	0	−2× MD
1	0	1	−1× MD
1	1	0	−1× MD
1	1	1	0× MD

Tab. 2: Radix-8 Booth Encoding.

B_{i+2}	B_{i+1}	B_i	B_{i-1}	Encoding
0	0	0	0	0× MD
0	0	0	1	1× MD
0	0	1	0	1× MD
0	0	1	1	2× MD
0	1	0	0	2× MD
0	1	0	1	3× MD
0	1	1	0	3× MD
0	1	1	1	4× MD
1	0	0	0	−4× MD
1	0	0	1	−3× MD
1	0	1	0	−3× MD
1	0	1	1	−2× MD
1	1	0	0	−2× MD
1	1	0	1	−1× MD
1	1	1	0	−1× MD
1	1	1	1	0× MD

2.2. Booth-Encoded Wallace Tree Multiplier

Several architectures have been proposed in the past with the aim of boosting the speed of the multiplier. One such architecture is the Wallace tree architecture which is an improved version of tree based multipliers [16] and [17]. The advantage offered by Wallace tree multiplier in terms of speed gets amplified when a higher multiplier units are constructed, generally 16-bit or higher. In Wallace tree architecture, the bits of all of the partial products in each column are added together by employing counters or compressors in parallel without carry propagation. A tree of Carry Save Adders (CSA) is thus used until a final 2 row matrix is generated. The final 2 rows are added using a ripple carry adder or carry look-ahead adder. The main advantage of Wallace tree architecture is the speed since the time complexity for the addition of N partial products is reduced from $O(N^2)$ to $O(\log N)$.

Furthermore, a higher order $N \times N$ bit multiplier unit can be further optimized by splitting it into smaller $\frac{N}{2} \times \frac{N}{2}$ bit multiplier units. Thus, an N-bit MD can be split into two $\frac{N}{2}$ bit words which are MD_H and MD_L. Similarly, the N-bit MR can be decomposed into MR_H and MR_L. Thus, the product for an N-bit by N-bit multiplication can be attained by employing four smaller $\frac{N}{2} \times \frac{N}{2}$ bit multiplier units. This can be better understood from Fig. 3. The products of the four multiplier units (M1, M2, M3, and M4) are given as:

$$a = MD_H \times MR_H, \tag{2}$$

$$b = MD_L \times MR_L, \tag{3}$$

$$c = MD_H \times MR_L, \tag{4}$$

$$d = MD_L \times MR_H. \tag{5}$$

©2021 ADVANCES IN ELECTRICAL AND ELECTRONIC ENGINEERING

Here, c and d are added together using A1 (adder) and then the sum is left shifted by $\frac{N}{2}$ bits to generate output e given as follows:

$$e = 2^{\frac{N}{2}}(c+d). \tag{6}$$

Now, a is also shifted left by N bits and added with e and b, using adder A2 to obtain the final product given by the following equation:

$$Product = 2^{N}a + 2^{\frac{N}{2}}(c+d) + b. \tag{7}$$

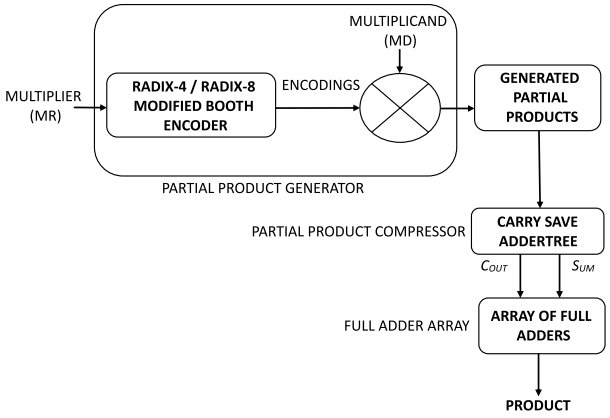


Fig. 4: Multiplier Architecture.

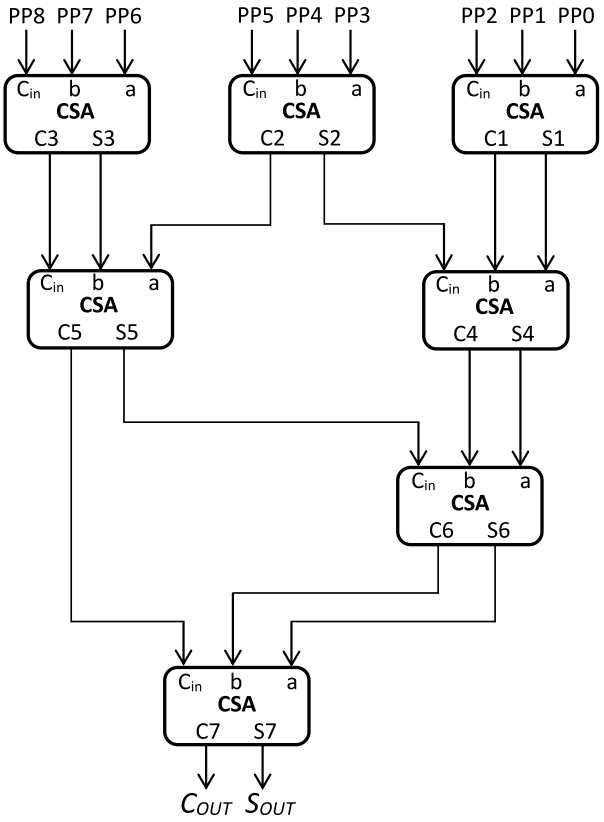


Fig. 5: CSA Tree Architecture to add 9 partial products.

Each of the 4 $\frac{N}{2}$ bit multipliers (denoted as M1, M2, M3, and M4 in Fig. 3) employ the processing steps as highlighted in Fig. 4. Various blocks involved in the multiplier architecture and their respective functionality will be explained further.

Partial Product Generator: The first processing block of the multiplier architecture, as depicted in Fig. 4, is the Partial Product Generator block that groups the MR as per the Fig. 1 or Fig. 2 depending on the Radix-4 or Radix-8 encoding. Furthermore, it generates the partial products using the encoding tables as described in Tab. 1 or Tab. 2 employing the steps enumerated in Subsec. 2.1.

Partial Product Compressor: The generated partial products are fed into the Partial Product Compressor block which is the second processing block. It employs a Wallace tree Architecture using 3:2 Carry Save Adders (CSA). For instance, the Fig. 5 represents the CSA tree architecture for the addition of 9 partial products obtained from Radix-4 Booth encoding of 16-bit multiplier. The CSA takes three inputs (a, b and cin) and generates two outputs, namely carry and sum. CSA(s) are used to improve on the delay ($O(\log N)$) and area ($O(N)$). The compression of partial products using CSA is much faster than their conventional addition. The C_{out} and S_{out} which represent the carry-out and sum generated by the last CSA block are passed onto the next block of the architecture i.e. Full Adder Array.

Full Adder Array: The Full Adder Array block takes in the C_{out} and S_{out} generated from the Partial Product Compressor block to generate the desired N-bit product by using an array of full adders, commonly known as ripple carry adder. This block can also be realized using carry look-ahead adder.

Thus, the Booth Encoded Wallace Tree Multiplier offers advantage of both Booth Encoding for reduction of Partial Products using Radix-4 and Radix-8 algorithms and faster accumulation of the generated partial products by reducing the levels of addition using the Wallace Tree structure. Thus, it finds use in various DSP applications like the computation of FFT for biomedical applications [25], and convolution operations [26].

3. Proposed Algorithm: Booth-Encoded Karatsuba Multiplier

The Karatsuba multiplication algorithm was proposed in 1962 by Karatsuba and Ofman [27]. It was the first attempt at finding efficient techniques for multiplication of large integers. The naive approach of multi-

plying two numbers has a time complexity of $O(N^2)$ for a N-bit multiplication. However, the Karatsuba algorithm leverages the divide and conquer approach and has a time complexity of $O(N^{\log_2 3})$ which is a significant improvement over the naive algorithm. The Karatsuba Algorithm can be employed in several applications that aim at achieving enhanced performance and efficient hardware accelerators, such as biomedical devices to address healthcare issues, human-computer interfaces for human assistance and cryptographic applications like RSA and DSA [28], [29] and [28]. These applications require high accuracy and demand for compact hardware designing. Recently, an approximate Karatsuba multiplier was proposed [7] in the literature which focused on the error-resilient applications. However, in this paper, we intend to introduce an accurate multiplier which integrates the Karatsuba algorithm with Booth-Encoded multiplier in order to achieve an optimized multiplier for applications where accuracy is critical and cannot be compromised.

The flow of the algorithm has been depicted in Fig. 6. In order to multiply two N-bit numbers MD and MR, the Karatsuba algorithm can be understood as described below. Let subscript H represent the upper half bits of (MD/MR) and subscript L represent the lower half bits of (MD/MR), then

$$a = MD_H \times MR_H, \tag{8}$$

$$b = MD_L \times MR_L, \tag{9}$$

At the same time,

$$c = MD_H + MD_L, \tag{10}$$

$$d = MR_H + MR_L, \tag{11}$$

Now, a and b are subtracted from the product of c and d. It can be computed as follows:

$$\begin{aligned} e &= (c \times d) - (a + b) = \\ &= (MD_H + MD_L)(MR_H + MR_L) + \\ &-(MD_H \times MR_H)(MD_L \times MR_L), \end{aligned} \tag{12}$$

After simplifying Eq. (12) we get,

$$e = MD_H MR_L + MD_L MR_H. \tag{13}$$

Thus, final product can be written as:

$$Product = 2^N a + 2^{\frac{N}{2}} e + b. \tag{14}$$

From the above equations, it can be seen that for N-bit multiplication, Karatsuba multiplier requires two $\frac{N}{2}$ bit multipliers and one $\frac{N}{2} + 1$ bit multiplier. Therefore, a total of 3 small size multipliers are required along with a series of addition and subtraction operations. In comparison to the Wallace tree multiplier as described in Subsec. 2.2. which utilises 4 small multipliers (denoted by M1, M2, M3, and M4), the Karatsuba multiplier utilises only 3 such multipliers units (M1, M2, and M3), as shown in Fig. 6. Thus, the Karatsuba multiplier performs the same N-bit multiplication as the Wallace Tree multiplier utilizing one lesser multiplier.

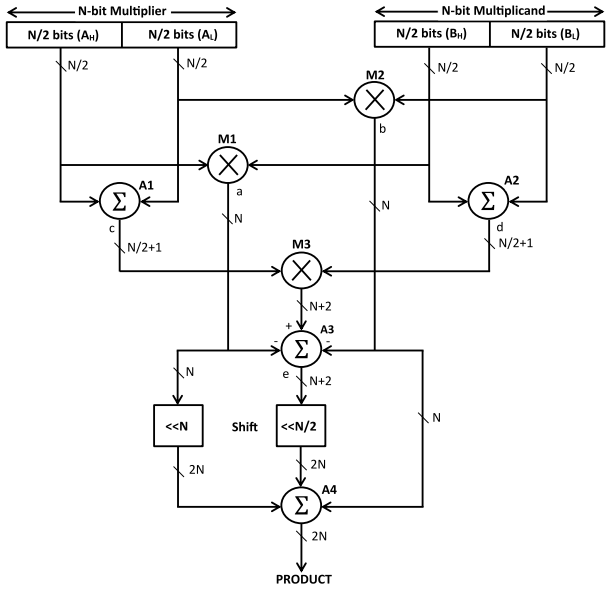


Fig. 6: Karatsuba Multiplier Architecture [7].

To further enhance the performance of the existing Karatsuba multiplier, we have proposed the hybridized architecture in which all the smaller size multipliers of Karatsuba algorithm are realized to exploit the advantages of reduced number of partial products obtained from modified Booth algorithm and the fast, parallel addition offered by Wallace Tree structure. Thus, to overcome the drawback of a conventional multiplier in which $N \times N$ multiplication results in N partial products, Booth encoding with Radix-4 and Radix-8 has been used which greatly reduces the number of partial products generated. Furthermore, these generated partial products are then added using Wallace tree addition scheme. The multiplier architecture used for BoothEncoded Karatsuba algorithm is same as described in Fig. 4. Working of multiplier architecture with respect to Booth-Encoded Karatsuba algorithm will be described below.

Partial Product Generator: The first processing block in Fig. 4 is the Partial Product Generator which uses Radix-4 or Radix-8 Booth encoding to generate partial products. For an N-bit multiplication, the proposed algorithm requires two $\frac{N}{2}$ bit multipliers and one $\frac{N}{2} + 1$ bit multiplier. In order to develop a 32-bit mul-

Tab. 3: Comparison of Total Hardware Utilization.

Size	Encoding used	Structure	Hardware	Percentage decrease
64-bit × 64-bit	Radix-4	Booth Karatsuba	2478	18.59 %
		Booth Wallace	3044	
	Radix-8	Booth Karatsuba	1815	13.57 %
		Booth Wallace	2100	
32-bit × 32-bit	Radix-4	Booth Karatsuba	777	14.05 %
		Booth Wallace	904	
	Radix-8	Booth Karatsuba	576	10.00 %
		Booth Wallace	640	
16-bit × 16-bit	Radix-4	Booth Karatsuba	252	5.97 %
		Booth Wallace	268	
	Radix-8	Booth Karatsuba	192	−14.28 %
		Booth Wallace	168	

Tab. 4: Comparison of PDP.

Size	Encoding used	Structure	Power	Delay	PDP	Percentage increase
64-bit × 64-bit	Radix-4	Booth Karatsuba	0.715	36.36	25.99	1.86 %
		Booth Wallace	0.713	35.80	25.52	
	Radix-8	Booth Karatsuba	0.719	37.40	26.89	3.73 %
		Booth Wallace	0.715	36.25	25.92	
32-bit × 32-bit	Radix-4	Booth Karatsuba	0.636	14.41	9.16	2.57 %
		Booth Wallace	0.635	14.07	8.93	
	Radix-8	Booth Karatsuba	0.639	15.23	9.73	6.45 %
		Booth Wallace	0.638	14.33	9.14	
16-bit × 16-bit	Radix-4	Booth Karatsuba	0.612	10.39	6.36	3.08 %
		Booth Wallace	0.611	10.10	6.17	
	Radix-8	Booth Karatsuba	0.612	11.52	7.05	11.46 %
		Booth Wallace	0.610	10.37	6.32	

tiplier using Radix-4 Booth encoding, two 16-bit multipliers and one 17-bit multiplier are used. For the 16-bit and 17-bit multipliers the number of partial products generated would be 9 using the Eq. (1). For the 32-bit multiplier using Radix-8 Booth encoding, the 16-bit and 17-bit multipliers would result in the generation of 6 partial products. Similarly, in order to develop a 64-bit multiplier, two 32-bit multipliers and one 33-bit multiplier are used.

Partial Product Compressor: The partial products generated in the Partial Product Generator block are then added and compressed to generate two rows of C_{out} and S_{out} using the CSA tree as described in Fig. 5.

Full Adder Array: Finally, an array of full adders is used to generate the product using C_{out} and S_{out} obtained from the Partial Product Compressor block.

4. Results

Architectures of the proposed multiplier and BoothEncoded Wallace tree multiplier [16], as discussed in Sec. 3. and Subsec. 2.2. respectively, have been realized for 16-, 32- and 64-bit multiplication using Radix-4 as well as Radix-8 Booth encoding. All the architectures have been implemented in Verilog HDL using Xilinx Vivado Design Suite. They are then evaluated in terms of hardware utilization and PDP, as discussed in this section.

Hardware Utilization: Hardware consumption for all the architectures has been compared in terms of the total number of Full Adders (FA) required. The percentage decrease in hardware of proposed architecture with respect to Booth-Encoded Wallace tree multiplier [16] have been recorded for all architectures in Tab. 3. Following observations can be derived from Tab. 3:

- For 16-, 32- and 64-bit multiplication, Radix-4 encoding, the proposed Booth-Encoded Karatsuba multiplier utilizes less hardware compared to Booth-Encoded Wallace tree multiplier.
- For 32- and 64-bit multiplication, Radix-8 encoding, the proposed Booth-Encoded Karatsuba multiplier utilizes less hardware compared to Booth Encoded Wallace tree multiplier.
- There is one outlier: for 16-bit multiplication, Radix-8 encoding, the proposed Booth-Encoded Karatsuba multiplier utilizes more hardware compared to Booth-Encoded Wallace tree multiplier. The reason for above discrepancy is that the hardware addition due to extra addition/subtraction

operations overshadows the hardware savings due to 1 less multiplier.

- Rate of decrease in hardware utilization from Booth-Encoded Wallace tree multiplier to Booth-Encoded Karatsuba multiplier increases with increase in size. Thus, system performance, in terms of hardware utilization, increases with increase in size.

- For better visualization, graphical representation has been added in Fig. 7 and Fig. 8. It can be seen that plot corresponding to Booth-Encoded Karatsuba multiplier always lies below the plot corresponding to Booth-Encoded Wallace tree multiplier - for both Radix-4 and Radix-8. Moreover, the gap between two plots keeps increasing with increasing multiplier size which shows that improvement in terms of hardware increases with increasing multiplier size.

Power Delay Product (PDP): The proposed Booth-Encoded Karatsuba architecture has also been compared with the Booth-Encoded Wallace [16] architecture in terms of the PDP. The results are summarized in the Tab. 4 from where certain observations can be made:

- For all the architectures, the proposed Booth-Encoded Karatsuba multiplier has slightly more PDP compared to Booth-Encoded Wallace tree multiplier.

- Lowest and highest PDP is seen in case of Booth-Encoded Wallace tree multiplier with Radix-4 encoding and Booth-Encoded Karatsuba multiplier with Radix-8 encoding, respectively. This trend is consistent for 16-bit, 32-bit as well as 64-bit multiplication.

- Percentage increase in PDP from Booth-Encoded Wallace tree multiplier to Booth-Encoded Karatsuba multiplier decreases with increase in size. Thus, system performance, in terms of PDP, increases with increase in size.

- For better visualization, graphical representation has been added in Fig. 9 and Fig. 10. It can be seen that plot corresponding to Booth-Encoded Karatsuba multiplier is always close to the plot corresponding to Booth-Encoded Wallace tree multiplier- for both Radix-4 and Radix-8. Thus, it is evident that there is no significant degradation in terms of PDP with varying multiplier size.

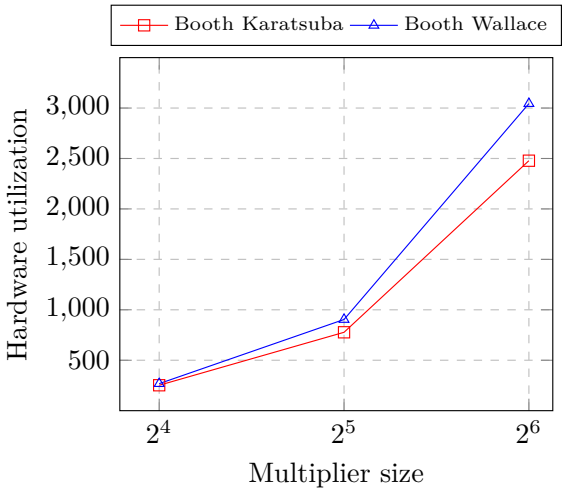


Fig. 7: Hardware Utilization varying with multiplier size for Radix-4 encoding.

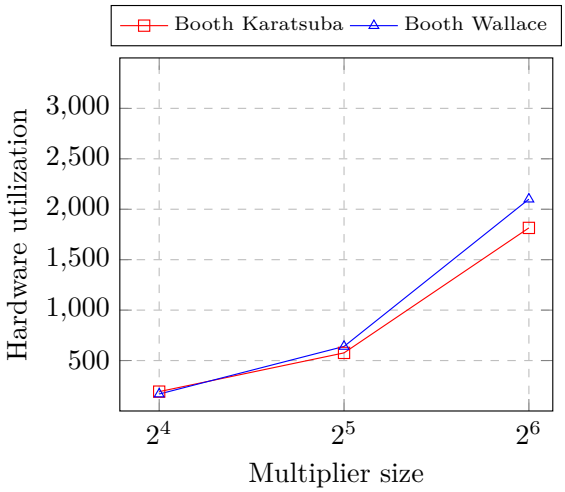


Fig. 8: Hardware Utilization varying with multiplier size for Radix-8 encoding.

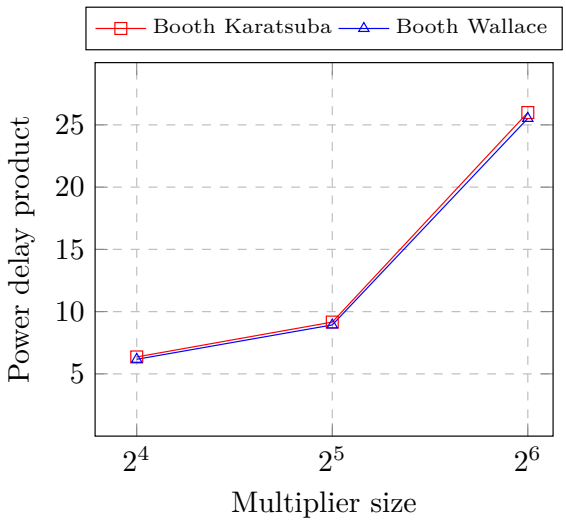


Fig. 9: PDP varying with multiplier size for Radix-4 encoding.

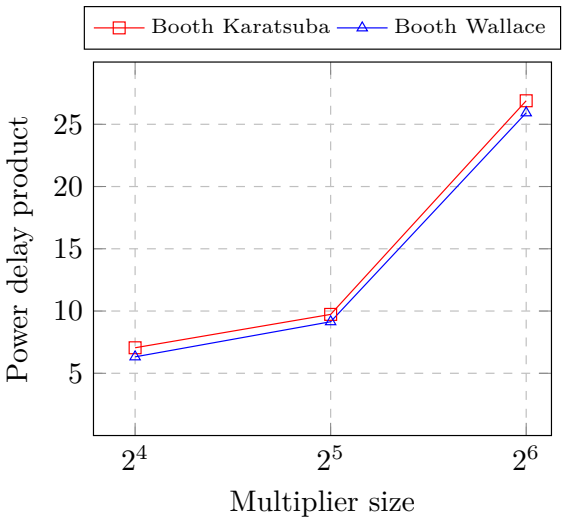


Fig. 10: PDP varying with multiplier size for Radix-8 encoding.

Hence, on the basis of above observations, we can state that there is an overall performance improvement of the proposed algorithm for 16-, 32- and 64 bit multipliers. The system's performance improvement increases with increasing size due to significant reduction in hardware utilization with slight increase in PDP. This can be accounted by the fact that percentage decrease in hardware increases with increase in multiplier size. At the same time, percentage increase in PDP decreases with increase in multiplier size. One more thing to be noted here is that when comparing different architectures with respect to encoding, Radix-4 architectures shows better performance compared to Radix-8 architectures.

5. Conclusion

An accurate multiplier based on the hybridized architecture of modified Booth encoding and Karatsuba multiplication algorithm was proposed in this paper. It was compared with an existing accurate BoothEncoded Wallace tree multiplier [16]. For 64-bit and 32-bit multiplier, there is an average improvement of 16.08 % and 12.025 % in terms of hardware while slight increase of 2.79 % and 4.51 % of PDP respectively has been encountered. Similar trend was seen for 16-bit Radix-4 multipliers. However, there is an exception seen for 16-bit Radix-8 which does not show any improvement in terms of hardware and PDP compared to the Booth-Encoded Wallace tree multiplier. Thus, on the basis of results encapsulated in previous section, it can be concluded that with increasing multiplier size, our proposed algorithm shows significant improvement. In future, the presented work can be extended by hybridizing Karatsuba multiplication method with other existing multipliers to optimize it further and leverage it in different applications. Furthermore, it can be extended to approximate computing operations to be employed in error-resilient applications.

Author Contributions

R.J. conceived the presented idea. R.J. encouraged K.P. to investigate the literature and formulate the plan for the proposed architecture. Both R.J. and K.P. developed the methodology and implemented the algorithms. R.J. performed data curation while K.P. wrote the original draft. Both R.J. and K.P. contributed to the final version of the manuscript. N.P. administered and supervised the project and contributed towards the modification of the final layout. All authors discussed the results and contributed to the final manuscript.

References

[1] GOSWAMI, S. S. P., B. PAUL, S. DUTT and G. TRIVEDI. Comparative Review of Approximate Multipliers. In: *30th International Conference Radioelektronika (RADIOELEKTRONIKA)*. Bratislava: IEEE, 2020 pp. 1–6. IBSN 978-1-7281-6469-4. DOI: 10.1109/RADIOELEKTRONIKA49387.2020.9092370.

[2] LIU, W., L. QIAN, C. WANG, H. JIANG, J. HAN and F. LOMBARDI. Design of Approximate Radix-4 Booth Multipliers for Error-Tolerant Computing. *IEEE Transactions on Computers*. 2017, vol. 66, iss. 8, pp. 1435–1441. ISSN 1557-9956. DOI: 10.1109/TC.2017.2672976.

[3] JIANG, H., C. LIU, L. LIU, F. LOMBARDI and J. HAN. A Review, Classification, and Comparative Evaluation of Approximate Arithmetic Circuits. *ACM Journal on Emerging Technologies in Computing Systems*. 2017, vol. 13, iss. 4, pp. 1–34. ISSN 1550-4832. DOI: 10.1145/3094124.

[4] MASADEH, M., O. HASAN and S. TAHAR. Comparative Study of Approximate Multipliers. In: *Proceedings of the 2018 on Great Lakes Symposium on VLSI (GLSVLSI)*. New York: ACM, 2018, pp. 415–418. ISBN 978-1-4503-5724-1. DOI: 10.1145/3194554.3194626.

[5] JIANG, H., C. LIU, N. MAHESHWARI, F. LOMBARDI and J. HAN. A comparative evaluation of approximate multipliers. In: *IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*. Beijing: IEEE, 2016. ISBN 978-1-4503-4330-5. DOI: 10.1145/2950067.2950068.

[6] YAMAMOTO, T., I. TANIGUCHI, H. TOMIYAMA, S. YAMASHITA and Y. HARA-AZUMI. A systematic methodology for design and analysis of approximate array multipliers. In: *IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*. Jeju: IEEE, 2016 pp. 352–354. ISBN 978-1-5090-1570-2. DOI: 10.1109/APCCAS.2016.7803973.

[7] JAIN, R. and N. PANDEY. Approximate Karatsuba multiplier for error-resilient applications. *AEU - International Journal of Electronics and Communications*. 2021, vol. 130, iss. 1, pp. 1434–8411. ISSN 1434-8411. DOI: 10.1016/j.aeue.2020.153579.

[8] LIU, X., Y. ZHENG, M. W. PHYU, F. N. ENDRU, V. NAVANEETHAN and B. ZHAO. An Ultra-Low Power ECG Acquisition and Monitoring ASIC System for WBAN Applications. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*. 2012, vol. 2, iss. 1, pp. 60–70. ISSN 2156-3365. DOI: 10.1109/JETCAS.2012.2187707.

[9] CHEN, Y., E. YAO and A. BASU. A 128-Channel Extreme Learning Machine-Based Neural Decoder for Brain Machine Interfaces. *IEEE Transactions on Biomedical Circuits and Systems*. 2016, vol. 10, iss. 3, pp. 679–692. ISSN 1940-9990. DOI: 10.1109/TBCAS.2015.2483618.

[10] SHAIK, N. B. *Novel Implementation of Finite Field Multipliers over GF(2m) for Emerging Cryptographic Applications*. Dayton, 2017. Master thesis. Wright State University. Supervisor Jiafeng Xie, Ph.D.

[11] GAUBATZ, G. and B. SUNAR. Robust finite field arithmetic for fault-tolerant public-key cryptography. In: *Proceedings of the Third international conference on Fault Diagnosis and Tolerance in Cryptography (FDTC'06)*. Berlin: Springer, 2006, pp. 196–210. ISBN 978-3-540-46251-4. DOI: 10.1007/11889700_18.

[12] SAVAS, E. and C. K. KOC. Finite field arithmetic for cryptography. *IEEE Circuits and Systems Magazine*. 2010, vol. 10, iss. 2, pp. 40–56. ISSN 1558-0830. DOI: 10.1109/mcas.2010.936785.

[13] WARIS, H., C. WANG and W. LIU. Hybrid Low Radix Encoding-Based Approximate Booth Multipliers. *IEEE Transactions on Circuits and Systems II: Express Briefs*. 2020, vol. 67, iss. 12, pp. 3367–3371. ISSN 1558-3791. DOI: 10.1109/TCSII.2020.2975094.

[14] HE, Y. and C. CHANG. A New Redundant Binary Booth Encoding for Fast 2^n-Bit Multiplier Design. *IEEE Transactions on Circuits and Systems I: Regular Papers*. 2009, vol. 56, iss. 6, pp. 1192–1201. ISSN 1558-0806. DOI: 10.1109/TCSI.2008.2008503.

[15] KUANG, S.-R., J.-P. WANG and C.-Y. GUO. Modified Booth Multipliers With a Regular Partial Product Array. *IEEE Transactions on Circuits and Systems II: Express Briefs*. 2009, vol. 56, iss. 5, pp. 404–408. ISSN 1558-3791. DOI: 10.1109/TCSII.2009.2019334.

[16] ASIF, S., and Y. KONG. Performance analysis of Wallace and radix-4 Booth-Wallace multipliers. In: *Electronic System Level Synthesis Conference (ESLsyn)*. San Francisco: IEEE, 2015, pp. 17–22. ISBN 979-1-0922-7912-2.

[17] RAO, M. J. and S. DUBEY. A high speed and area efficient Booth recorded Wallace tree multiplier for fast arithmetic circuits. In: *Asia Pacific Conference on Postgraduate Research in Microelectronics and Electronics*. Hyderabad: IEEE, 2012, pp. 220–223. ISBN 978-1-4673-5067-9. DOI: 10.1109/PrimeAsia.2012.6458658.

[18] SHARMA, K., N. and S. RAVI. Modified Booth Multiplier using Wallace Structure and Efficient Carry Select Adder. *International Journal of Computer Applications*. 2013, vol. 68, no. 13, pp. 39–42. ISSN 0975-8887. DOI: 10.5120/11643-7130.

[19] SUREKA, N., R. PORSELVI and K. KUMUTHAPRIYA. An efficient high speed Wallace tree multiplier. In: *International Conference on Information Communication and Embedded Systems (ICICES)*. Chennai: IEEE, 2013, pp. 1023–1026. ISBN 978-1-4673-5788-3. DOI: 10.1109/ICICES.2013.6508192.

[20] LAKSHMANAN, T, M. OTHMAN and M. A. M. ALI. High performance parallel multiplier using Wallace-Booth algorithm. In: *Proceedings of the 9th International Conference on Neural Information Processing. Computational Intelligence for the E-Age (IEEE Cat. No.02EX575)(CONIP '02)*. Penang: IEEE, 2002, pp. 433–436. ISBN 0-7803-7578-5. DOI: 10.1109/SMELEC.2002.1217859.

[21] BHARADWAJ K., P. S. MANE and J. HENKEL. Power- and area-efficient Approximate Wallace Tree Multiplier for error-resilient systems. In: *Fifteenth International Symposium on Quality Electronic Design*. Santa Clara: IEEE, 2014, pp. 263–269. ISBN 978-1-4799-3946-6. DOI: 10.1109/ISQED.2014.6783335.

[22] MACSORLEY, O. L. High-Speed Arithmetic in Binary Computers. *Proceedings of the IRE*. 1961, vol. 49, iss. 1 pp. 67–91. ISSN 2162-6634. DOI: 10.1109/JRPROC.1961.287779.

[23] KAUR, N. and R. K. PATIAL. Implementation of Modified Booth Multiplier using Pipeline Technique on FPGA. *International Journal of Computer Applications*. 1913, vol. 68, iss. 16 pp. 38–41. ISSN 0975-8887. DOI: 10.5120/11666-7261.

[24] BEWICK, G. and M. J. FLYNN. *Binary Multiplication using Partially Redundant Multiples*. Stanford: Stanford University. 1992.

[25] AJAY, A. and M. REGEENA. VLSI Implementation of an Improved Multiplier for FFT Computation in Biomedical Applications. In: *IEEE Computer Society Annual Symposium on VLSI*. Montpellier: IEEE, 2015, pp. 68–73. ISBN 978-1-4799-8719-1. DOI: 10.1109/ISVLSI.2015.104

[26] MOHANTY, J. P., S. and R. DAS and S. K. PANDA. Design and Simulation of Convolution using Booth Encoded Wallace Tree Multiplier. *IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)*. 2016, pp. 42–46. ISSN 2278-2834.

[27] KARATSUBA, A. and Y. OFMAN. Multiplication of Multidigit Numbers on Automata. *Doklady Akadademii Nauk SSSR*. 1962, vol. 145, iss. 2, pp. 293–294.

[28] KASHIF, M., I. CICEK and M. IMRAN. A Hardware Efficient Elliptic Curve Accelerator for FPGA Based Cryptographic Applications. In: *11th International Conference on Electrical and Electronics Engineering (ELECO)*. Bursa: IEEE, 2019, pp. 362–366. ISBN 978-605-01-1275-7. DOI: 10.23919/ELECO47770.2019.8990437.

[29] CHOW, G. C. T., K. EGURO, W. LUK and P. LEONG. A Karatsuba-Based Montgomery Multiplier. In: *International Conference on Field Programmable Logic and Applications*. Milan: IEEE, 2010, pp. 434–437. ISBN 978-1-4244-7843-9. DOI: 10.1109/FPL.2010.89.

[30] MANIKANDAN, S. and C. PALANISAMY. Design of an Efficient Binary Vedic Multiplier for High Speed Applications using Vedic Mathematics with Bit Reduction Technique. *Circuits and Systems*. 2016, vol. 7, no. 9, pp. 2593–2602. ISSN 2153-1293. DOI: 10.4236/cs.2016.79224.

About Authors

Riya JAIN received her B.Tech in the discipline Electronics and Communication Engineering from Delhi Technological University (DTU) in 2020. She is a research enthusiast who aims to contribute in the field of Electronics and Communication in coming future. Her areas of interest are: Digital Very Large Scale Integration (VLSI) design, Digital Design, and Computer Architecture. She has been contributing to the field of Electronics since 2016 and has been rigorously involved in research since 2018. Being a novice researcher, she is still exploring multiple domains in the field of Electronics and is motivated enough for further contribution to technological advancement in the same.

Khushbu PAHWA received her B.Tech from Delhi Technological University (DTU) in 2020 in the field of Electrical and Electronics Engineering (Gold Medallist). Her research interests span across the broad domain of Very Large Scale Integration (VLSI) (Digital and Analog Circuits), Internet of Things (IoT), Artificial Intelligence, and Wireless Sensor Networks. She has gained valuable experience in these domains through her research internships and publications. She aspires to work on hardware optimization which can be leveraged to serve as the workhorse for future complex Deep Learning Models for deployment at the edge.

Neeta PANDEY received her M.E. in Microelectronics from Birla Institute of Technology and Sciences, Pilani in 1991 and Ph.D. from Guru Gobind Singh Indraprastha University, Delhi in 2009. She has served in Central Electronics Engineering Research Institute, Pilani, Indian Institute of Technology, Delhi, Priyadarshini College of Computer Science, Noida and Bharati Vidyapeeth's College of Engineering, Delhi in various capacities. At present, she is a professor in the ECE department, Delhi Technological University. Her teaching and research interests include analog and digital VLSI design. A life member of Indian Society for Technical Education (ISTE), and senior member of Institute of Electrical and Electronics Engineers (IEEE), USA, she has coauthored over 100 papers in international, national journals of repute and conferences.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/355650259

Centellosides: pharmaceutical applications and production enhancement strategies

Article *in* Plant Cell Tissue and Organ Culture · October 2021

DOI: 10.1007/s11240-021-02187-z

CITATIONS
0

READS
95

3 authors:



Ritika Luthra
Delhi Technological University
5 PUBLICATIONS **10** CITATIONS
SEE PROFILE



Garvit Satija
Jamia Hamdard University
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE



Arpita Roy
Sharda University
104 PUBLICATIONS **567** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Antibacterial studies of medicinal plant View project

Medicinal Plants Research View project

All content following this page was uploaded by Ritika Luthra on 26 October 2021.

The user has requested enhancement of the downloaded file.

Plant Cell, Tissue and Organ Culture (PCTOC)
https://doi.org/10.1007/s11240-021-02187-z

REVIEW

Centellosides: pharmaceutical applications and production enhancement strategies

Ritika Luthra[1] · **Garvit Satija**[2] · **Arpita Roy**[3]

Received: 5 August 2021 / Accepted: 18 October 2021
© The Author(s), under exclusive licence to Springer Nature B.V. 2021

Abstract

Demand for naturally-derived compounds is growing steadily, owing to a broad spectrum of medicinal benefits posed by them. Triterpenoid centellosides, obtained from *Centella asiatica*, is being utilized as an important component in a variety of products such as pharmaceutics. Centellosides possess various therapeutic properties like antidiabetic, anti-inflammatory, antioxidant, anticancer, wound healing, neuroprotective properties, etc. However, *C. asiatica* is reportedly undergoing rapid depletion and is endangered due to the increasing demand of this plant for its secondary metabolites. Apart from this cultivation of *C. asiatica* needs a long time period, along with extensive labour for controlling the quality despite the unfavourable environmental conditions. Therefore, a substitute method is required for its cultivation and improved production of centellosides. For biosynthesis of centellosides at higher levels, various biotechnological methods have now been established, such as shoot culture, callus and cell suspension culture and hairy root culture. Further, many more strategies like use of elicitors and precursors for enhanced production as well as bioreactors for commercial scaling have also been incorporated. This review thus provides an understanding of biological applications of centellosides, biotechnological approaches as well as biotechnological improvements for its increased production. Information about various biotechnological approaches has also been collected after an extensive literature study.

Key message

Centellosides are important bioactive compounds from Centella asiatica which possess various therapeutic properties like antidiabetic, anti-inflammatory, antioxidant, anticancer, wound healing, neuroprotective properties, etc. For biosynthesis of centellosides at higher levels, various biotechnological methods have been established, such as shoot culture, callus and cell suspension culture, and hairy root culture. This review provides an understanding of biological applications of centellosides, biotechnological approaches as well as biotechnological improvements for its increased production.

Keywords Centellosides · Pharmaceutical application · Extraction · Biotechnological approaches · Bioreactors

Introduction

Several secondary metabolites are produced by medicinal herbs which play an essential part in their defence mechanisms as well as possess various biological applications (Roy et al. 2018a, b; Roy and Bharadvaja 2018a, b; Roy 2021). Centellosides, derived from *Centella asiatica* of the Apiaceae family, native to Southeast Asian countries, are one of these essential bioactive triterpenoids (Roy et al. 2016, 2017a, 2017b). This medicinal plant is widely distributed along warmer areas across the globe, including India, Indonesia, Malaysia, Sri Lanka, portions of China, Mexico, Columbia, South Africa, Madagascar etc. (Gray et al. 2018; Katiyar et al. 2021). It is found to grow well in

Communicated by Patricia Marconi.

This article belongs to the Topical Collection: Secondary Metabolites and Medicinal Plant Biotechnology.

✉ Arpita Roy
arbt2014@gmail.com

[1] Department of Biotechnology, Delhi Technological University, Delhi, India

[2] School of Pharmaceutical Education and Research, Jamia Hamdard University, New Delhi, India

[3] Department of Biotechnology, School of Engineering & Technology, Sharda University, Greater Noida, India

Published online: 26 October 2021

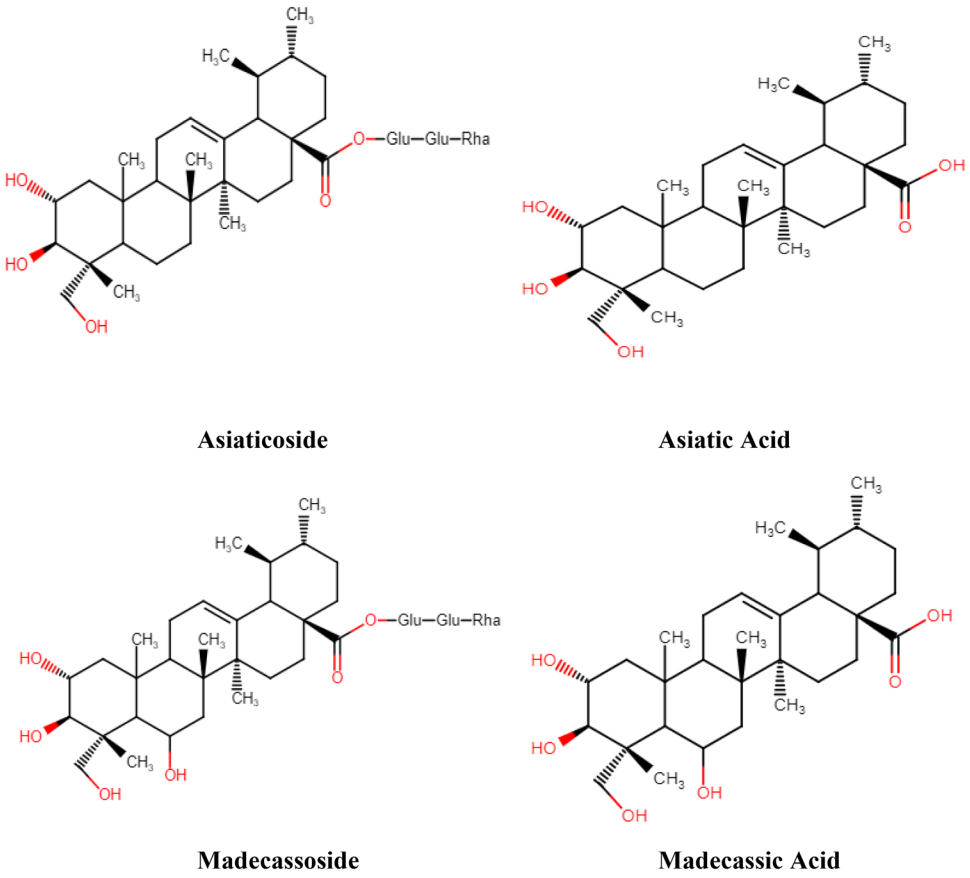


Fig. 1 Structures of major centellosides. (PubChem)

moist, gloomy areas upto an altitude of 7000 ft and may be seen along the sides of streams, rivers, ponds, and irrigated fields. (Gohil et al. 2010) Among this group of secondary metabolites, the most prominent triterpenoid centellosides include asiaticoside, asiatic acid, madecassoside and madecassic acid (Kundu et al. 2016; Roy and Bharadvaja 2017a, b) (Fig. 1). Since ages, it has been used in Ayurveda and Chinese traditional medicine. Studies reveal that these compounds show a broad-spectrum of health benefits, including memory enhancement, anti-inflammatory, antibacterial, anticancer, antidiabetic, antiproliferative, antioxidant, and wound-healing properties (Stafford et al. 2008; Chauhan et al. 2010; Arumugam et al. 2011; Mutua et al. 2013; Kundu et al. 2016; Park et al. 2017) as well as can also be used as tonics for blood circulation and overall, wellbeing. As per the reports of the Export and Import Bank of India, *C. asiatica* has immense importance in the international market of medicinal plants trade, with Asia Pacific (excluding Japan) being one of the most prominent regions. The market is also expected to grow significantly in the US and Europe, owing to the changing lifestyle and increased inclination towards plant-based supplements (https://www.transparencymarketresearch.com/centella-asiatica-market.html). However, there is a significant reduction in the natural production of this herb due to unregulated exploitation and minimal growth, which is attributable to its versatile therapeutic activities (Roy and Bharadvaja 2017a, b). As a result, it has been classified as an endangered species by the International Union for Conservation of Nature and Natural Resources (IUCN) and is presently vulnerable to extinction because of unnecessary uprooting (Bhavna and Jyoti 2011; Seevaratnam et al. 2012).

The herb harvested from its natural habitat viz. Marshy lands and ponds are often regarded as 'dirty herb' attributed to the presence of undesirable levels of heavy metals, microbial infestations, noxious waste, as well as other toxic substances. The centellosides are synthesised in the leaves and shoots of *C. asiatica* (Prasad et al. 2019). The chemical synthesis of centellosides, however, is either exceptionally challenging or uneconomical due to its high structural complexity, lack of high yielding varieties and inadequate agro-technology. As a result, the scope of plant culture technologies for efficient production of secondary metabolites has gained a lot of interest (Gallego et al. 2014). Biotechnological approaches pose many benefits, such as producing bioactive compounds in a regulated environment, increasing efficiency, reducing the risk of contamination by microbiological pathogens, and so on (Roy et al. 2018a, b; Roy and Bharadvaja 2019; Roy 2021). A few research groups all over the world have developed *C. asiatica* cell cultures giving increased yield of centellosides, yet no such

biotechnological procedures developed till now that has emerged to be commercially viable. In-vitro culturing as well as somatic embryogenesis, which include shoot, callus, cell suspension and hairy root cultures are examples of biotechnological techniques that are being used to establish new approaches for restoration of pharmaceutically important plant species (Luthra et al. 2021). Use of in-vitro culture provides the ability to improve propagation of plants and produce bioactive compounds in greater amounts. In in-vitro conditions plant cells encounter biological as well as phenotypic variations in the presence of various substances such as elicitors, which can be microbes, physical or chemical agents. Elicitation, a well-recognised approach for production of secondary metabolites, has been successfully incorporated for centelloside biosynthesis (Roy and Bharadvaja 2018a, b, 2019; Kantayos et al. 2021; Mahendran et al. 2021). They stimulate or enhance synthesis of particular components that play a vital role in the stress responses in plants (Mahendran et al. 2021). Improved production of desirable substances can now be achieved by optimising the process parameters. Therefore, the present review provides an overview of biological applications of centellosides and different biotechnological methods for its increased production. Recent literature on the same along with techniques for scaling up the production aided by bioreactors and future prospects have also been addressed.

Synthesis pathway

Biosynthesis of centelloid triterpenes takes place via isoprenoid pathway, which begins from complete condensation of a five-carbon compound, 3-isopentenyl pyrophosphate (IPP), with another five-carbon compound, dimethylallyl pyrophosphate (DMAPP) to yield a C_{10} compound, geranyl pyrophosphate (GPP). DMAPP and IPP are produced in plants either through mevalonate pathway within the cytosol or through plastidial methylerythritol phosphate route from pyruvate and phosphoglyceraldehyde (James et al. 2009; Augustin et al. 2011). Further, inclusion of another unit of IPP into GPP results in the formation of farnesyl pyrophosphate (FPP; C_{15}). This sequential condensation takes place under the catalytic influence of an enzyme, farnesyl diphosphate synthase (FPS), which is encoded by a single gene called FPS. Thereafter, two units of FPP undergo a head-to-head joining dimerization reaction to give a linear intermediate compound i.e., squalene (C_{30}) that is subsequently oxygenated to form 2,3-oxidosqualene (OSQ) under the influence of squalene synthase (SQS). Studies advocate that OSQ is the natural precursor of each phytosterols and triterpenoid saponins and plays a critical role in their production. Next, OSQ undergoes enzymatic cyclization to produce cyclic triterpenoids. This step is regulated with the aid of enzyme 2,3-oxidosqualene cyclases (OSCs). Hereby, inner bonds are included within the oxidosqualene backbone via double bond migrations, leading to the formation of polycyclic compounds with numerous 5 and 6-membered rings. The cyclization of OSQ to either protosteryl or dammarenyl cationic species is considered to be a key point of divergence in the biosynthetic pathway of triterpenoid saponins and sterols and is chiefly regulated by OSCs. The oxirane ring is broken down in presence of acidic medium and two conformational structures are procured after a couple of rearrangements. A chair-boat-chair-boat shaped structure is obtained in case of protosteryl cation whereas a chair-chair-chair-boat shaped structure is obtained in case of dammarenyl cation. The former, i.e., protosteryl cation, is the predecessor of sterol in plants and gets transformed into cycloartenol by the enzyme cycloartenol synthase (CYS) while the latter one, i.e., dammarenyl cation is the predecessor of phytosterols. Dammarenyl cations are then converted into various intermediate cations like baccharenyl, lupenyl and olenayl or dammarene-like sapogenins with the help of dammarenediol synthase. Thereafter, lupeol or lupine-type sapogenins are formed from lupenyl cations and oleane, the most copious sapogenin, is synthesized by transformation of oleanyl cation to α or β-amyrin (ursane or oleane skeleton, respectively) with the help of α/β-amyrin synthases. Even though the chemical structure of asiatic acid and madecassic acid suggests their α-amyrin origin, they are apparently obtained from β-amyrin. This part of centelloside biosynthesis still remains undisclosed. The final steps of the pathway are yet to be enlightened, but they are likely to include common plant enzymes and reactions (James et al. 2009). Products of cyclization undergo oxidation, glycosylation, hydroxylation and other substitution reactions with the help of various enzymes to form diverse secondary metabolites. For instance, oligosaccharide chains in saponins are most likely formed by subsequent addition of monosaccharides to the aglycone part through glycosylation reaction and is catalysed by enzyme cytochrome-P450 (Augustin et al. 2011). Similarly, madecassic acid and asiatic acid are glycosylated to procure madecassoside and asiaticoside with the help of enzyme UDP-glycosyltransferases (UGTs) (Kim et al. 2014). The synthesis pathway for centelloside production has been graphically presented in Fig. 2.

Methods for extraction of bioactive compounds

Till date, a variety of procedures have been employed to extract high-quality biologically active compounds from *C. asiatica* but the presence of relatively small amount of bioactive components in this herb makes it an essential task to select the method of extraction very precisely, keeping

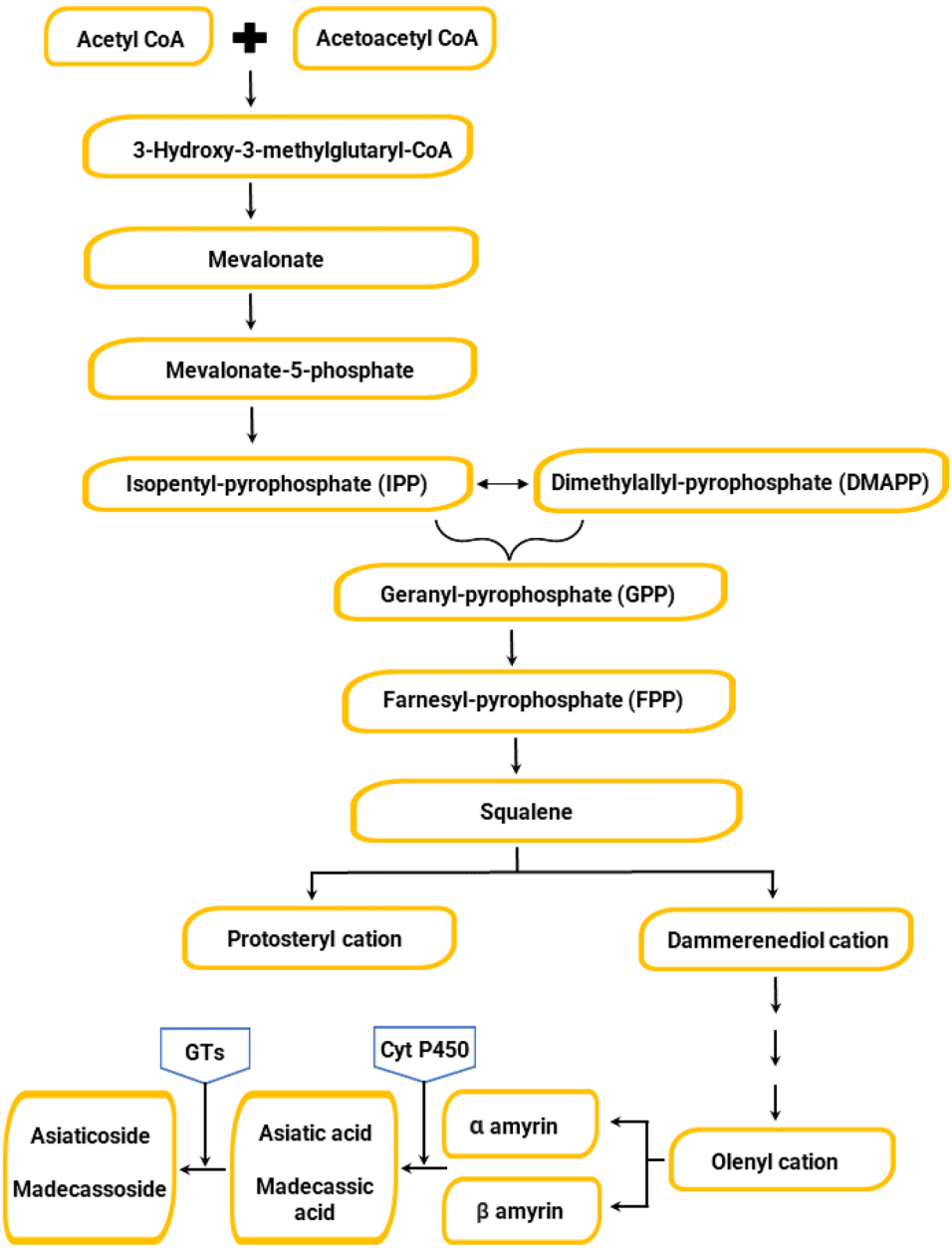


Fig. 2 Route of synthesis of Centellosides (Prasad et al. 2019)

in mind the economic feasibility, shorter time duration and appropriateness of process for the target compound. Generally, plant material is extracted by washing the analyte from matrix within a solvent and allowing it to diffuse through the cell wall (Zhang et al. 2018). However, with the increasing interest of researchers in this plant, several other extraction techniques have also been reported (Fig. 3).

Maceration is one such technique in which plant material is soaked in solvents like methanol, ethanol or a water-alcohol mixture for a specific time duration at a particular temperature, followed by concentrating the extract using a rotary evaporator to obtain desired solvent-free extract. Monton et al. (2019) reported highest yield of madecassoside (0.855%), madecassic acid (0.053%), asiaticoside (0.174%) and asiatic acid (0.025%) using this procedure under optimal temperature conditions (60 °C) for 120 min. Maceration has also been widely used in the cosmetic industry with solvents like propylene glycol and water. However, it is a time-consuming process that requires a high degree of temperature and a considerable amount of solvent.

Distillation is another efficient technique to obtain the best quality of essential oils from *C. asiatica* extracts. While steam distillation yields more sesquiterpenoid hydrocarbons (Idris and Mohd Nadzir 2021) water distillation has been proved to be an effective technique for extraction of caryophyllene and monoterpenoid hydrocarbons. In astudy,

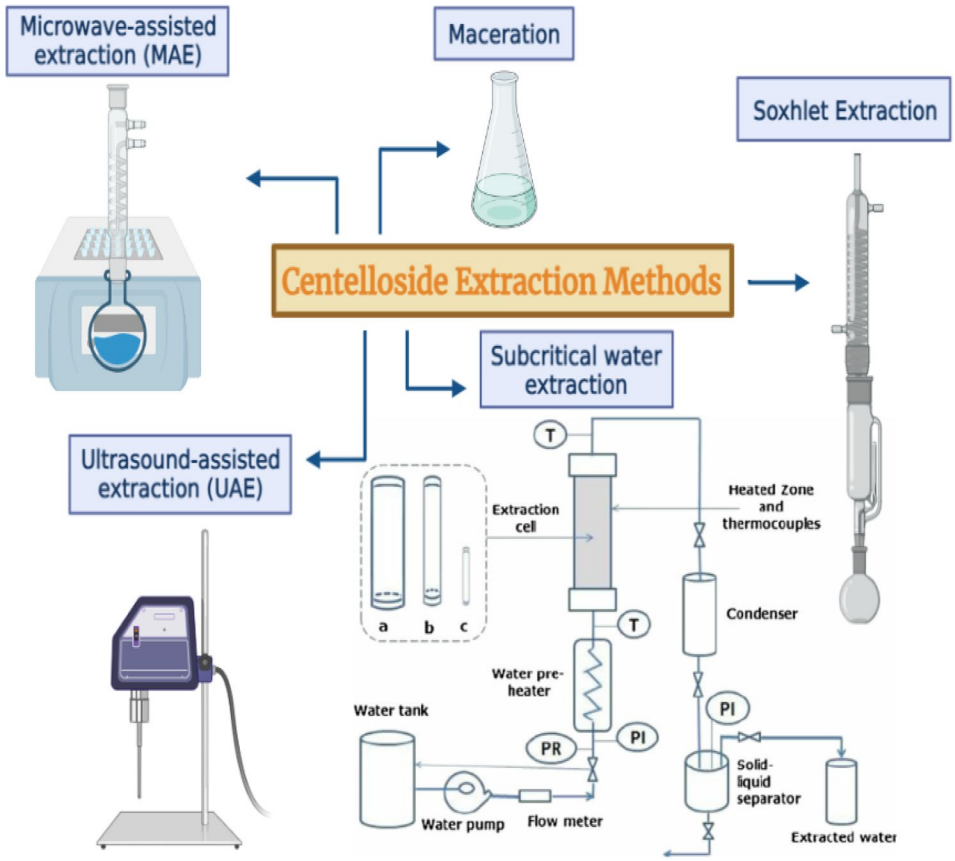


Fig. 3 Methods of extraction of Centellosides

Paudel et al. (2017) obtained a yield of 0.05% essential oils containing 33 compounds by distillation of 85 g of dry leaves of *C. asiatica* for 4 h. However, water distillation consumes a significant amount of time and energy in addition to a large volume of water.

Soxhlet extraction is another widespread approach used for extraction of semi-volatile or non-volatile oils from this herb using solvents like methanol (Vadlapudi et al. 2012), acetone (Jayaprakash and Nagarajan 2016), hexane, petroleum ether, chloroform and water (Sarumathi et al. 2013). In a study conducted by Byakodi et al. (2018) presence of phenols, flavonoids, tannins, saponins, terpenoids and alkaloids was observed in the methanolic extract of *C. asiatica* obtained by the Soxhlet technique. Another study reported that subjecting 500 g of powdered herb to soxhlet extraction for 8 h with 1:4 ethanol to solid ratio yielded an extract which contained alkaloids, terpenoids, saponins and phenols (Thamarai et al. 2012). However, this method is not suitable for thermolabile substances, time-consuming, requires a large volume of solvents and may lead to environmental pollution.

Ultrasound-assisted extraction (*UAE*) is another method of extraction. In a study UAE resulted in a greater yield than soxhlet and maceration extraction procedures for an extract containing 0.47% saponins, 0.34% terpenoids, 0.11% flavonoids and 0.03% alkaloids (Sellathoroe et al. 2019). According to Nithyanandam et al. (2014), UAE is one of the best extraction techniques for considerable recovery of antioxidant compounds like betacyanin from *C. asiatica*. It is comparatively less time-consuming, requires a lesser volume of solvents and suitable for thermolabile substances (Mutalib et al. 2015; Zhang et al. 2018). However, it may lead to degradation of the sample due to long exposure to ultrasound waves (Borhan et al. 2013).

Subcritical water extraction is another harmless, less time-consuming eco-friendly extraction method used to obtain high quality extracts. Although only few studies have reported use of this method. Kim et al. (2009) reported an increase in the amount of asiatic acid from 0 mg/g to 7 mg/g and asiaticoside from 1.1 mg/g to 8.4 mg/g on increasing temperature from 100 to 250 °C in subcritical water extraction technique.

Microwave-Assisted Extraction (*MAE*) method uses microwave energy for rapid heating of aqueous samples at high temperature and an oxygen rich environment. When compared to traditional extraction procedures, it is one of the most effective approaches to minimise time and solvent used for extraction of target compounds (Shen et al. 2009). In a study, Shen et al. (2009) reported that maximum yield was obtained from *C. asiatica* extract on application of

90% methanol solution for 20 min in MAE method. This technique has been enhanced via application of solvent-free microwave extraction (SFME), which eliminates the need for organic solvents as well as allows faster isolation because of direct irradiation of heat to the sample. Another study on combined enzymolysis with MAE, resulted in higher efficacy, lesser time consumption and an eco-friendlier extraction procedure (Wang et al. 2009). However, long exposure to microwaves may lead to degradation of desired constituents and use of enzymes in this technique may increase cost of extraction. Thus, there is an utmost requirement of a suitable method of extraction to obtain high quality yield without damaging the target compound.

Therapeutic applications of centellosides

The tropical perennial ancient herb, *C. asiatica*, has been extensively used since time immemorial for various medicinal purposes due to its high nutritional value (Roy and Bharadvaja 2017a, b). It contains an array of bioactive phytochemicals including asiatic acid, madecassic acid, asiaticoside, madecassoside, collectively known as centellosides, which make it a plant of immense medicinal importance in a wide range of health problems (Orhan 2012; Mohapatra et al. 2021) (Fig. 4). Its leaves appeared to carry tremendous antioxidant properties when investigated for their in-vitro superoxide free radical action, ability to inhibit linoleic acid peroxidation as well as 2,2-diphenyl-1-picryl-hydrazyl (DPPH) radical scavenging ability (Wong et al. 2021a, b). In an experiment performed on Sprague-Dawley rats, *C. asiatica* extracts were found to cause a significant decrease in the levels of oxidative stress and ROS generation (Hussin et al. 2007). Another study revealed that continuous administration of crude methanolic extract in lymphoma bearing mice for 14 days led to a rise in antioxidant enzyme levels and decreased the level of ascorbic acid (Gohli et al. 2010). In another research, it was found that asiaticosides possess the ability to promote and accelerate the process of wound healing by increasing the percentage of collagen in fibroblast cells within the walls of vascular tissue (Seevaratnam et al. 2012). It induces an antioxidative action at some phase during the preliminary action and for that reason performs an essential role in the wound recovery process. In an experiment on open wounds in rats, aqueous extracts of *C. asiatica* was found to speed up the recovery process by promoting cellular proliferation and synthesis of collagen leading to increase in the tensile strength (Gohli et al. 2010). Various reports proposed that alkaloids and flavonoids obtained from different extracts of *C. asiatica* showed antimicrobial action against an extensive range of pathogenic microorganisms Idris and Mohd Nadzir 2021; Wong et al. 2021a, b). In an attempt to assess the in vitro antibacterial activity of methanol hot extract obtained from leaves of this herb, minimum inhibitory concentration (MIC) value and zone of inhibition was determined using disc diffusion technique (2 µg/disc). The efficacy against wild type methicillin resistant S. aureus and Staphylococcus aureus ATCC 25923 was demonstrated by zones of inhibition of 7 mm and 5 mm, respectively (Zaidan et al. 2005). Studies also indicated the potential application of these plant extracts in the treatment of cardiovascular diseases due to the cardioprotective, antihypertensive, anti-atherosclerotic and antihyperlipidemic properties of triterpenes present in it (Razali et al. 2019). Extracts also proved to increase the strength of impaired veins and defend the endothelium of vascular tissue from detrimental effects of venous stasis. Apoptosis is the principal cause of myocardiocyte death in Myocardial Ischemia (Korshunova et al. 2021), which could be reduced with the aid of asiatic acid via lowering the levels of Bax/Bcl-2, caspase-9 and caspase-3 apoptotic factors (Cai et al. 2021; Huang et al. 2016). Presence of bioactive compounds like terpenoids, flavonoids and polyphenols have a significant therapeutic application in prevention and treatment of diabetes mellitus (Tran et al. 2020). Even though the exact mechanism is not clear, studies suggested that asiatic acid stimulates secretion of insulin from β cells of pancreas by inhibiting the adenosine triphosphate (ATP) dependent potassium channels and thus helps in management of diabetes mellitus. In a glucose tolerance test (GTT) performed on alloxan-induced diabetic rats, methanolic and ethanolic extracts of *C. asiatica* showed substantial protection and reduced blood glucose to normal values (Chauhan et al. 2010; Kabir et al. 2014) further reported about the anti-hyperglycemic property of this herb in rats with type II diabetes. It has also been reported that extracts obtained from *C. asiatica* containing centelloids

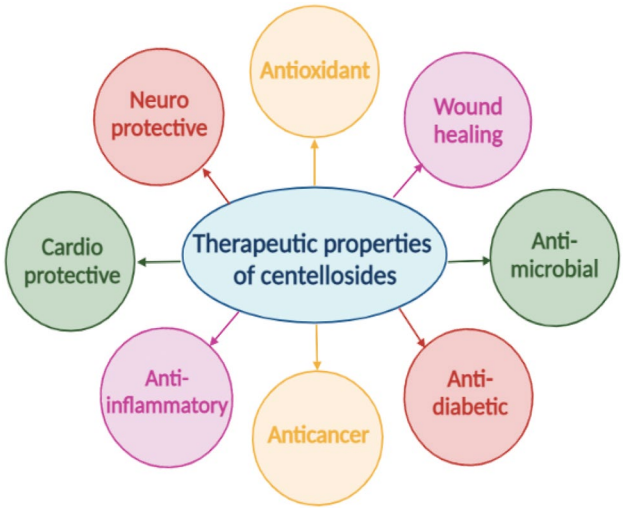


Fig. 4 Therapeutic properties of centellosides

showed a positive impact on the human nervous system and enhances its functioning (Luthra and Roy 2021). Triterpenes present in it help to mitigate various neurological disorders, especially Alzheimer's and Parkinson's disease (Sun et al. 2020). It has also been recognised as a memory enhancer (Orhan et al. 2012) and for its ability to rejuvenate or revive damaged brain and nerve cells (Seevaratnam et al. 2012). In a study conducted on rat's brains, *C. asiatica* extract effectively shielded the brain from age related oxidative damage and exerted a strong neuroprotective action (Sun et al. 2020). Studies also explored the antidepressant and anxiolytic potential of crude extracts as well as its ability to enhance cognitive conduction in humans (Kalshetty et al. 2012; Mala and Tulika 2015). Sirichoat et al. (2015) discovered that asiatic acid promotes hippocampus neurogenesis, which may act as a powerful cognitive enhancer. Another study revealed that supplementing gotu-kola extract at a certain dosage for weeks was beneficial in the therapy of cognitive function impairment post stroke (Farhana et al. 2016). Additionally, studies reported its anti-inflammatory action using human RBC membrane stabilization procedure in vitro Membrane stabilisation was assessed for diclofenac sodium and methanolic extracts of *C. asiatica* extracts at various concentrations, and was reported to be maximum (94.7%) at 2000 g/ml dose (Chippada et al. 2011). Another study revealed that the aqueous and alcoholic extracts of *C. asiatica* inhibited edema by 46.31% and 71.18%, respectively after 3 h of administration. (George et al. 2009) Madecassoside has been reported to be responsible for anti-inflammatory properties. Furthermore, studies also revealed that asiatic acid was very effective in treatment of endocrine system disorders like obesity (Sun et al. 2020) and madecassoside could be a promising therapeutic candidate for the treatment of osteolytic diseases (Wang et al. 2019). A partially purified methanolic extract of this herb was found to inhibit the growth of human breast cancer cells by promoting apoptosis in MCF-7 cells, which was demonstrated by loss of mitochondrial membrane potential, nuclear condensation and DNA breaks identified via TUNEL reactivity (Babykutty et al. 2009). Considerable amount of DNA damage and apoptotic cell death was observed on human HepG2 cell lines when tested with *C. asiatica* juice using MTT assay (Hussin et al. 2014). In ovarian cancer cells, treatment with 40 µg/ml asiatic acid, reduced the viability by approximately 50%, arrested the cell cycle in the G0/G1 phase as well as enhanced apoptosis by about 7–10 times (Ren et al. 2016). Therapeutic response of extracts may also be seen in treatment of digestive ailments, witnessed by improvement in gastric mucosal layer damage and reduction in *H. pylori* bacteria gastric colonization (Sun et al. 2020). Nonetheless, it also finds its application in the cosmetic industry where its extract is used in preparation of antiaging and skin moisturizing formulations as well as in treatment of dry and sensitive skin. Emulsions and gel formulations prepared from it have also been known to prevent acne, dermatitis and oxidative skin damage (Sawatdee et al. 2016; Shen et al. 2019).

Biotechnological methods for centelloside production

Many products obtained from plants are structurally complex, making their synthesis challenging and expensive, and most plants that produce these useful natural ingredients are vulnerable or grow in undesirable ecosystems. As a result, biotechnological methods are a viable and cost-effective way to obtain a safe and consistent supply of bioactive metabolites. Uncontrolled exploitation and growing commercial need have triggered production of centellosides in in-vitro cultures of *C. asiatica* to establish a persistent inventory (Dakah et al. 2014). For synthesising bioactive compounds, various methods are established, including in-vitro cell culture, organ culture as well as hairy root culture of plants (Roy and Bharadvaja 2018a, b; Luthra et al. 2021) (Fig. 5). The ability to synthesize and increase concentrations of these compounds in plant tissue cultures is enormous (Pagare et al. 2015; Roy et al. 2016; Roy and Bharadvaja 2019; Roy 2021).

Shoot culture is described by the induction of shoots via explant without the presence of any intervening callus. Direct development of shoots from an appropriate explant poses a lot of significance for clonal multiplication. Shoot cultures have the capacity to produce bioactive molecules and have shown genetic integrity (Gantait et al. 2018). Similarly, adventitious roots are also used for production of secondary metabolites from different plant species (Le et al. 2018; Roy and Bharadvaja 2019). Some studies have been reported on formation of secondary metabolites via tissue culturing. The concentration of centellosides varies from

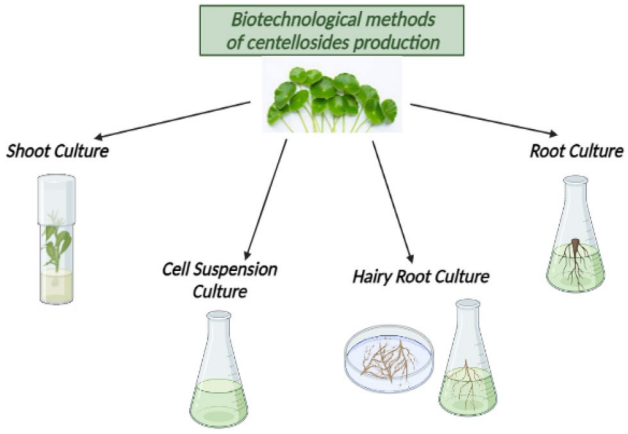


Fig. 5 Biotechnological methods for production of Centellosides (Luthra et al. 2021)

tissue to tissue and among two phenotypes in-vitro and in-vivo. Two separate phenotypic variants of leaf i.e., fringed and smooth shaped, were utilized for examining the terpenoid composition in three explants (namely leaf, roots, and petiole) in *C. asiatica*. In contrast to the roots and petiole, leaves showed maximum concentration of terpenoids (Aziz et al. 2007). Smooth leaves consisted of 1.15 ± 0.10% and 1.65 ± 0.01% dry weight (DW) of asiaticosides and madecassoside, respectively, whereas fringed leaves comprised of 0.79 ± 0.03% asiaticosides as well as 0.97 ± 0.06% madecassoside. In the petiole, smooth phenotypic variant constituted 0.16 ± 0.01% and 0.18 ± 0.14% DW of asiaticoside and madecassoside, respectively, while in the roots, the S phenotypic variant comprised 0.12 ± 0.01% DW of asiaticoside (Aziz et al. 2007).

Callus as well as cell suspension culture were developed by many research groups and they have proven to be an effective, although less productive method for generation of centellosides. In a study cell suspension cultures of *C. asiatica* possessed substantially increased concentrations of asiaticoside i.e., 45.35 mg/g DW in comparison to in-vitro shoot cultures having only 10.55 mg/g DW (Loc et al. 2010). Gandi and Giri (2013) also determined production of asiatic acid, and it was found that callus contained a larger proportion than shoot buds. Asiatic acid levels in regenerative shoot buds were found to be 1.02 ± 0.03 mg/g and 0.47 ± 0.08 mg/g fresh weight (FW), respectively, whereas callus comprised 1.46 ± 0.06 mg/g of FW (Gandi and Giri 2013). Upon adjusting the parameters of callus culture for an increased yield, centelloside production was evaluated in suspension cultures of *C. asiatica.* This was obtained from callus isolated from fresh leaf and petiole. Madecassoside was the key component, accompanied by asiaticoside, madecassic acid, and asiatic acid, with an overall production of about 900 g/g DW (Mangas et al. 2008). Poor generation of centellosides in such culture may be explained by lower expression of CabAS as well as CaSQS genes in callus than that in the leaves. Centelloside synthesis using suspension cultures can also be increased through feeding studies. *C. asiatica* cells have the ability to turn -amyrin to centellosides, as revealed by the fact that 7 days after augmenting the cultured cells with this possible precursor, 84% of cells were successfully produced centellosides (Hernandez-Vazquez et al. 2010).

There are limited studies conducted on production of centellosides from the hairy root culture of *C. asiatica.* Hairy root disease is caused by *Agrobacterium rhizogenes* infecting the plants. In contrast to untransformed roots, agrobacterium modified roots develop robust production of bioactive compounds in hormone-free medium (Roy 2021). Bacteria which yield higher levels of bioactive compounds can benefit from elicitation or modification in the growth medium (Bensaddek et al. 2008; Chandra and Chandra 2011). Post transformation with *Agrobacterium rhizogenes*, asiaticoside production increased by 166 to 172% in *C. asiatica* that could be supplemented by treating with elicitor molecules (Ruslan et al. 2012). Studies have been conducted using elicitors, precursors and stimulants for enhanced production. Different studies that have been conducted for production of centellosides have been mentioned in Table 1.

Genetically transformed cells and use of precursors for improve centellosides production

Many genes in *C. asiatica* have been found to be overexpressed by the use of precursors and elicitors. *C. asiatica* genomes already have successfully transcribed and cloned, albeit their use in in-vitro methods utilizing metabolic engineering approaches is still confined. Scientists will be able to better investigate the centellosides pathway, discover specific genes encoding critical enzymes, as well as manipulate its expression, due to the presence of sophisticated new methods for genomic investigations. Cytochrome P450 enzyme, which is expressed by a class of genes is engaged in a variety of primary as well as supplementary metabolic activities, and its functioning suggests that they play a key role in centellosides pathway's alterations. CYP93E1 in *Glycine max* (Shibuya et al. 2006), CYP51H10 (Sad2) in *A. strigosa* (Qi et al. 2006), and CYP88D6 as well as CYP9E3 in *Glycine uralensis* (Seki et al. 2008) are reported for being engaged in triterpenoid-saponin production. A hairy root cDNA sample of this herb treated with MeJA was used to create an expression sequencing tagging repository (Kim et al. 2014), out of which six P450 genes got examined using RT-PCR. MeJA, which is thought for boosting triterpene conversion, was found to enhance transcriptional activity of CYT50B04, CYT47G03, as well as CYT46B07 genes (Kim et al. 2014), implying that these serve a significant part in the synthesis. Furthermore, upon MeJA administration, RT-PCR examination of four UGTs by this dataset showed a rise in the expression of a 35F01 gene, indicating that it is involved inside the relationship between glucose and centellosides synthesis. Jisha et al. (2018) reported enhanced asiaticoside amount in *Centella asiatica* (L.) Urban with the cell wall extract from *P. indica* (PiCWE). They found that PiCWE play important role in enhancing transcript level expression of key genes which is involved in asiaticoside pathway. It has also been observed that *A. rhizogenes* rol gene showed rise in synthesis of plants' secondary metabolite in modified tissues, and in such cases increased asiaticoside content of modified roots might be as a result of the existence of rol genes (Rogowska and Szakiel 2021).

Table 1 Biotechnological approaches for centelloside production

Type of culture	Culture conditions	Plant growth regulators	Precursor/ elicitor	Yield of centellosides	References
Whole plant culture	B5 medium, 3% sucrose, 25 ± 2 °C, 16 h light 8 h dark, 10 weeks	–	5mM $CdCl_2$	3.98 mg/g DW asiaticoside	Kim et al. (2004)
			5mM $CuCl_2$	4.2 mg/g DW asiaticoside	
			0.1 g/L Yeast extract	5.91 mg/g DW asiaticoside	
			0.01mM MeJA	6.74 mg/g DW asiaticoside	
Cell suspension culture	MS medium, 150 rpm, 25 ± 2 °C, 55–60% relative humidity	6-benzylaminopurine, α-naphthalene acetic acid, 2,4-dichlorophenoxyacetic acid	–	494.6 mg/g DW asiaticoside	Nath et al. (2005)
Whole plant	MS medium, 23 °C, continuous light	–	–	1.7 mg/g DW centellosides	Mangas et al. (2006)
Callus culture		2 mg/L 6-benzylaminopurine	0.1 mg/L indole-3-butyric acid, 2 mg/L 4PU-30	0.85 mg/g DW centellosides	
Hairy root culture	MS medium, 3% sucrose, 0.8% agar at 23 ± 2 °C, 16 h light period	–	0.1 mM MeJA	7.12 mg/g DW asiaticoside	Kim et al. (2007)
Callus and cell suspension culture	MS medium, vitamins, 30 g/L sucrose, 1 g/L casein hydrolysate	2,4-Dichlorophenoxyacetic acid, 6-Benzylaminopurine	–	0.15–0.25% DW Asiatic acid	James et al. (2008)
Whole plant culture	MS medium	–	100 µM MeJA	In leaves, 1.4 mg/g DW asiaticoside, 0.6 mg/g DW madecassoside, 0.3 mg/g DW asiatic acid, 0.25 mg/g DW madecassic acid	Mangas et al. (2008)
Hairy root culture	MS medium, 3% sucrose, 0.8% agar, 23 ± 2 °C, 16 h photoperiod	–	100 µM MeJA	0.17% DW madecassoside, 0.38% DW asiaticoside	Kim et al. (2010)
Cell suspension culture	MS medium, 30 g sucrose, 100 rpm, dark at 25 °C.	2 mg/L 2,4-Dichlorophenoxyacetic acid, 0.1 mg/L 6-benzyladenine	100 µM MeJA	1.1 mg/g DW asiaticoside and madecassoside	Bonfill et al. (2011)
Whole plant culture	50ml half-strength MS medium, 3% sucrose,16 h/day light irradiation, 5weeks	–	0.025 mg/L Thidiazuron and 0.1 mM MeJA	2.40- and 2.44 times increase in madecassoside and asiaticoside	Yoo et al. (2011)
Cell suspension culture	MS medium, 20 g/L sucrose, light intensity: 2500 lx, 10 h daylight (photoperiod)	1 mg/L 6-benzylaminopurine, 1 mg/L 2-(1-naphthyl) acetic acid	100 µM Salicylic acid	102.6 mg/g DW asiaticoside	Loc and Giang (2012)
			4 g/L Yeast extract	165.4 mg/g DW asiaticoside	
Callus culture	MS medium, 4% (w/w) sucrose, 0.9% agar	1.0 mg/L indole-3-acetic acid, 1.0 mg/L 6-benzylaminopurine	Pectin	31% Increase in asiaticoside	Ruslan et al. (2012)
Cell suspension culture			50 µM MeJA	171% Increase in asiaticoside than explant	
			25 µM copper ion	144% Increase in asiaticoside than explant	

Table 1 (continued)

Type of culture	Culture conditions	Plant growth regulators	Precursor/ elicitor	Yield of centellosides	References
Cell suspension culture	MS medium, vitamins, 30 g/ Lsucrose, 1 g/L casein hydrolysate	2,4-Dichlorophenoxyacetic acid, 6-Benzylaminopurine	0.2 mM MeJA	10.6 µg/g FW madecassoside, 20.9 µg/g FW asiaticoside, 8.0 µg/g FW madecassic acid, 4.9 µg/g FW asiatic acid	James et al. (2013)
Shoot culture	MS medium, agar gel, 30 g/L sucrose, 2.5 mg/L kinetin, pH 5.8 ± 0.04, 16 h photoperiod	–	3% v/v trichoderma harzianum culture filtrate	2.35-times increase in asiaticoside	Prasad et al. (2013)
Hairy root culture	MS medium, 1week	–	30ppm Chito-oligosaccharide	5.97 mg/g FW asiaticoside	Zahanis et al. (2016)
Whole plant culture	MS medium, 30 g/L sucrose, dark, 25 ºC, 100 rpm	2 mg/L 2,4-dichlorophenoxyacetic acid, 0.1 mg/L 6-benzylaminopurine	Elicitor: 1µM coronatine, Precursor: 125 mg/L copal extract	10.9-times increase in centelloside	Hidalgo et al. (2016)
			Elicitor: 1 µM coronatine, Precursor: 125 mg/L MER	8.7-times increase in centelloside	
Cell suspension culture	MS medium	2 mg/L 2,4-dichlorophenoxyacetic acid, 3 mg/L thidiazuron	100 µM Salicylic acid	114.25 mg/g DW madecassoside	Loc et al. (2017)
Hairy root culture (diploid)	MS medium, 3.0% sucrose, 5.5 g/l agargel, pH 5.8, 25 °C, 16 h light/8 h dark	–	Control	0 µg/mg DW triterpenoid	Nguyen et al. (2019)
			50 µM MeJA	27.25 ± 0.27 µg/mg DW triterpenoid	
Hairy root culture (tetraploid)			50 µM MeJA	16.29 ± 6.32 µg/mg DW triterpenoid	
			100 µM MeJA	16.31 ± 9.24 µg/mg DW triterpenoid	
Hairy root culture	100 mL MS medium, 50 g/L sucrose, dark, 24 ± 1 °C, 100 rpm	–	400 µM MeJA	60.25 mg/g DW of Triterpenoids	Baek et al. (2020)
			2.5 mM squalene	57.53 mg/g DW of triterpenoids	
			5 mM pyruvic acid	29.13 mg/g DW of triterpenoids	
Axenic shoot cultures	MS medium, 1.3% phytagel 3% sucrose, 25 ± 2 °C under 12 h	1.0 mg/L BAP and 0.2 mg/L NAA	Control	0.21 mg/g DW	Ranjith et al. (2021a, b)
			1 mg/L salicylic acid	1.34 mg/g DW asiaticoside	
			2 mg/L salicylic acid	1.32 mg/g DW asiaticoside	
			1 mg/L AgNO3	1.26 mg/g DW asiaticoside	
			2 mg/L AgNO3	1.10 mg/g DW asiaticoside	

Biotechnological improvements and scale up strategies

Despite the fact that demand for continuous production has been rising rapidly over the years, the availability of secondary metabolites continues to remain restricted. Thus, there is a dire need to address this concern and scale-up the production of these secondary metabolites to meet the current requirements. Several effective researches have been conducted to establish alternative methods for regulating the supply of these biologically active compounds irrespective of climate and plant abundance (Monica et al. 2013). Improving the synthesis of secondary metabolites in cell culture by supplying the growth media with various elicitor molecules is a widely used technique, suggesting that several such molecules are associated with plants' defense mechanisms (Kundu et al. 2016; Krishnan et al. 2019).

Elicitors may be regarded as low molecular weight moieties that stimulate plant immune system response by activation of signal cascade when exposed to certain biotic and abiotic factors (Baenas et al. 2014). They may be plant or pathogen derived (termed as endogenous and exogenous elicitors, respectively) and have been shown to induce aggregation of secondary metabolites in plants as well as plant cell cultures as a part of their defence mechanism (Namdeo et al. 2007; Luthra et al. 2021). They are categorised into biotic and abiotic elicitors. The most extensively used biotic elicitors include polysaccharides, enzymes and other complex molecules obtained from a living source (Vasconsuelo et al. 2007). Abiotic elicitors, on the other hand, include UV rays, osmotic stress induced by compounds (like $AgNO_3$, NaCl, KCl, $CdCl_2$, $CuCl_2$), sunlight, temperature shifts, intracellular signaling molecules (like methyl jasmonate, jasmonic and salicylic acid) and other physical as well as chemical factors. Precursor feeding is another well-established approach utilised for enhancing secondary metabolite synthesis in *in-vitro* conditions. Many organic substances are being applied to the growth media for promoting production of bioactive compounds based on understanding of the biosynthesis pathway. Addition of biosynthetic precursors like copal extracts, squalene, pyruvic acid, etc. has shown to boost the production of desirable substances (Namdeo et al. 2007; Hidalgo et al. 2016).

According to a study, 0.1 mM MeJA is adequate enough to increase the concentration of squalene synthase mRNA and β-amyrin synthase mRNA, and can increase the triterpene saponin production in callus and shoot cultures (Kim et al. 2007). In comparison to plant leaf, control cell culture, and 2-month-old callus cultures, asiaticoside production was enhanced by around 31% in case of callus treated with pectin and nearly 171%, 125%, and 494% in cell cultures elicited with MeJA. In another study, application of 100 M MeJA to the growth media substantially increased asiaticoside content in hygromycin-resistant hairy roots. Such findings are linked to a rise in CabAS expression following elicitation, establishing the gene's function in asiaticoside biosynthetic pathway (Kim et al. 2007). In a research study, it was observed that adding 100 µM MeJA to a cell suspension culture of *C. asiatica* boosted the centelloside content to seven folds; with madecassoside being the most prevalent one (Bonfill et al. 2010). Similar findings were also reported by Krishnan et al. (2019) suggesting that when MS media was elicited with 100 µM MeJa, growth was inhibited in callus, shoot, and cell suspension cultures, but the volume of asiaticoside increased to 69-times in callus culture, 39-times in shoot culture, and 1.9-times in cell suspension culture. In a study, effects of various potential metabolite stimulants (PMS) such as salicylic acid, meat peptone, casein acid hydrolysate, silver nitrate, and copper sulphate were investigated on axenic shoot culture of *C. asiatica.* After 3 weeks of treatment, silver nitrate was found to exhibit the best results, boosting up the madecassoside concentration by nearly 6 times and 50 times higher gene expression of genes like β-amyrin synthase (βAS), which code for key enzymes involved in the synthesis of triterpenoids (Ranjith et al. 2021a, b).

Prasad et al. (2012) recorded the growth patterns and metabolite synthesis of this essential herb in a hydroponic system across a 70-days cycle, collecting data at an interval of every 7 days. The results depicted that the plant followed an exponential growth pattern beginning from day 21 and reached its peak on day 42. Thereby, indicating that use of hydroponics to cultivate *C. asiatica* for 6 to 7 weeks could be another potential approach to ensure satisfactory output; which in this case was found to be 5.5% triterpenoids including 3.6% madecassic acid, 1.1% madecassoside, 0.63% asiatic acid and 0.17% asiaticoside respectively. Hidalgo et al. (2016) recommended another approach to increase the production of centellosides using a 2 L cell bag to check one of the techniques that delivered high quality output in shake flasks, namely CORO (1 µM) elicitation and MER (125 mg/L) feeding. It was found that *C. asiatica* developed extremely well under such conditions, achieving a growth index of 4.8 post 10-days of culture and yielded 4.2 times higher centelloside content (i.e., 7.3 m/g DW) than a shake flask under comparable conditions.

While numerous techniques for increasing the production on a small scale have been established, the most common issue is scaling up the whole process to commercial level. One of the approaches to overcome this obstacle is by using bioreactors, which help in maintaining a bio-active ecosystem to achieve wide-scale production in an economically feasible manner (Gutierrez-Valdes et al. 2020; Marchev and Georgiev 2020; Roy 2021). Although there are various different types of bioreactors in the market, factors including

the growth, developmental behaviour pattern, low shear stress, morphology of the culture as well as the operational factors such as pH, temperature, O_2 and substrate concentration play a significant role in determining the most appropriate bioreactor for best results. Studies recommend the use of disposable bioreactors as they offer a larger capacity up to 200 L and ensure a higher yield of secondary metabolites (Lehmann et al. 2014; Loc and Nhat 2013) grew cell suspension culture of *C. asiatica* in a 5 L bioreactor using an inoculum of size 30 g agitated at 120 rpm with an aeration rate of 2 L/min and reported highest yield of cell biomass i.e.,116.17 g of FW and 12.24 g of DW, on the 24th day of experiment. Thereafter, cell growth slowed down and biomass on the 30th day was just 94.76 g FW and 9.01 g DW. Similarly, concentration of asiaticoside was examined which revealed that total yield increased from day 1 until day 24, with maximum yield of 56.21 mg/g DW, and then declined from day 27 to day 30. In an attempt to examine the influence of culture conditions on production, cell suspension was moved into a 14 L bioreactor under varied conditions including size of inoculum extending from 30 to 100 g, rate of agitation varying from 120 rpm to 200 rpm and aeration rate ranging from 2 to 3 L/min. Findings revealed that these factors played a vital role in culturing and optimum productivity was noted when a 100 g inoculum was cultured at 150 rpm with an aeration rate of 2.5 L/min (Loc et al. 2013). In another study, outcomes of agitated cultures and greenhouse crops were contrasted to the aggregation of specific bioactive compounds of *C. asiatica* shoot cultures, which were grown in bioreactor RITA® (R-1 and R-5) as well as Plantform TM(P-5). After harvesting the biomass on 23rd day of culture, it was observed that supplementing the culture with MeJa resulted in a substantial improvement in accumulation of metabolites i.e., 3.6-times in P-5 and nearly 3-times in AC, R-1, and R-5, when compared to greenhouse-grown species. Centellosides were found in greater abundance in greenhouse-grown plants (Skrzypczak-Pietraszek et al. 2019).

Future prospects and concluding remarks

Centellosides have been gaining popularity since recent times due to their extensive medicinal and nutraceutical benefits. These triterpenoid saponins have pharmaceutical attributes that have been useful in the treatment of a wide range of illnesses. This triterpenoid is employed in the cosmetics industry in addition to its pharmaceutical value. Overexploitation of medicinal herb *C. asiatica* for these secondary metabolites has exhausted its natural supply, necessitating the development of new methods for their production. With the help of in-vitro cultures and genetic modifications, biotechnology allows the cells, tissue as well as the whole plant to be exploited on a sustainable basis for the production of bioactive compounds like centellosides. In-vitro production allows long-term use of herbs as well as formation of bioactive compounds including triterpenoids, and also their amplification by the use of different agents such as elicitors. Hairy root culture promotes secondary metabolite production mostly in the aerial portion of a crop; nevertheless, some studies have solved this challenge by combining hairy root culture and elicitation process. Very few studies mention cell suspension cultures for *C. asiatica* in bioreactors, which is used for large-scale synthesis of secondary metabolites. As a result, suitable bioreactors with a lesser shear rate impeller must be installed, as well as the culture conditions must be optimised leveraging statically strategies that boost the production of these compounds. Elicitor molecules play an essential part in enhancing the content of secondary metabolites, but little is known about how they modify their mechanism in-vitro. Latest research relating to the genomics of *C. asiatica* have discovered a set of genes that are dysregulated in response to elicitors, despite the fact that many genes are already sequenced as well as reproduced using the functional approaches like genomics, metabolomics and transcriptomics. Attempts to address the metabolic processes of these bioactive metabolites which are elicited must be bolstered. Thus, research is intended to investigate mechanisms that have been affected by use of elicitor molecules and identify genes encoding essential enzymes.

Declarations

Conflict of interest No conflict of interest is observed between any of the authors.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Arumugam T, Ayyanar M, Pillai YJK (2011) Phytochemical screening and antibacterial activity of leaf and callus extracts of *Centella asiatica* Bangladesh J Pharmacol 6:55–60

Augustin JM, Kuzina V, Andersen SB, Bak S (2011) Molecular activities, biosynthesis and evolution of triterpenoid saponins. Phytochemistry 72:435–457

Aziz ZA, Davey MR, Power JB et al (2007) Production of asiaticoside and madecassoside in *Centella asiatica* in vitro and in vivo. Biol Plant 51:34–42

Babykutty S, Padikkala J, Sathiadevan P, Vijayakurup V, Azis T, Srinivas P, Gopala S (2009) Apoptosis induction of *Centella asiatica* on human breast cancer cells. Afr J Tradit Complement Altern Med. https://doi.org/10.4314/ajtcam.v6i1.57068

Baek S, Ho TT, Lee H, Jung G, Kim YE, Jeong CS, Park SY (2020) Enhanced biosynthesis of triterpenoids in *Centella asiatica* hairy root culture by precursor feeding and elicitation. Plant Biotechnol Rep 14:45–53

Baenas N, Garcia-Viguera C, Moreno DA (2014) Elicitation: a tool for enriching the bioactive composition of foods. Molecules 19:13541–13563

Bensaddek L, Villarreal ML, Fliniaux MA (2008) Induction and growth of hairy roots for the production of medicinal compounds. Electron J Integr Biosci 3:2–9

Bhavna D, Jyoti K (2011) *Centella asiatica*: the elixir of life. Int J Res Ayurveda Pharm 2:431–438

Bonfill M, Cusido R, Moyano E et al (2010) Centelloside production in *Centella asiatica* cell suspension cultures elicited with methyl jasmonate. Planta Med 76:342

Bonfill M, Mangas S, Moyano E et al (2011) Production of centellosides and phytosterols in cell suspension cultures of *Centella asiatica* Plant Cell Tissue Organ Cult 104:61–67

Borhan MZ, Ahmad R, Rusop M, Abdullah S (2013) Green extraction: enhanced extraction yield of asiatic acid from *Centella asiatica* (L.) nanopowders. J Appl Chem 2013:1–7

Byakodi MK, Bagewadi ZK, Muddapur UM (2018) Phytoconstituents profiling and evaluation of antimicrobial and antioxidant attributes of methanolic extract of *Centella asiatica* Res J Pharm Biol Chem Sci 9:493–500

Cai X, He L, Zhou G, Li S, Liao X (2021) Mogroside IIe ameliorates cardiomyopathy by suppressing cardiomyocyte apoptosis in a Type 2 diabetic model. Front Pharmacol. https://doi.org/10.3389/fphar.2021.650193

Chandra S, Chandra R (2011) Engineering secondary metabolite production in hairy roots. Phytochem Rev 10:371–395

Chauhan PK, Pandey IP, Dhatwalia VK (2010) Evaluation of antidiabetic effect of ethanolic and methanolic extracts of *Centella asiatica* leaves on alloxan induced diabetic rats. Adv Biol Res 4:27–30

Chippada SC, Volluri SS, Bammidi SR, Vangalapati M (2011) In vitro anti-inflammatory activity of methanolic extract of *Centella asiatica* by HRBC membrane stabilisation. Rasayan J Chem 4(2):457–460

Dakah A, Zaid S, Suleiman M et al (2014) In vitro propagation of the medicinal plant *Ziziphora tenuior* L. and evaluation of its antioxidant activity. Saudi J Biol Sci 21:317–323

Farhana KM, Malueka RG, Wibowo S, Gofir A (2016) Effectiveness of GotuKola extract 750 mg and 1000 mg compared with folic acid 3 mg in improving vascular cognitive impairment after stroke. Evid Based Complement Alternat Med 2016:2795915

Gallego A, Ramirez-Estrada K, Vidal-Limon HR et al (2014) Biotechnological production of centellosides in cell cultures of *Centella asiatica* (L) Urban. Eng Life Sci 14(6):633–642

Gandi S, Giri A (2013) Production and quantification of asiatic acid from in vitro raised shoots and callus cultures of *Centella asiatica* (L.) Urban. Ann Phytomed 2:95–101

Gantait S, El-Dawayati MM, Panigrahi J, Labrooy C, Verma SK (2018) The retrospect and prospect of the applications of biotechnology in *Phoenix dactylifera* L. Appl Microbiol Biotechnol 102:8229–8259

George M, Joseph L, Ramaswamy (2009) Anti-allergic, anti-pruritic, andanti-inflammatory activities of *Centella asiatica* extracts. Afr J Tradit Complement Altern Med 6(4):554–559

Gohil KJ, Patel JA, Gajjar AK (2010) Pharmacological review on *Centella asiatica*: a potential herbal cure-all. Indian J Pharm Sci 72(5):546–556

Gray NE, Magana AA, Lak P, Wright KM, Quinn J, Stevens JF, Soumyanath A (2018) *Centella asiatica*: phytochemistry and mechanisms of neuroprotection and cognitive enhancement. Phytochem Rev 17(1):161–194

Gutierrez-Valdes N, Häkkinen ST, Lemasson C, Guillet M, Oksman-Caldentey KM, Ritala A, Cardon F (2020) Hairy root cultures—a versatile tool with multiple applications. Front Plant Sci 11:33

Hernandez-Vazquez L, Bonfill M, Moyano E et al (2010) Conversion of α-amyrin into centellosides by plant cell cultures of *Centella asiatica* Biotechnol Lett 32:315–319

Hidalgo D, Steinmetz V, Brossat M, Tournier-Couturier L, Cusido RM, Corchete P, Palazon J (2016) An optimized biotechnological system for the production of centellosides based on elicitation and bioconversion of *Centella asiatica* cell cultures. Eng Life Sci 17(4):413–419

Huang X, Zuo L, Lv Y, Chen C, Yang Y, Xin H, Qian Y (2016) Asiatic acid attenuates myocardial ischemia/reperfusion injury via Akt/GSK-3β/HIF-1α signaling in rat H9c2 cardiomyocytes. Molecules 21(9):1248

Hussin M, Abdul-Hamid A, Mohamad S, Saari N, Ismail M, Bejo MH (2007) Protective effect of *Centella asiatica* extract and powder on oxidativestress in rats. Food Chem 100(2):535–541

Hussin F, Eshkoor SA, Rahmat A, Othman F, Akim A (2014) The *Centella asiatica* juice effects on DNA damage, apoptosis and gene expressionin hepatocellular carcinoma (HCC). BMC Complement Altern Med 14:32

Idris FN, Mohd Nadzir M (2021) Comparative studies on different extraction methods of *Centella asiatica* and extracts bioactive compounds effects on antimicrobial activities. Antibiotics 10(4):457

James JT, Dubery IA (2009) Pentacyclic triterpenoids from the medicinal herb, *Centella asiatica* (L.) Urban. Molecules 14:3922–3941

James JT, Meyer R, Dubery IA (2008) Characterisation of two phenotypes of *Centella asiatica* in Southern Africa through the composition of four triterpenoids in callus, cell suspensions and leaves. Plant Cell Tissue Organ Cult 94:91–99

James JT, Tugizimana F, Steenkamp PA, Dubery IA (2013) Metabolomic analysis of methyl jasmonate-induced triterpenoidproduction in the medicinal herb *Centella asiatica* (L.) Urban. Molecules 18:4267–4281

Jayaprakash SB, Nagarajan N (2016) Studies on the bioactive compounds and antimicrobial activities of medicinal plant *Centella asiatica* (Linn). J Med Plants Stud 4:181–185

Jisha S, Gouri PR, Anith KN, Sabu KK (2018) *Piriformospora indica* cell wall extract as the best elicitor for asiaticoside production in Centella asiatica (L.) Urban, evidenced by morphological, physiological and molecular analyses. Plant Physiol Biochem 125:106–115

Kabir AU, Samad MB, D'Costa NM, Akhter F, Ahmed et al (2014) Anti-hyperglycemic activity of *Centella asiatica* is partly mediatedby carbohydrase inhibition and glucose-fiber binding. BMC Complement Altern Med 14:31

Kalshetty P, Aswar U, Bodhankar S, Sinnathambi A, Mohan V, Thakurdesai P (2012) Antidepressant effects of standardized extract of *Centella asiatica* L in olfactory bulbectomy model. Biomed Aging Pathol 2(2):48–53

Kantayos V, Kim JS, Baek SH (2021) Enhancing resveratrol bioproduction and anti-melanogenic activities through elicitation in DJ526 cell suspension. Plants 10(8):1653

Katiyar CK, Kanjilal S, Narwaria A (2021) Brain and mental health in Ayurveda. Nutraceuticals in brain health and beyond. Academic Press, Cambridge, pp 81–112

Kim O, Kim M, Hong M et al (2004) Stimulation of asiaticoside accumulation in the whole plant cultures of *Centella asiatica* (L.) Urban by elicitors. Plant Cell Rep 23:339–344

Kim OT, Bang KH, Shin YS et al (2007) Enhanced production of asiaticoside from hairy root cultures of *Centella asiatica* (L.) Urban elicited by methyl jasmonate. Plant Cell Rep 26:1941–1949

Kim W, Kim J, Veriansyah B, Kim J, Lee Y, Oh S, Tjandrawinata RR (2009) Extraction of bioactive components from *Centella asiatica* using subcritical water. J Supercrit Fluids 48:211–216

Kim OT, Kim SH, Ohyama K et al (2010) Upregulation of phytosterol and triterpene biosynthesis in *Centella asiatica* hairy roots overexpressed ginseng farnesyl diphosphate synthase. Plant Cell Rep 29:403–411

Kim OT, Um Y, Jin ML et al (2014) Analysis of expressed sequence tags from *Centella asiatica* (L.) Urban hairy roots elicited by methyl jasmonate to discover genes related to cytochrome P450s and glucosyltransferases. Plant Biotechnol Rep 8:211–220

Korshunova AY, Blagonravov ML, Neborak EV, Syatkin SP, Sklifasovskaya AP, Semyatov SM, Agostinelli E (2021) BCL2regulated apoptotic process in myocardial ischemiareperfusion injury. Int J Mol Med 12:1–9. https://doi.org/10.3892/ijmm.2020.4781

Krishnan ML, Roy A, Bharadvaja N (2019) Elicitation effect on the production of asiaticoside and asiatic acid in shoot, callus, and cell suspension culture of *Centella asiatica*. J App Pharm Sci 9(06):067–074

Kundu K, Roy A, Saxena G, Kumar L, Bharadvaja B (2016) Effect of different carbon sources and elicitors on shoot multiplication in accessions of *Centella asiatica* Med Aromat Plants 5:251

Le KC, Im WT, Paek KY, Park SY (2018) Biotic elicitation of ginsenoside metabolism of mutant adventitious root culture in Panax ginseng. Appl Microbiol Biotechnol 102:1687–1697

Lehmann N, Dittler I, Lämsä M et al (2014) Disposable bioreactors for cultivation of plant cell cultures. In: Murthy PKY, Zhong HN et al (eds) Production of Biomass and bioactive compounds using bioreactor technology. Springer, Dordrecht, pp 17–46

Loc NH, An NTT (2010) Asiaticoside production from Centella (*Centella asiatica* L. Urban) cell culture. Biotechnol Bioprocess Eng 15:1065–1070

Loc NH, Giang NT (2012) Effects of elicitors on the enhancement of asiaticoside biosynthesis in cell cultures of centella (*Centella asiatica* L. Urban). Chem Pap 66:642–648

Loc NH, Nhat NTD (2013) Production of asiaticoside from centella (*Centella asiatica* L. Urban) cells in bioreactor. Asian Pac J Trop Biomed 3(10):806–810

Loc NH, Giang NT, Huy ND, Lan TTP (2017) Accumulation of madecassoside-a major component of centelloside-in centella (Centella asiatica (L) Urban) cells elicited by salicylic acid. Period Biol. https://doi.org/10.18054/pb.v119i1.4386

Luthra R, Roy A (2021) Role of medicinal plants against neurodegenerative diseases. Curr Pharm Biotechnol. https://doi.org/10.2174/1389201022666210211123539

Luthra R, Roy A, Pandit S, Prasad R (2021) Biotechnological methods for the production of ginsenosides. S Afr J Bot 141:25–36

Mahendran G, Iqbal Z, Kumar D, Verma SK, Rout PK, ur Rahman L (2021) Enhanced gymnemic acids production in cell suspension cultures of *Gymnema sylvestre* (Retz.) R. Br. ex Sm. through elicitation. Ind Crops Prod 162:113234

Mala A, Tulika T (2015) Therapeutic efficacy of *Centella asiatica* (L.) and *Momordica charantia*: as traditional medicinal plant. J Plant Sci 3:1–9

Mangas S, Bonfill M, Osuna L, Moyano E, Tortoriello J, Cusido RM, Palazón J (2006) The effect of methyl jasmonate on triterpene and sterol metabolisms of *Centella asiatica*, *Ruscus aculeatus* and *Galphimia glauca* cultured plants. Phytochemistry 67(18):2041–2049

Mangas S, Moyano E, Osuna L et al (2008) Triterpenoid saponin content and the expression level of some related genes in calli of *Centella asiatica* Biotechnol Lett 30:1853–1859

Marchev AS, Georgiev MI (2020) Plant in vitro systems as a sustainable source of active ingredients for cosmeceutical application. Molecules 25(9):2006

Mohapatra P, Ray A, Sandeep IS, Parida R, Mohanty S (2021) Genetic and biochemical stability of in vitro raised and conventionally propagated *Centella asiatica*—a valuable medicinal herb. S Afr J Bot. https://doi.org/10.1016/j.sajb.2021.01.004

Monica J, Ritika R, Anamika M (2013) Enhancement of secondary metabolite biosynthesis in *Bacopa monnieri* an in vitro study. Res J Recent Sci 2:13–16

Monton C, Settharaksa S, Luprasong C, Songsak T (2019) An optimization approach of dynamic maceration of *Centella asiatica* to obtain the highest content of four centelloids by response surface methodology. Braz J Pharmacogn 29:254–261

Mutalib LY (2015) Comparison between conventional and modern methods for extraction of *Rosmarinus officinalis* leaves. Zanco J Med Sci 19:1029–1034

Mutua PM, Gicheru MM, Makanya AN et al (2013) Anti-proliferative activities of *Centella asiatica* extracts on human respiratory epithelial cells in vitro. Int J Morphol 31(4):1322–1327

Namdeo AG (2007) Plant cell elicitation for production of secondary metabolites: a review. Pharmacogn Rev 2007 1:69–79

Nath S, Buragohain A (2005) Establishment of callus and cell suspension cultures of *Centella asiatica* Biol Plant 49:411–413

Nguyen KV, Pongkitwitoon B, Pathomwichaiwat T, Viboonjun U, Prathanturarug S (2019) Effects of methyl jasmonate on the growth and triterpenoid production of diploid and tetraploid *Centella asiatica* (L.) Urb. hairy root cultures. Sci Rep 9(1):1–8

Nithyanandam R, Shapheri MR, Nasir MH (2014) Antioxidant potential of Malaysian Herb *Centella asiatica* Int Conf Environ Chem Biol 78:56–60

Orhan IE (2012) *Centella asiatica* (L.) Urban: from traditional medicine to modern medicine with neuroprotective potential. Evid Based Complement Altern Med 2012:946259

Pagare S, Bhatia M, Tripathi N et al (2015) Secondary metabolites of plants and their role overview. Curr Trends Biotechnol Pharm 9:293–304

Park JH, Choi JY, Son DJ et al (2017) Anti-inflammatory effect of titrated extract of *Centella asiatica* in phthalic anhydride-induced allergic dermatitis animal model. Int J Mol Sci 18:738

Paudel P, Satyal P, Dosoky NS, Setzer WN (2017) Chemical composition and biological activity of *Centella asiatica* essential oil from Nepal. Am J Essent Oils Nat Prod 5:5–8

Prasad A, Pragadheesh VS, Mathur A, Srivastava NK, Singh M, Mathur AK (2012) Growth and centelloside production in hydroponically established medicinal plant-*Centella asiatica* (L.). Ind Crops Prod 35(1):309–312

Prasad A, Mathur A, Kalra A et al (2013) Fungal elicitor mediated enhancement in growth and asiaticoside content of *Centella asiatica* L. shoot cultures. Plant Growth Regul 69:265–273

Prasad A, Mathur AK, Mathur A (2019) Advances and emerging research trends for modulation of centelloside biosynthesis in *Centella asiatica* (L.) Urban—a review. Ind Crops Prod 141:111768

Qi X, Bakht S, Qin B, Leggett M et al (2006) A different function for a member of an ancient and highly conserved cytochrome P450 family: From essential sterols to plant defense. Proc Natl Acad Sci USA 103:18848–18853

Ranjith GP, Jisha S, Hemanthakumar AS, Saji CV, Shenoi RA, Sabu KK (2021a) Impact of potential stimulants on asiaticoside and madecassoside levels and expression of triterpenoid-related genes in axenic shoot cultures of *Centella asiatica* (L.) Urb. Phytochemistry 186:112735

Ranjith GP, Jisha S, Hemanthakumar AS, Saji CV, Shenoi RA, Sabu KK (2021b) Impact of potential stimulants on asiaticoside and madecassoside levels and expression of triterpenoid-related genes in axenic shoot cultures of *Centella asiatica* (L.) Urb. Phytochemistry 186:112735

Razali NNM, Ng CT, Fong LY (2019) Cardiovascular protective effects of *Centella asiatica* and its triterpenes: a review. Planta Med 85(16):1203–1215

Ren L, Cao QX, Zhai FR, Yang SQ, Zhang HX (2016) Asiatic acid exertsanticancer potential in human ovarian cancer cells via suppression ofPI3K/Akt/mTOR signalling. Pharm Biol 54(11):2377–2382

Rogowska A, Szakiel A (2021) Enhancement of phytosterol and triterpenoid production in plant hairy root cultures—simultaneous stimulation or competition? Plants 10(10):2028

Roy A (2021) Hairy root culture an alternative for bioactive compound production from medicinal plants. Curr Pharm Biotechnol 22(1):136–149

Roy A, Bharadvaja N (2017a) *Centella asiatica*: a pharmaceutically important medicinal plant. Curr Trends Biomed Eng Biosci 5(3):1–5

Roy A, Bharadvaja N (2017b) Effect of different culture medias on shoot multiplication and stigmasterol content in accessions of *Centella asiatica*. Int J Ayurvedic Herb Med 7:2643–2650

Roy A, Bharadvaja N (2018a) Biotechnological approaches for the production of pharmaceutically important compounds: plumbagin. Curr Pharm Biotechnol 19(5):372–381

Roy A, Bharadvaja N (2018b) Effect of various culture conditions on shoot multiplication and GC–MS analysis of Plumbago zeylanica accessions for plumbagin production. Acta Physiol Plant 40(11):1–11

Roy A, Bharadvaja N (2019) Establishment of root suspension culture of *Plumbago zeylanica* and enhanced production of plumbagin. Ind Crops Prod 137:419–427

Roy A, Kundu K, Saxena G, Kumar L, Bharadvaja N (2016) Effect of different media and growth hormones on shoot multiplication of in vitro grown *Centella asiatica* accessions. Adv Tech Biol Med. https://doi.org/10.4172/2379-1764.1000172

Roy A, Jauhari N, Bharadvaja N (2018a) Medicinal plants as a potential source of chemopreventive agents. Anticancer plants: natural products and biotechnological implements. Springer, Singapore, pp 109–139

Roy A, Krishnan L, Bharadvaja N (2018b) Qualitative and quantitative phytochemical analysis of *Centella asiatica*. Nat Prod Chem Res 6(4):1–4

Ruslan K, Selfitri AD, Bulan SA, Rukayadi Y (2012) Effect of *Agrobacterium rhizogenes* and elicitation on the asiaticoside production in cell cultures of *Centella asiatica*. Pharmacogn Mag 8:111–115

Sarumathi A, Anbu S, Saravanan N (2013) Antibacterial activity of *Centella asiatica* (Linn.,) leaves. Int J Chem Pharm Sci 1:393–395

Sawatdee S, Choochuay K, Chanthorn W, Srichana T (2016) Evaluation of the topical spray containing *Centella asiatica* extract and its efficacy on excision wounds in rats. Acta Pharm 66(2):233–244

Seevaratnam V, Banumathi P, Premalatha MR et al (2012) Functional properties of *Centella asiatica* (L.) a review. Int J Pharm Pharm Sci 4:8–14

Seki H, Ohyama K, Sawai S, Mizutani M et al (2008) Licorice β-amyrin 11-oxidase, a cytochrome P450 with a key role in the biosynthesis of the triterpene sweetener glycyrrhizin. Proc Natl Acad Sci USA 105:14204–14209

Sellathoroe S, Marimuthu S, Ramays TR (2019) Comparison of different extraction methods to study the antimicrobial activity of *Centella asiatica* leaf extracts. Int J Adv Res 7:344–347

Shen Y, Liu A, Ye M, Wang L, Chen J, Wang X, Han C (2009) Analysis of biologically active constituents in *Centella asiatica* by microwave-assisted extraction combined with LC–MS. Chromatographia 70:431–438

Shen X, Guo M, Yu H, Liu D, Lu Z, Lu Y (2019) Propionibacterium acnes related anti-inflammation and skin hydration activities of madecassoside, a pentacyclic triterpene saponin from *Centella asiatica* Biosci Biotechnol Biochem 83(3):561–568

Shibuya M, Hoshino M, Katsube Y, Hayashi H et al (2006) Identification of β-amyrin and sophoradiol 24-hydroxylase by expressed sequence tag mining and functional expression assay. FEBS J 273:948–959

Sirichoat A, Chaijaroonkhanarak W, Prachaney P, Pannangrong W, Leksomboon R, Chaichun A et al (2015) Effects of asiatic acid on spatialworking memory and cell proliferation in the adult rat hippocampus. Nutrients 7:8413–8423

Skrzypczak-Pietraszek E, Urbańska A, Żmudzki P, Pietraszek J (2019) Elicitation with methyl jasmonate combined with cultivation in the Plantform^{TM} temporary immersion bioreactor highly increases the accumulation of selected centellosides and phenolics in *Centella asiatica* (L.) Urban shoot culture. Eng Life Sci 19(12):931–943

Stafford GI, Pedersen ME, Van Staden J, Jager AK (2008) Review on plants with CNS-effects used in traditional South African medicine against mental diseases. J Ethnopharmacol 119:513–527

Sun B, Wu L, Wu Y, Zhang C, Qin L, Hayashi M, Liu T (2020) Therapeutic potential of *Centella asiatica* and its triterpenes: a review. Front Pharmacol 11:1373

Thamarai Selvi P, Senthil Kumar M, Rajesh R, Kathiravan T (2012) Antidepressant activity of ethanolic extract of leaves of *Centella asiatica*. Linn by in vivo methods. Asian J Res Pharm Sci 2:76–79

Tran N, Pham B, Le L (2020) Bioactive compounds in anti-diabetic plants: from herbal medicine to modern drug discovery. Biology 9(9):252

Vadlapudi V, Behara M, Kaladhar DSVGK, Suresh Kumar SVN, Seshagiri B, John Paul M (2012) Antimicrobial profile of crude extracts *Calotropis procera* and *Centella asiatica* against some important pathogens. Indian J Sci Technol 5:3132–3136

Vasconsuelo A, Boland R (2007) Molecular aspects of the early stages of elicitation of secondary metabolites in plants. Plant Sci 172:861–875

Wang C, Han W, Fan L, Wang C (2009) Enzymatic pretreatment and microwave extraction of asiaticoside from *Centella asiatica*. J Biomed Sci Eng 2:526–531

Wang Q, Yao L, Xu K, Jin H, Chen K, Wang Z, Xu J (2019) Madecassoside inhibits estrogen deficiency-induced osteoporosis by suppressing RANKL-induced osteoclastogenesis. J Cell Mol Med 23(1):380–394

Wong JH, Barron AM, Abdullah JM (2021) Mitoprotective effects of *Centella asiatica* (L.) Urb.: anti-inflammatory and neuroprotective opportunities in neurodegenerative disease. Front Pharmacol. https://doi.org/10.3389/fphar.2021.687935

Wong JX, Ramli S, Desa S, Chen SN (2021) Use of *Centella asiatica* extract in reducing microbial contamination and browning effect in fresh cut fruits and vegetables during storage: a potential alternative of synthetic preservatives. LWT 151:112229

Yoo NH, Kim OT, Kim JB et al (2011) Enhancement of centelloside production from cultured plants of *Centella asiatica* by combination of thidiazuron and methyl jasmonate. Plant Biotechnol Rep 5:283–287

Zahanis Z, Mansyurdin M, Zozy Aneloi Noli ZAN, Amri Bakhtiar AB (2016) Production of asiaticoside from hairy roots culture of pegagan (*Centella asiatica* (l.)) urban using chitosan and its derivates as elisitors. J Chem Pharm Res 8(2):808–812

Zaidan MR, Noor Rain A, Badrul AR, Adlin A, Norazah A, Zakiah I (2005) In vitro screening of five local medicinal plants for antibacterial activityusing disc diffusion method. Trop Biomed 22(2):165–170

Zhang QW, Lin LG, Ye WC (2018) Techniques for extraction and isolation of natural products: a comprehensive review. Chin Med 13:1–26

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

View publication stats

2021 8th International Conference on Signal Processing and Integrated Networks (SPIN) | 978-1-6654-3564-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SPIN52536.2021.9565960

Circularly Polarized Antenna for ISM (5.8 GHz), Satellite Communications and UWB Applications

Navneet Sharma
Department of Electronics and Communication Engineering
Shobhit Institute of Engineering & Technology, (Deemed to be University)
Meerut, India
navneet1979@gmail.com

Anubhav Kumar
Department of Electronics and Communication Engineering
Shobhit Institute of Engineering & Technology, (Deemed to be University)
Meerut, India
rajput.anubhav@gmail.com

Asok De
Department of Electronics and Communication Engineering
Delhi Technological University (DTU)
Delhi India
asok.de@gmail.com

Rakesh Kumar Jain
Department of Electronics and Communication Engineering
Shobhit Institute of Engineering & Technology, (Deemed to be University)
Meerut, India
rakesh.jain@shobhituniversity.ac.in

Abstract—A compact and dual-band circularly polarized antenna with resonant frequency of 5.8GHz(ISM) and 7.6 GHz is proposed for biomedical, satellite and specific UWB applications. The 10 dB impedance bandwidth (IBW) of the antenna varies from 5.55 GHz to 5.94 GHz and 6.78 GHz to 8.78 GHz. The tilted arc-shaped radiator is used to perturb the current which is responsible for lower frequency band as well as circular polarization with the axial ratio extending from 5.77 GHz to 5.93 GHz, which covers 41% of the lower frequency band. The antenna is analyzed for wearable applications on a three-layer skin phantom model and the SAR value obtained is 0.2 W/Kg with Source power of 10mW, which is below the maximum permissible limit of 1.6W/Kg.

Keywords— Dual Band, Circularly Polarized, ISM Band, Skin Phantom model, Specific Absorption Rate(SAR).

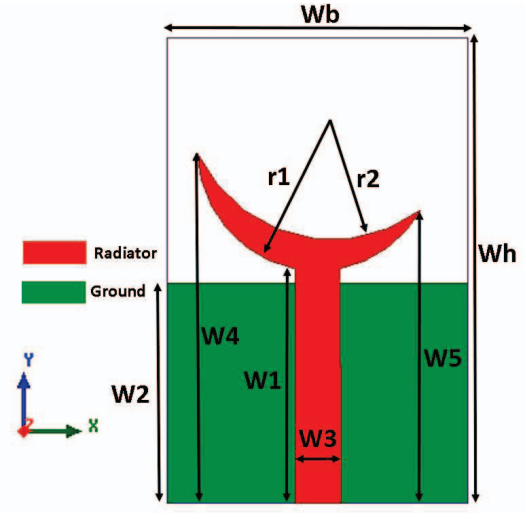
I. INTRODUCTION

Wearable biomedical devices are in high demand to regularly monitor the health parameters for sports, military, biomedical as well as in pandemic situations such as COVID-19. Compact, circularly polarized antennas are indispensable for wearable applications because of their small size, which can be embedded in the compact structure whereas it transmits or receives the EM waves in different planes.

In literature, research on circularly polarized antenna has been presented [1-9]. In [1], a semi-circled radiating patch with an L-shaped stub and slot in-ground is used to perturb the current and is responsible for circular polarization. In [2], inverted and tilted U-shaped with metallic strips are used to achieve CP. In [3], a rectangular patch is used with a tilted slot for surface current variation and to achieve circular polarization. In [4], a circularly polarized patch with a bandpass filter is used for ISM band application. In [5], U-shaped radiator with a rectangular slot in the ground is used to achieve the circular polarization in two bands. In [6], a rectangular patch with co-axial feed with a fractal-shaped slot in the ground is responsible for circular polarization is used. In [7], a square-shaped radiator with the slot is used to improve the impedance bandwidth as well as circular polarization. In [8], a rectangle patch with coaxial feed with the defected ground is used, where fractal-shaped defected ground and position of the feed is responsible for circular polarization. In [9], an H-shaped conducting strip with rectangular-shaped DRA is used to accomplish CP. In the proposed antenna, a tilted arc shaped radiator (TASR) with partial ground is used to achieve dual-band and CP. The antenna is a good choice to be used for wearable applications in 5.55 GHz to 5.94 GHz band, resonating at 5.8 GHz with circular polarization as well as 6.78 GHz to 8.78 GHz for satellite and specific UWB applications [10].

II. ANTENNA DESIGN

A compact and optimized geometry of CP antenna, designed on FR4 material (Dielectric Constant, $\varepsilon_r = 4.4$) is illustrated in Fig. 1(a) Overall dimensions(in mm) of dual band CP antenna is 30 x 20 x 1.6. The optimized dimensions of antenna are as follows: Wh=30mm, Wb=20 mm, W1=15 mm, W3 = 3 mm, W2= 14.2 mm, W5= 19 mm ,W4 = 22.6 mm, r2= 8 mm and r1= 10 mm. The antenna is excited through SMA port designed for 50 ohms input impedance to match the feed line.

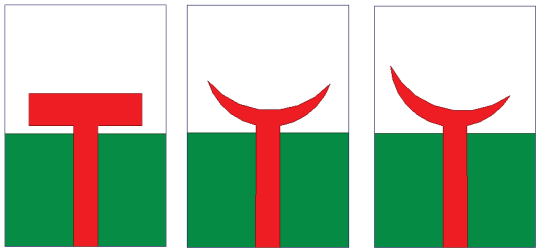


(a)

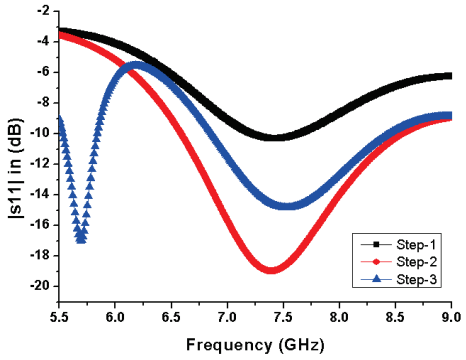
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:46:36 UTC from IEEE Xplore. Restrictions apply.

III. EVOLUTION OF CP ANTENNA

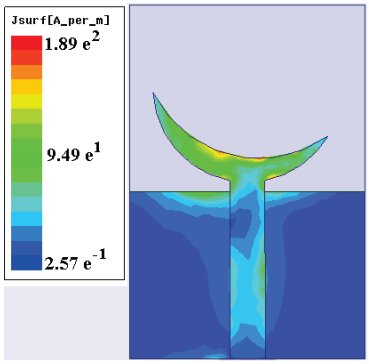
The proposed dual-band antenna is discussed in three steps as illustrated in Fig. 1(b) and s-parameters depicted in Fig. 1(c). In step-1, microstrip feed line with 50 ohms' impedance and 3 mm width is used with rectangle shaped radiator where partial ground is used to achieve 10 dB impedance bandwidth(IBW) that varies from 7.23 GHz to 7.61 GHz. The rectangular radiator is modified to an arc-shaped radiator to improve the impedance matching and enhances the bandwidth which varies from 6.59 GHz to 8.55 GHz in step-2, the height of both the edges of arc-shaped radiator are same. In step-3, arc-shaped radiator is tilted where the difference between ground plane and lower edge of radiator is 5.8 mm whereas it is 8.4 mm from upper edge. The tilted arc shaped radiator perturbs the surface current and is responsible for lower frequency band as well as 3 dB axial ratio. The antenna represents RHCP and LHCP waves which is analyzed in next section. The modified radiator changes the surface current variation, which provides the double resonant frequency, which has been illustrated in surface current at 5.8 GHz, and 7.6 GHz in Fig. 1(d) and Fig. 1(e). The 10 dB impedance bandwidth at this step exists from 5.55 GHz to 5.94 GHz as well as 6.78 GHz to 8.78 GHz



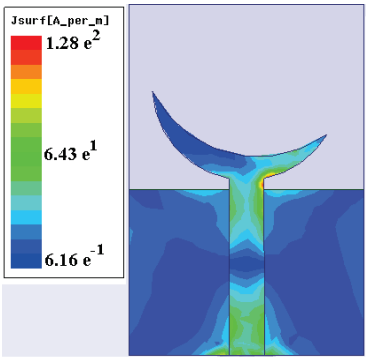
(b)



(c)



(d)



(e)

Fig.1 (a) Proposed CP Antenna, (b) Evolution Steps, (c) |S11| Parameters of Evolution Steps.(d) J Surf at 5.8 GHz (e) J Surf at 7.6 GHz.

IV. CP ANALYSIS

CP mechanism is studied by observing the time-varying electric field of 0°, 90°, 180°, and 270° phase at 5.8 GHz. On observing the current path, it is apparent that the antenna exhibits both Right Hand Circular Polarization(RHCP) and Left Hand Circular Polarization(LHCP) properties suggested by the movement of arrows in clockwise as well as the anti-clockwise direction in the ground plane and the radiator. These movements are depicted in Fig. 2.

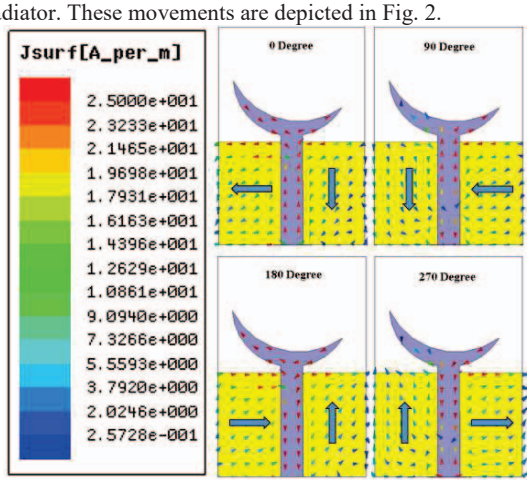


Fig.2. CP Analysis at 5.8 GHz

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:46:36 UTC from IEEE Xplore. Restrictions apply.

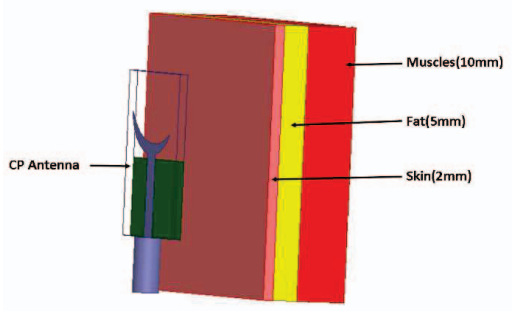
V. PHANTOM MODEL ANALYSIS FOR BIOMEDICAL APPLICATIONS

The proposed antenna is designed and is intended to be used for wearable applications, for practical considerations, a skin phantom model with three layers of Skin, Fat and Muscles is used with width 2mm, 5mm and 10 mm respectively as depicted in Fig. 3(a). As the parameters of body tissues such as relative permittivity(ε_0), Loss Tangent(σ) and Conductivity(S/m) changes, with change in operating frequency, relevant values of these parameters are taken to simulate the model. The values along with the resonant frequency is depicted in Table-1.

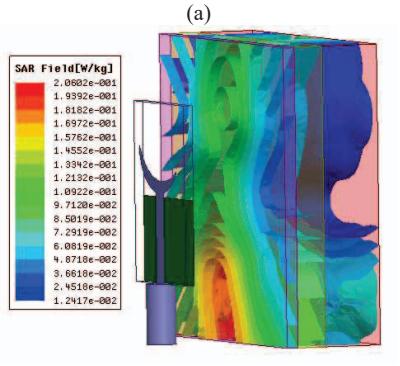
Table-1: Material properties of tissues in Skin Phantom Model

Freq	Tissue	Relative Permittivity(ε_0)	Loss Tangent(σ)	Conductivity) (S/m)
5.8 GHz	Skin	35.114	0.32807	3.717
	Fat	4.9549	0.18335	0.29313
	Muscles	48.485	0.31715	4.9615
7.6 GHz	Skin	33.548	0.38162	5.4129
	Fat	4.7959	0.20467	0.41502
	Muscles	46.045	0.37266	7.2549

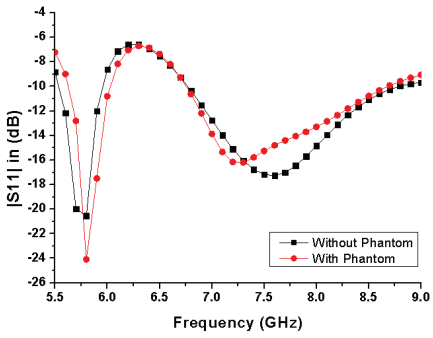
The skin phantom is positioned behind the antenna at a distance of 10 mm for SAR analysis illustrated in Fig. 3(b) where SAR value is observed 0.2 W/Kg with 10mW input power, and 1 g of human tissue, which is below the maximum recommended value of 1.6W/kg. All the parametric difference with and without placement of skin phantom is depicted in Fig. 3(c-e). It is evident that when the three layered skin phantom is positioned at a distance of 10 mm from the antenna, a gain enhancement of 20.8 % and 127 % is observed at 5.8 GHz and 7.6 GHz respectively. There is no shift at 5.8 GHz and there is marginal shift at the resonance of 7.6 GHz but the antenna maintains its bandwidth effectively. The radiation efficiency and realized gain are depicted in Fig. 3(d) and Fig. 3(e) with and without phantom, where efficiency reduces due to coupling between antenna and phantom model. The enhancement in realized gain is observed on phantom model with 10 mm distance, where phantom model controls the grating lobes and improves the gain at higher frequency.



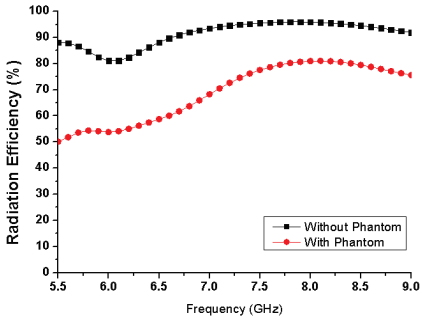
(a)



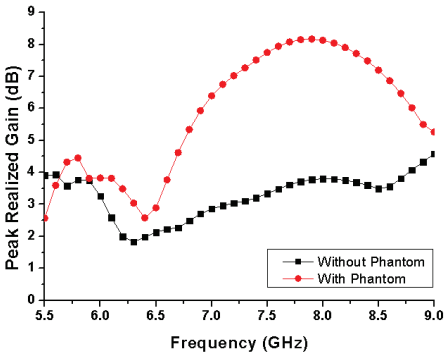
(b)



(c)



(d)



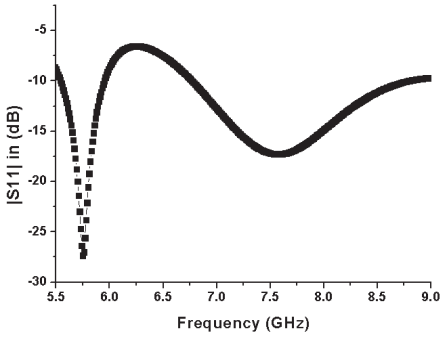
(e)

Fig. 3(a) Phantom model structure (b) SAR analysis at 10mW (c) Comparative reflection Coefficient(dB) (d) Comparative Efficiency(%) (e) Comparative realized gain(dB).

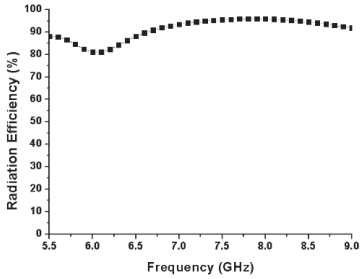
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:46:36 UTC from IEEE Xplore. Restrictions apply.

VI. RESULTS AND DISCUSSIONS

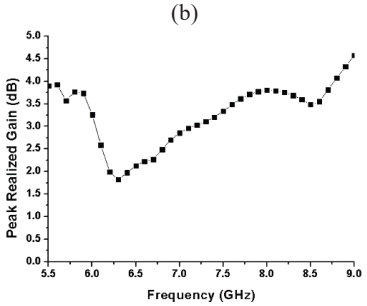
The proposed dual band CP antenna works effectively based on tilted arc-shaped radiator is analyzed in this section. The first operating band 10dB bandwidth(IBW) varies from 5.55 GHz to 5.94 GHz as depicted in Fig. 4(a) with axial ratio extending from 5.77 GHz to 5.93 GHz as given in Fig. 4(d). The radiation efficiency of the proposed antenna at this band is greater than 80% whereas realized gain is 3.93 dB as depicted in Fig. 4(b) and Fig. 4(c). The second band extends from 6.78 GHz to 8.55 GHz. The radiation efficiency is more than 95% with maximum gain of 4.57 dB in this band. The radiation patterns at 5.8 GHz and 7.6 GHz in xz and yz planes are depicted in Fig. 4(e) and Fig. 4(f), where stable radiation pattern is observed at both the bands. Table.2 shows the compared results of the existing CP work with the proposed antenna, where proposed antenna has compact size(20mm X 30mm), whereas all other related work has a much larger size, which may not be considered suitable for wearable applications. The bandwidth obtained is wider, with circular polarization is at 41% of the bandwidth still the antenna is able to maintain its gain of 3.93 and 4.97 at two bands, which is appreciable at such small size of antenna.



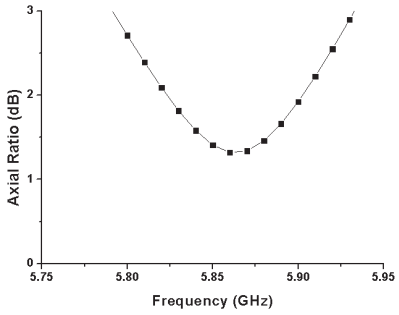
(a)



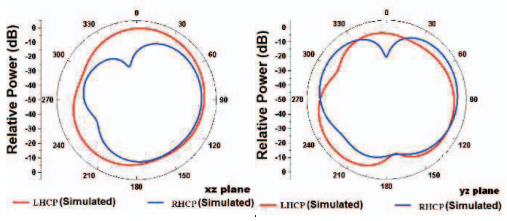
(b)



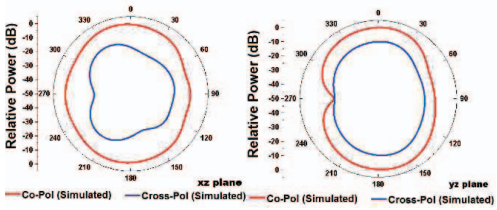
(c)



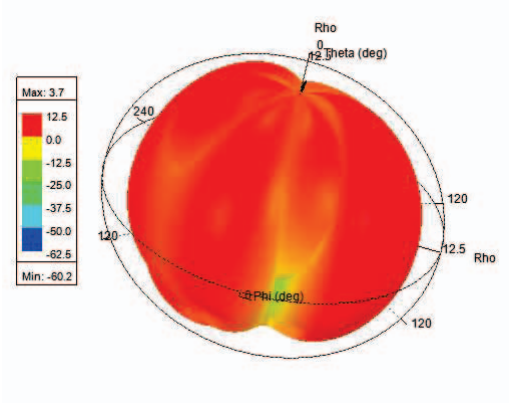
(d)



(e)



(f)



(g)

Fig.4(a) Reflection Coefficient(dB), (b) Radiation Efficiency (%), (c)Realized Gain(dB), (d) Axial ratio(dB),(e)Radiation Pattern LHCP-RHCP(5.8 GHz), (f) Radiation Pattern, xz and yz plane (7.6 GHz)(g) 3-D Radiation Pattern at 5.8 GHz.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:46:36 UTC from IEEE Xplore. Restrictions apply.

Table 2: Proposed antenna as compared to related published works

Ref	Size(λ_0)	Bandwidth (GHz)	Axial Ratio %	Gain(dB)	SAR(tissue is W/Kg)
[1]	0.29 X 0.21	2.4, 5.8	10,34	2, 1.3-4.5	NA
[2]	0.42 X 0.44	2.4,3.5,5.8	5.4,8.3,3.7	6.92	NA
[3]	0.46 X 0.26	2.45	23.3	1.8	NA
[4]	0.29 X 0.32	2.45	NA	3	NA
[5]	0.30 X 0.30	2.3 – 6.0	40.54 & 18.91	2.91 & 3.81	0.8 & 0.95
[6]	0.23 X 0.23	1.558-1.588	20	1.7-2.2	NA
[7]	0.24 X 0.24	0.961-1.204	31	5.2	NA
[8]	0.48 X 0.48	2.354–2.463	23.80	4.62	NA
Proposed	0.37 X 0.56	5.8, 7.6	41	3.93, 4.97	0.2

VII. CONCLUSION

A compact dual-band antenna is presented with circular polarization at 5.8 GHz (ISM), whereas satellite and specific UWB applications for the band from 6.78 GHz to 8.78 GHz. The tilted arc shape radiator perturbs the surface current and responsible for the dual operating band as well as circular polarization. It is observed that the antenna has good efficiency and exhibits stable radiation patterns in both bands. The antenna can be used for biomedical at ISM band, wearable applications, satellite communications as well as specific UWB applications.

ACKNOWLEDMENTS

The antenna is designed on Ansys HFSS EM wave simulator at the “Optical, Antenna and Microwave Engineering” facility at ABES Engineering College, Ghaziabad.

REFERENCES.

[1] Liang, W., Jiao, Y. C., Luan, Y., & Tian, C. "A dual-band circularly polarized complementary antenna." IEEE Antennas and Wireless Propagation Letters 14: 1153-1156, 2015

[2] Hoang, T. V., Le, T. T., Li, Q. Y., & Park, H. C. "Quad-band circularly polarized antenna for 2.4/5.3/5.8-GHz WLAN and 3.5-GHz WiMAX applications." IEEE antennas and wireless propagation letters 15: 1032-1035, 2015.

[3] Li, Jianxiong, Yuchen Jiang, and Xiaoming Zhao. "Circularly polarized wearable antenna based on NinjaFlex-embedded conductive fabric." International Journal of Antennas and Propagation, 2019

[4] Jiang, ZhiHao, Micah D. Gregory, and Douglas H. Werner. "Design and experimental investigation of a compact circularly polarized integrated filtering antenna for wearable biotelemetric devices." IEEE Transactions on Biomedical Circuits and Systems 10.2: 328-338, 2015

[5] Yadav, Ashok, Vinod Kumar Singh, and Himanshu Mohan. "Design of a U-shaped circularly polarized wearable antenna with DGS on a fabric substrate for WLAN and C-band applications." Journal of Computational Electronics 18.: 1103-1109, 2019

[6] Wei, Kun, J. Y. Li, L. Wang, R. Xu, and Z. J. Xing. "A new technique to design circularly polarized microstrip antenna by fractal defected ground structure." IEEE Transactions on Antennas and Propagation 65.7: 3721-3725, 2017

[7] Chen, Xi, Long Yang, Jia-yue Zhao, and Guang Fu. "High-efficiency compact circularly polarized microstrip antenna with wide beamwidth for airborne communication." IEEE Antennas and Wireless Propagation Letters 15: 1518-1521, 2016.

[8] Pandey, Shashi Kant, Ganga Prasad Pandey, and P. M. Sarun. "Circularly polarized micro-strip antenna with fractal trees loaded ground plane." Electromagnetics 39.7 : 505-523, 2019.

[9] Illahi, U., Iqbal, J., Sulaiman, M. I., Alam, M. M., Su'Ud, M. M., Jamaluddin, M. H., & Yasin, M. N. M. "Design of new circularly polarized wearable dielectric resonator antenna for off-body communication in WBAN applications." IEEE Access 7": 150573-150,582, 2019.

[10] Verma, Rahul Kumar, Anubhav Kumar, and R. L. Yadava. "Compact Multiband CPW Fed Sub 6 GHz Frequency Reconfigurable Antenna for 5G and Specific UWB Applications." J. Commun 15: 345-349, 2020.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:46:36 UTC from IEEE Xplore. Restrictions apply.

2021 8th International Conference on Signal Processing and Integrated Networks (SPIN) | 978-1-6654-3564-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SPIN52536.2021.9565976

Comparative Analysis of Non-Isolated DC-DC Converters for Solar-Photovoltaic System

Shivangi Yadav
Dept. of Electrical Engineering
Delhi Technological University
Delhi, India
yadavshivangi26@gmail.com

Saurabh Mishra
Dept. of Electrical Engineering
Delhi Technological University
Delhi, India
saurabh.mishra1410@gmail.com

Garima
Dept. of Electrical Engineering
Delhi Technological University
Delhi, India
garima.dtu@gmail.com

***Abstract*— On account of the continuous consumption of fossil fuels and its adverse effects on the environment, the renewable sources of energy are becoming increasingly popular. Hence, solar energy is one of the most reliable solution as it is available in abundance. A typical photovoltaic system comprises of the PV array/panel, DC-DC converters and the load. DC-DC converters are the power conditioning units which perform buck or boost operations in accordance to the requirement at the output. In this paper four DC-DC converter topologies i.e SEPIC, CUK, ZETA and Landsman have been implemented with a PV array rating of 250 Watts. The simulation of these PV based system involving 4 different converter topologies has been performed in MATLAB/Simulink environment. Perturb and Observe algorithm has been implemented for derivation of maximum power from the PV array/panel. The result obtained help us to compare the performance of different topologies when MPPT is applied.**

Keywords*— Zeta, CUK, Single Ended Primary Inductor Converter (SEPIC), Landsman, Perturb and Observe algorithm (P&O).

I. INTRODUCTION

The demand for electricity has risen dramatically in recent years and also the conventional sources of energy are depleting at a very fast pace due to the increased consumption and limited supply, thereby making power production a challenge. Owing to this, the popularity of the renewable sources of energy has increased manifold. Among all the available sources of energy solar energy has emerged to be the most suitable as it is universally available, eco-friendly, cheaper when compared to other renewable sources.

The PV cell is a non-linear source of electrical energy therefore, it is required that we extract maximum possible power from them even in varying atmospheric conditions precisely varying temperature and irradiance [1]. With regard to the derivation of maximum power from the PV array, a substantial number of MPPT techniques have been suggested, some of them have been discussed in [2]. Among all these MPPT techniques, P&O has gained popularity because of its easy implementation and simplicity.

The characteristics of the PV array show that it has a low output dc voltage therefore, they are not suitable to be used in applications like the microgrids where there is requirement of a higher voltage. To resolve this problem and to obtain a higher voltage, PV modules are linked in series and parallel. Connecting the PV modules in large numbers require huge amount of physical space hence, to find a solution for this shortcoming, the photovoltaic arrays are interfaced with DC-DC converters which are designed for boosting the voltage to a higher value depending on the demand of the application. DC-DC converters that are frequently used with photovoltaic systems are boost, buck-boost, ZETA, CUK, Luo, Landsman, SEPIC etc. Comparative study of the DC-DC converters that are commonly being used nowadays in Photovoltaic Systems have been discussed in [3]-[6]. A detailed description of a wide variety of MPPT techniques that exist today for implementation to DC-DC converters in view to improve photovoltaic system's efficiency are discussed in [7].

This paper is organised in the following manner. Section II and III detail the circuit design and modelling of the PV array. Section IV describes the dc -dc converter topologies used. The control algorithm used has been briefly discussed in section V. At last, in Section VI simulation results are shown followed by conclusion in Section VII.

II. CIRCUIT CONFIGURATION

The Photovoltaic system comprises of the following components: PV array for the purpose of converting solar energy into electrical energy, DC-DC converters are utilized to optimize the obtained power, extracted from the solar PV array, Maximum Power Point Tracking (MPPT) algorithm block for providing suitable duty cycle to the DC-DC converter, to attain maximum amount of power from the PV array at all atmospheric conditions and lastly it consists of the load, as given in Fig.1. The load in this paper is considered to be resistive.

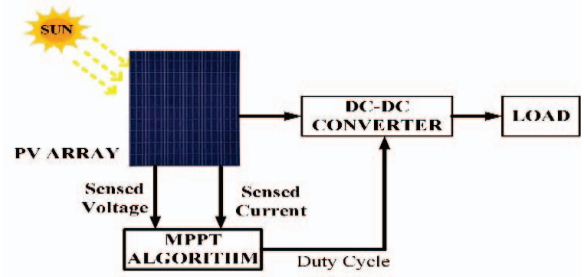


Fig.1 Block Representation [4]

III. PV ARRAY MODELLING

The PV array module that has been interfaced with each converter has been used from MATLAB/SIMULINK environment. The PV module that has been chosen for the purpose of modelling is **1Soltech-1STH-250-WH**. The specified ratings of the selected module have been given in table I. Table II presents the different irradiance values. Fig.2 below shows the P-V and I-V waveforms of PV array taken in the system.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:42:29 UTC from IEEE Xplore. Restrictions apply.

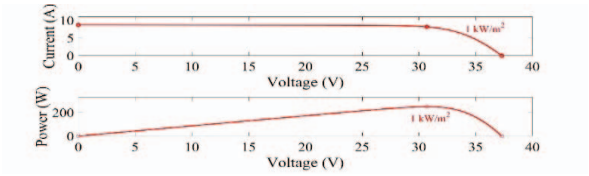


Fig.2 I-V and P-V curves for PV array [1]

TABLE I. The PV ARRAY MODULE PARAMETERS

DESCRIPTION	RATING
Maximum Power (P_{max})	**250.205 Watts**
Voltage at P_{max} (V_{mp})	**30.7V**
Current at P_{max} (I_{mp})	**8.15A**
Open-Circuit Voltage (V_{oc})	**37.3V**
Short-Circuit current (I_{sc})	**8.66A**
Number of PV Cells in series	**4**
Number of PV Cells in parallel	**1**

TABLE II. IRRADIANCE VALUES

IRRADIANCE	TIME
800 Watts per meter square	**0-0.3 sec**
1000 Watts per meter square	**0.3-0.7 sec**
700 Watts per meter square	**0.7-1 sec**

IV. CONVERTER TOPOLOGY

In this paper we have studied four DC-DC converter topologies that is SEPIC, CUK, ZETA and Landsman converter. The converter topology diagrams have been given in Fig.3 to Fig.6 respectively. The equations used for designing their parameters have been described below. The variables that have been used are: P_{mp} is the maximum power,V_0 is output voltage, D is duty ratio, f is the frequency, V_{in} is input voltage, V_{C1} is the capacitor C_1 voltage, V_{C2} is the capacitor C_2 voltage, I_{L1} is the inductor current, I_{L2} is the inductor current, ΔI_{L1} and ΔI_{L2} are the ripples in inductor currents and ΔV_{C1}, ΔV_{C2} and ΔV_0 are the ripples in voltages.

A. SEPIC Converter

The circuit topology of a SEPIC converter interfaced to the PV array along with P&O algorithm implemented is given below:

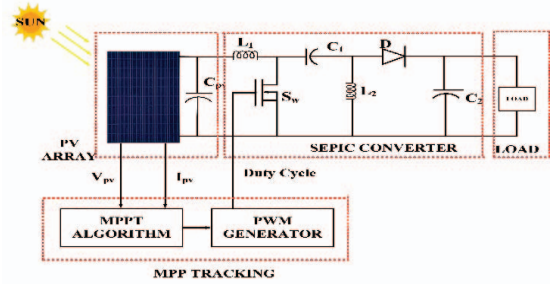


Fig.3 SEPIC Converter interfaced to PV array [4]

The procedure and equations to find the value of L_1, L_2, C_1, C_2 are:

$$P_{mp} = 1000W \quad (1)$$

$$V_{in} = 120V \quad (2)$$

$$V_0 = 40V \quad (3)$$

$$D = \frac{V_0}{V_0 + V_{in}} = \frac{40}{40+120} = 0.25 \quad (4)$$

$$I_{L_1} = \frac{P_{mp}}{V_{in}} = \frac{1000}{120} = 8.33A \quad (5)$$

$$I_{L_2} = \frac{P_{mp}}{V_0} = \frac{1000}{40} = 25A \quad (6)$$

$$\frac{\Delta V_0}{V_0} = 2\% \quad (7)$$

$$\Delta I_{L_1} = 40\% \; of \; I_{L_1} = 0.4*8.33 = 3.332A \quad (8)$$

$$\Delta I_{L_2} = 40\% \text{ of } I_{L_2} = 0.4*25 = 10A \quad (9)$$

$$R = \frac{V_0^2}{P_{mp}} = \frac{40*40}{1000} = 1.6\Omega \quad (10)$$

$$L_1 = \frac{V_{in}*D}{\Delta I_{L_1}*f} = \frac{120*0.25}{3.332*5000} = 1.8mH \quad (11)$$

$$L_2 = \frac{V_{in}*D}{\Delta I_{L_2}*f} = \frac{120*0.25}{10*5000} = 0.6mH \quad (12)$$

$$C_1 = C_2 = \frac{D}{R*\frac{\Delta V_0}{V_0}*f} = \frac{0.25}{1.6*0.02*5000} = 1.5625mF \quad (13)$$

B. CUK Converter

The following is the circuit architecture of the CUK converter with P&O algorithm implemented and interfaced to the array.

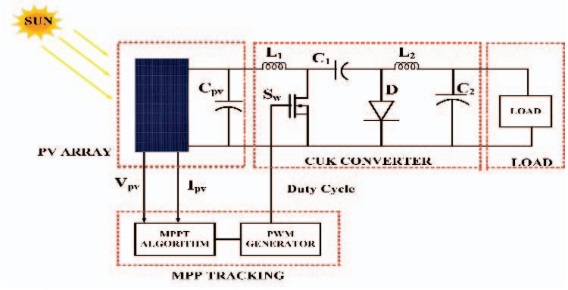


Fig.4 CUK Converter interfaced to PV array [5]

The design equations for determining the parameters of the converter are :

$$V_{in} = 120V \quad (14)$$

$$V_0 = 40V \quad (15)$$

$$P_{mp} = 1000W \quad (16)$$

$$D = \frac{V_0}{V_0 + V_{in}} = \frac{40}{40+120} = 0.25 \quad (17)$$

$$I_{L_1} = \frac{P_0}{V_{in}} = \frac{1000}{120} = 8.33A \quad (18)$$

$$\Delta I_{L_1} = 10\% \text{ of } I_{L_1} = 0.1*8.33 = 0.833A \quad (20)$$

$$\Delta I_{L_2} = 10\% \text{ of } I_{L_2} = 0.1*25 = 2.5A \quad (21)$$

$$V_{C_1} = V_{in} - V_0 = 120 + 40 = 160V \quad (22)$$

$$\Delta V_{C_1} = 5\% \text{ of } V_{C_1} = 0.05*160 = 8V \quad (23)$$

$$\frac{\Delta V_0}{V_0} = 1\% \quad (24)$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:42:29 UTC from IEEE Xplore. Restrictions apply.

$$L_1 = \frac{V_{in} * D}{f * \Delta I_{L_1}} = \frac{120 * 0.25}{5000 * 0.833} = 7.2mH \quad (25)$$

$$L_2 = \frac{V_{in} * D}{f * \Delta I_{L_2}} = \frac{120 * 0.25}{5000 * 2.5} = 2.4mH \quad (26)$$

$$C_1 = \frac{V_0 * D}{R * f * \Delta V_{C_1}} = \frac{40 * 0.25}{1.6 * 5000 * 8} = 156.25\mu F \quad (27)$$

$$C_2 = \frac{(1-D)}{\frac{\Delta V_0}{V_0} * 8 * L_2 * f^2} = \frac{1-0.25}{0.01 * 8 * 2.4 * 10^{-3} * 25 * 10^6} = 156.25\mu F \quad (28)$$

C. Zeta Converter

The circuit configuration of the ZETA converter interfaced with PV array and P&O MPPT algorithm implemented to it is given below:

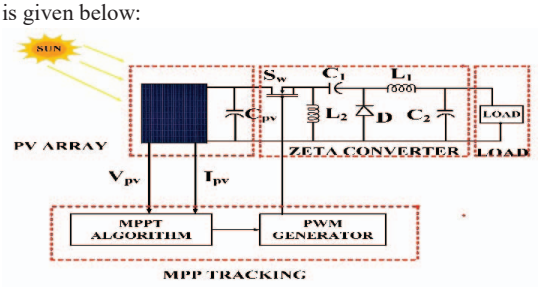


Fig.5 ZETA Converter interfaced to the PV array [4]

The equations for determining the parameters L_1, L_2, C_1, C_2 are given below:

$$P_{mp} = 1000W \quad (29)$$

$$V_{in} = 120V \quad (30)$$

$$V_0 = 40V \quad (31)$$

$$D = \frac{V_0}{V_0 + V_{in}} = \frac{40}{40+120} = 0.25 \quad (32)$$

$$I_{L_1} = \frac{P_{mp}}{V_{in}} = \frac{1000}{120} = 8.33A \quad (33)$$

$$I_{L_2} = \frac{P_{mp}}{V_0} = \frac{1000}{40} = 25A \quad (34)$$

$$\Delta I_{L_1} = 6\% \text{ of } I_{L_1} = 0.06 * 8.33 = 0.5A \quad (35)$$

$$\Delta I_{L_2} = 6\% \text{ of } I_{L_2} = 0.06 * 25 = 1.5A \quad (36)$$

$$V_{C_1} = V_0 = 40V \quad (37)$$

$$\Delta V_{C_1} = 10\% \text{ of } V_{C_1} = 0.01 * 40 = 4V \quad (38)$$

$$L_1 = \frac{D * V_{in}}{f * \Delta I_{L_1}} = \frac{0.25 * 120}{5000 * 0.5} = 12mH \quad (39)$$

$$L_2 = \frac{V_0 * (1-D)}{f * \Delta I_{L_2}} = \frac{40 * (1-0.25)}{5000 * 1.5} = 4mH \quad (40)$$

$$C_1 = \frac{D * I_0}{f * \Delta V_{C_1}} = \frac{0.25 * 25}{5000 * 4} = 312.5\mu F \quad (41)$$

$$C_2 = \frac{I_0}{6 * 2 * \pi * f * \Delta V_{C_1}} = \frac{25}{6 * 2 * \pi * 5000 * 4} = 33.15\mu F \quad (42)$$

D. Landsman Converter

The circuit topology of the Landsman converter that has been interfaced to the PV array and P&O algorithm implemented to it is given below:

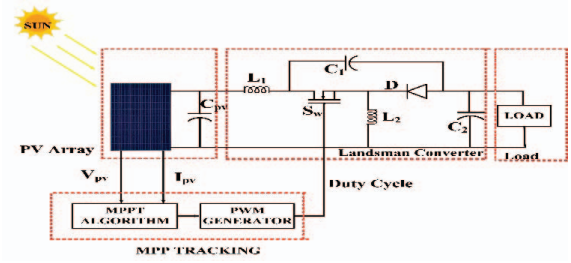


Fig.6 Landsman converter interfaced to the PV array [6]

The design equations for determining the parameters L1, L2, C1, C2 of the Landsman converter are:

$$P_{mp} = 1000W \quad (42)$$

$$V_{in} = 120V \quad (43)$$

$$V_0 = 40V \quad (44)$$

$$D = \frac{V_0}{V_0 + V_{in}} = \frac{40}{40+120} = 0.25 \quad (45)$$

$$V_{C_1} = V_{in} + V_0 = 120 + 40 = 160V \quad (46)$$

$$I_0 = \frac{P_{mp}}{V_0} = \frac{1000}{40} = 25A \quad (47)$$

$$\Delta V_{C_1} = 20\% \text{ of } V_{C_1} = 0.2 * 160 = 32V \quad (48)$$

$$I_L = I_0 + I_{in} = 25 + 8.33 = 33.33A \quad (49)$$

$$I_{L_1} = I_0 = 25A \quad (50)$$

$$\Delta I_L = 3\% \text{ of } I_L = 0.03 * 33.33 = 0.99A \quad (51)$$

$$\Delta I_{L_1} = 3\% \text{ of } I_{L_1} = 0.03 * 25 = 0.75A \quad (52)$$

$$\Delta V_0 = 4\% \text{ of } V_0 = 0.04 * 40 = 1.6V \quad (53)$$

$$C_1 = \frac{D * I_0}{f * \Delta V_{C_1}} = \frac{0.25 * 25}{5000 * 32} = 40\mu F \quad (54)$$

$$L_1 = \frac{D * I_0}{8 * f^2 * C_1 * \Delta I_{L_1}} = \frac{0.25 * 25}{8 * 25 * 10^6 * 40 * 10^{-6} * 0.75} = 1.04mH \quad (55)$$

$$L_2 = \frac{D * V_{in}}{f * \Delta I_L} = \frac{0.25 * 120}{5000 * 0.99} = 6.06mH \quad (56)$$

$$C = \frac{I_0}{6 * 2 * \pi * f * \Delta V_0} = \frac{25}{6 * 2 * \pi * 5000 * 1.6} = 82.89\mu F \quad (57)$$

V. CONTROL ALGORITHM

Perturb and Observe algorithm is being used very commonly with PV systems, for obtaining maximum power and this is because of its simplicity and its easiness in implementation procedure [2]. In P&O the voltage at the output of PV array is sensed first and then varied by a small value. This variation in the PV array voltage causes the output power also to change. This output power from the PV array is noted and it is compared at successive intervals. If after the perturbations in the output voltage, the output power is observed to increase then the output voltage is perturbed further in the same direction. This process of varying the output voltage is continued until the power difference is positive, which indicates that the output power is approaching the maximum value. In case the power difference results in a negative value, the direction of perturbation of output voltage is reversed. There are steady state oscillations in case of Perturb and Observe algorithm because of the continuous change in direction of perturbation to maintain the power at maximum value.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:42:29 UTC from IEEE Xplore. Restrictions apply.

The flowchart in Fig.7 presents the working of P&O.

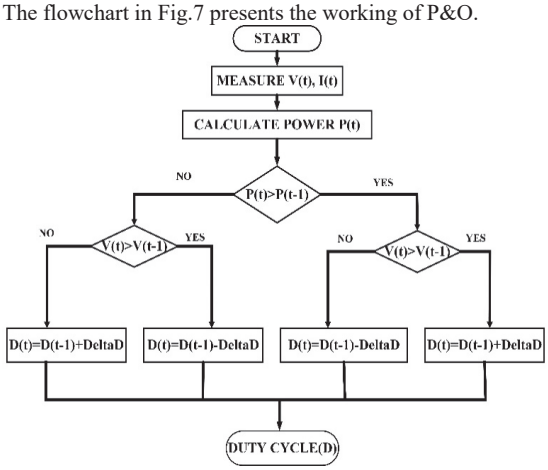
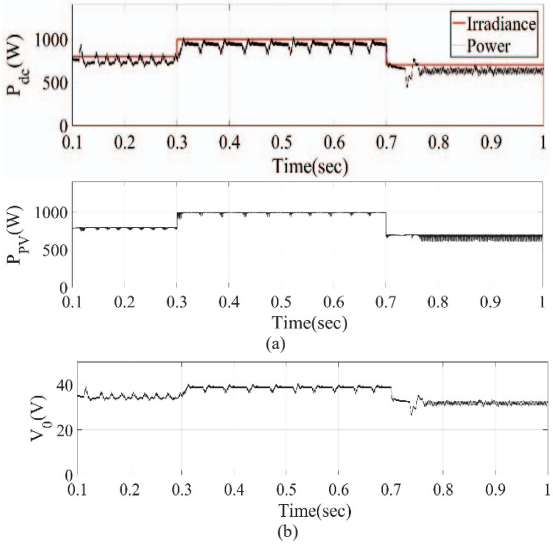


Fig.7 Flowchart for Perturb and Observe algorithm [7]

VI. SIMUTATION RESULTS

The waveforms obtained after simulation of four DC-DC converters interfaced to the PV array when perturb and Observe algorithm has been implemented for MPP tracking are given below from Fig.8 to Fig.11 respectively. The irradiance value has been varied in the range of 700 Watts/meter^2 to 1000 Watts/meter^2. The voltage is stepped down by the DC-DC converter from 120 V to 40V at a duty ratio of 0.25. The irradiance value changes at 0.3 sec and 0.7 sec where the waveforms of different elements of the converters show a transition which indicates that the MPPT algorithm is operating to reach the new maximum power point value. Converters that result in the same polarity of the output voltage as the input voltage are preferred. SEPIC and Zeta converter provide the same polarity whereas CUK and landsman provide inverted voltage at the output. Step size chosen for P&O can be varied to improve results. There are oscillations in all the waveforms as seen in Fig.8 to Fig.11.



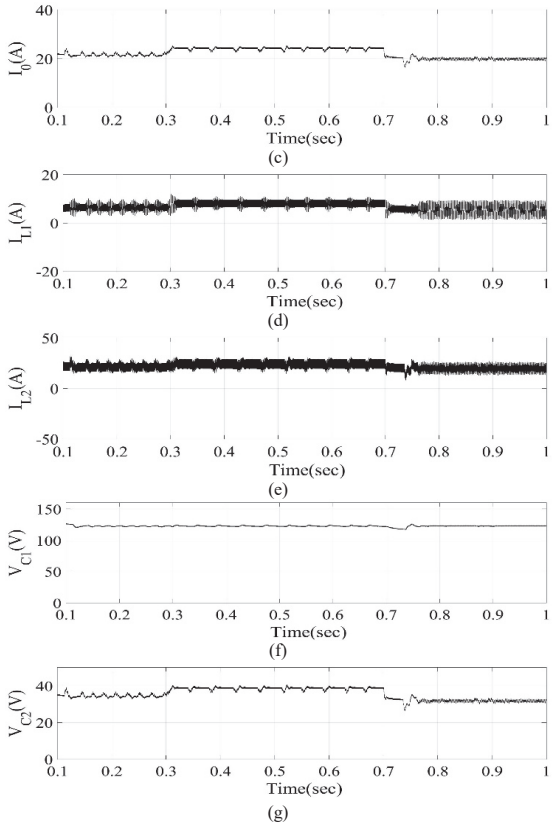
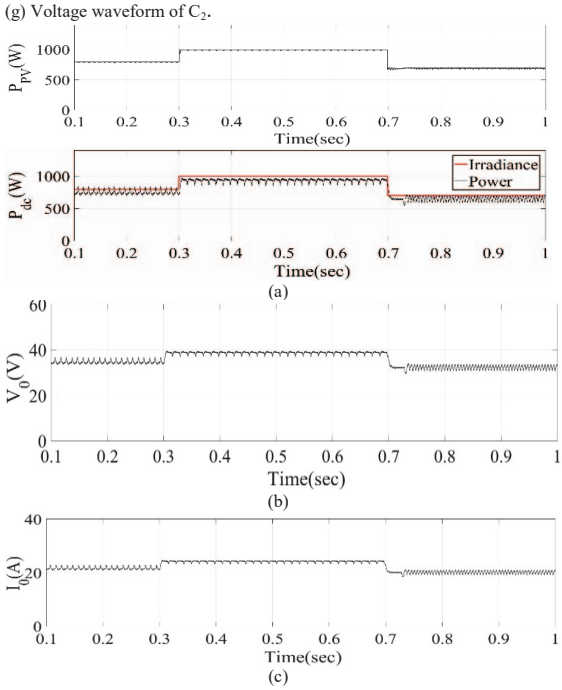


Fig.8 (a) Power waveform of SEPIC converter, (b)SEPIC converter's Output Voltage, (c)SEPIC converter's Output Current, (d) Current waveform of L_1, (e) Current waveform of L_2, (f) Voltage waveform of C_1, (g) Voltage waveform of C_2.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:42:29 UTC from IEEE Xplore. Restrictions apply.

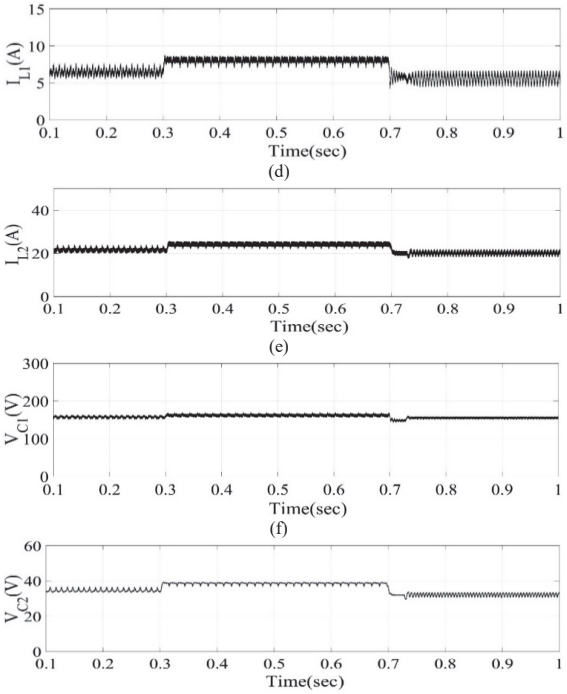
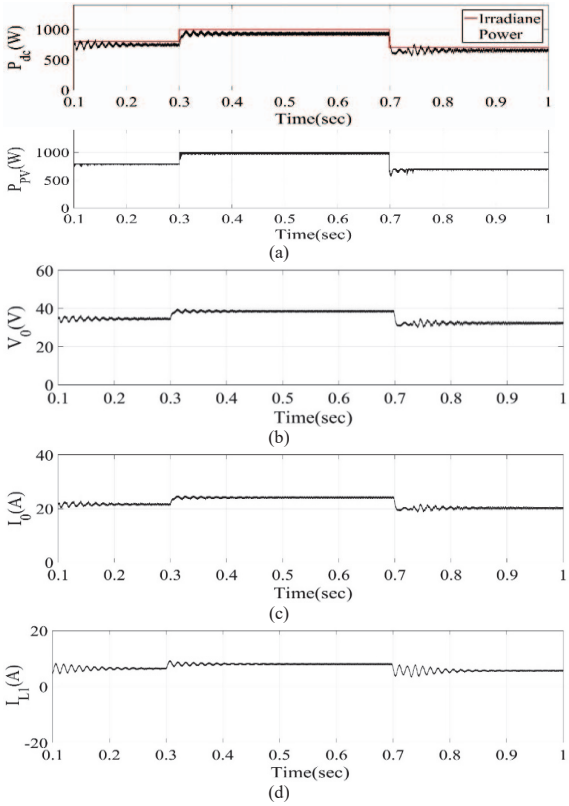


Fig.9. (a) Power waveform of CUK converter, (b)CUK converter's Output Voltage, (c)CUK converter's Output Current, (d) Current waveform of L_1, (e) Current waveform of L_2, (f) Voltage waveform of C_1, (g) Voltage waveform of C_2.



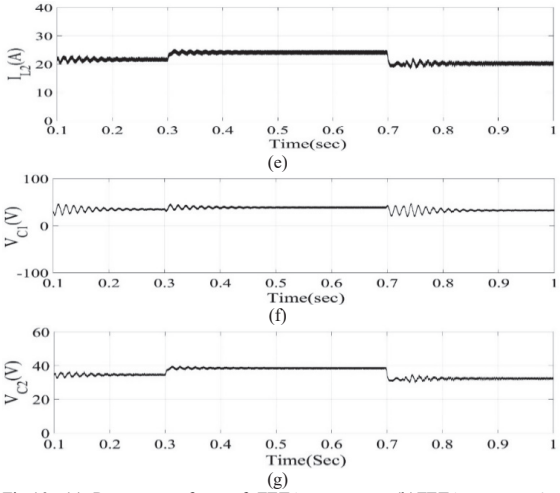
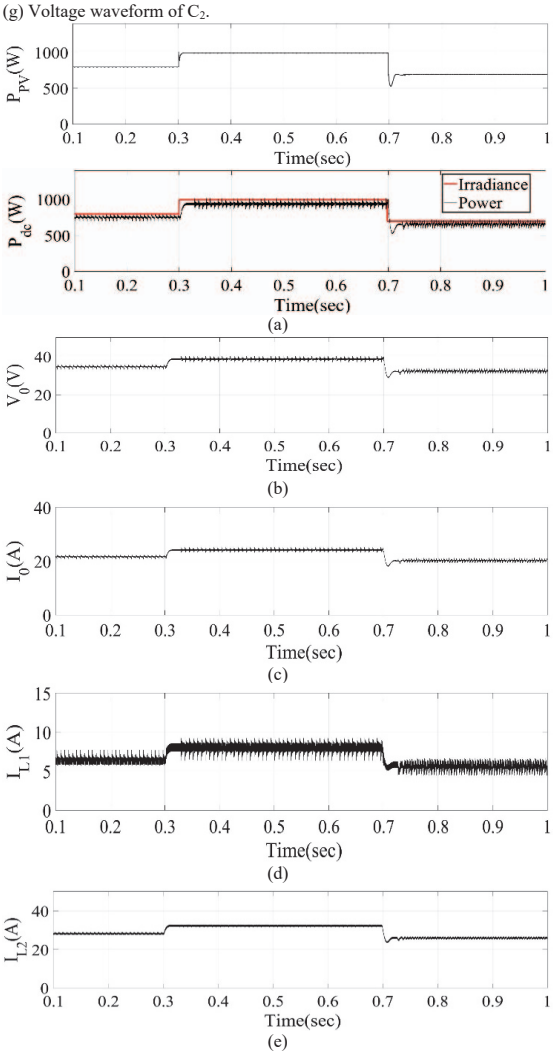


Fig.10. (a) Power waveform of ZETA converter, (b)ZETA converter's Output Voltage, (c)ZETA converter's Output Current, (d) Current waveform of L_1, (e) Current waveform of L_2, (f) Voltage waveform of C_1, (g) Voltage waveform of C_2.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:42:29 UTC from IEEE Xplore. Restrictions apply.

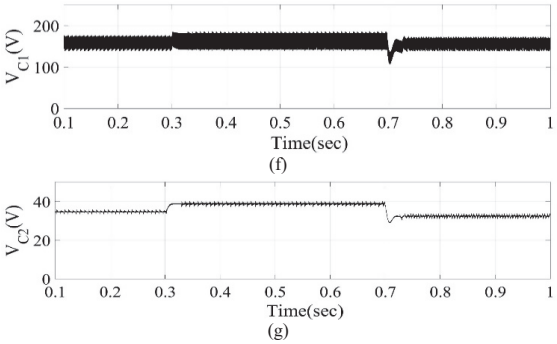
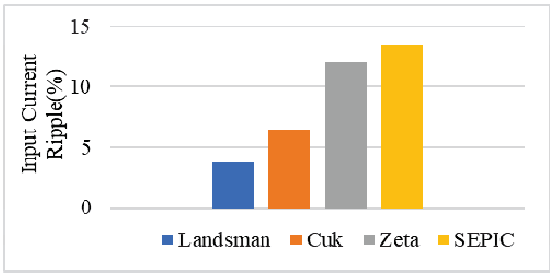
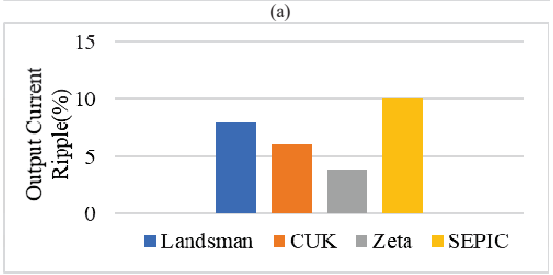


Fig.11. (a) Power waveform for LANDSMAN converter,(b)LANDSMAN converter's Output Voltage, (c)LANDSMAN converter's Output Current, (d) Current waveform of L_1, (e) Current waveform of L_2, (f) Voltage waveform of C_1, (g) Voltage waveform of C_2.

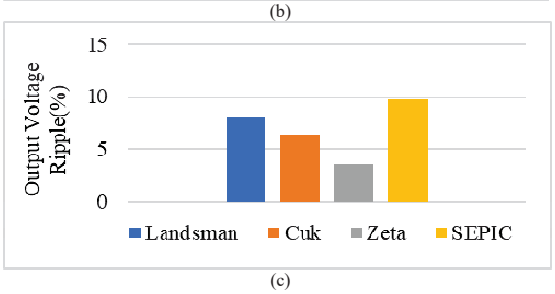
Fig.12(a) to Fig.12(c) shows the ripple percentage present in the waveforms of current and voltage for the four converters at an irradiance of 1000W/m^2. Fig.12(d) shows the efficiency of the converters with respect to different irradiances.



(a)



(b)



(c)

Fig.12(a) Input Current ripple for the converters, (b) Output Current Ripple for the converters, (c) Output Voltage Ripple.

As shown in Fig.12(b) and Fig.12(c), the output voltage and current waveforms have the least amount of ripple in case of Zeta Converter. Zeta converter interfaced to the PV array can be used for different solar applications owing to its response with lesser ripple.

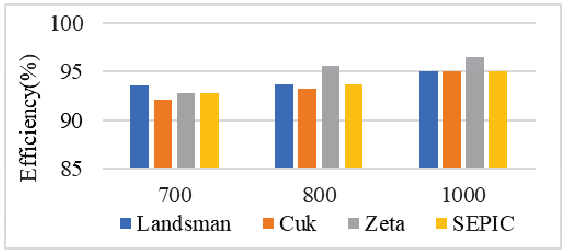


Fig. 13 Efficiency of the converters at different irradiance levels.

Fig.13 compares the efficiency of the four converters at three different irradiance levels. Zeta converter shows a consistent efficiency even when the irradiance levels are changed. Other three converters with the implementation of control (P&O) algorithm successfully extract the maximum possible power from PV array as they present a high efficiency in the above figure.

VII. CONCLUSION

In this paper, the designing and performance analysis of four different converter topologies that is SEPIC, CUK, ZETA, and Landsman have been considered. The equations for the design of the non-isolated converters have been mentioned in the paper. To extract maximum power, in case of variable parameters like the irradiance, a perturb and observe(P&O) maximum power point tracking method is implemented. The system has been tested for variable irradiance in the range of 700 watts/meter^2 to 1000 watts/ meter^2. MATLAB/Simulink tool has been used to simulate the system.

REFERENCES

[1] L. A. Soriano, P. Ponce and A. Molina, "Analysis of DC-DC converters for photovoltaic applications based on conventional MPPT algorithms," *14th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE),* Mexico City, pp. 1-6,2017.

[2] U. Sharma, B. Singh and S. Kumar, "A smart solar water pumping system with bidirectional power flow capabilities," *IEEE 7th Power India International Conference (PIICON)*, Bikaner, pp. 1-6,2016.

[3] S. Murali, K. P. Panda and G. Panda, "PV-HESS fed BLDC Driven Water Pumping System with PSO-based MPP Tracking Employing Zeta Converter," *IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia)*, Singapore, pp. 196-201,2018.

[4] Rashmi, Manohar J. and Rajesh K.S., "A comparative study and performance analysis of synchronous SEPIC Converter and synchronous Zeta Converter by using PV system with MPPT technique," *IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)*, Delhi, pp. 1-6,2016.

[5] M. Kumar Chaurasia and L. Gidwani, "Analysis of PV module with buck-boost and zeta converters with incremental conductance maximum power point tracking method," *International Conference On Smart Technologies For Smart Nation (SmartTechCon),* Bengaluru, India, 2017, pp. 1591-1596,2017.

[6] Singh, Bhim & Kumar, Rajan. "Solar photovoltaic array fed water pump driven by brushless DC motor using Landsman converter*," IET Renewable Power Generation*,2015.

[7] T. M. Chung, H. Daniyal, M. H. Sulaiman and M. S. Bakar, "Comparative study of P&O and modified incremental conductance algorithm in solar maximum power point tracking," *4th IET Clean Energy and Technology Conference (CEAT 2016)*, Kuala Lumpur, pp. 1-6,2016.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:42:29 UTC from IEEE Xplore. Restrictions apply.

2021 8th International Conference on Signal Processing and Integrated Networks (SPIN) | 978-1-6654-3564-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SPIN52536.2021.9566018

Comparative Analysis of Image Fusion Techniques for Infrared and Visible Images

Urvashi Rawat
Dept.of Electronics and Communication
Delhi Technological University
Delhi, India
urvashidpsgv99@gmail.com

Sudipta Majumdar
Dept.of Electronics and Communication
Delhi Technological University
Delhi, India
korsudipta@rediffmail.com

***Abstract*— Infrared Images have the ability to detect a target in low visibility or poor weather conditions. On the contrary, visible images reflect the background and has good spatial resolution. Fusion of these two images will yield a robust image having the advantages of structural details in visible image and thermal radiation in infrared image. In this paper four aspects are covered: comprehensive analysis of infrared and visible image fusion in spatial and transform domain, image fusion rules, application of infrared and visible images and performance criteria.**

***Index Terms:* Infrared image, visible image, image fusion, performance criteria.**

I. INTRODUCTION

Image fusion[1](IF)is a technique which aims at combing and enhancing the input images and generating an output image which is more informative and robust than any of the individual input images. In this method, images are obtained from different sensors of the same target and then the redundant information is removed and complementary information is integrated. In this way the resulting image becomes more suitable for human perception. Effective extraction of information and appropriate fusion algorithm are key to good fusion method[2] so that the important information from the input images is integrated without any artifacts.

There are various types of images that can be used for fusion, including visible images[1] (VIS), infrared images (IR), panchromatic images, computed tomography (CT), and magnetic resonance imaging (MRI). Out of the mentioned image sources, IR and VIS image fusion domain is quite superior in certain areas. They provide scene information with different aspects. The IR and VIS image comes from different modalities i.e the former captures the thermal radiation whereas the latter captures the target's reflected light. Fusing them gives us single modality image which integrates the complementary properties of both image type as shown in fig. 1. Also, IR and VIS images are acquired by simple equipments as compared to the medical images (CT and MRI).

There are several ways in which fusion schemes can be classified. Broadly they can be classified into spatial domain [3] and transform domain techniques. In spatial domain, the fusion algorithm is applied directly to the pixels of input IR and VIS images. This method is computationally simpler but lacks spatial information and output images suffer from distortion, contrast loss and also information loss. The input VIS and IR images are first divided into low frequency and high frequency components in the transform domain. The low frequency components provide us the details of image while the high frequency components provide us with line and edge information of the image.

With the advancement of technology, more visual comprehension and reliability is demanded from the fusion algorithms. Extensive research has been done in the field of IR and VIS image fusion[4]. This paper aims at comprehensively reviewing the existing algorithms for fusion since IR and VIS fusion approaches are given less priority in current reviews. The sections of this paper provides complete information about the various fusion methods, applications, fusion rules and performance criteria for IR and VIS image. Finally some experimental analysis has been done in order to compare typically used fusion methods and compared them on subjective and objective criteria.

The rest part of the paper is consisting of the following sections. A quick introduction to image fusion is given in section I. The background of IR and VIS images, as well as their applications, are described in Section II. The section III discusses the fusion methodologies in which fusion methods and fusion rules are reviewed briefly. Section IV contains the experimental analysis and the performance criteria. And section V concludes the paper followed by references.

II. BACKGROUND

A. Infrared and Visible Image Fusion

IR images are those images which uses the IR radiations radiated by ta body. The IR sensor senses the temperature of the target. If the temperature of the body is higher then we receive a strong IR spectrum, a brighter IR image and also a clear image[1]. It has a good target detection ability. But the demerit of IR image is that it has a low target resolution and a fuzzy background. Also less details of the target are present. VIS images are those which uses the visible light reflected from the target and captured by a VIS light image sensor. These images have high spatial resolution and good contrast information. It also reflects the background. Edge details and structure details are present. But in case of low visibility conditions or poor weather conditions, the image quality becomes poor.

The low frequency coefficients of an image also called as approximate coefficients[13] contain the major information of an IF and VIS image. Most of the energy is concentrated in the lower frequency sub-bands. Whereas, the high frequency coefficients also called detailed coefficients contain the detailed information of the IR and VIS images. This detailed information affects the clarity and resolution of the image. There have been many fusion rules which have been adopted for fusion of IR images with VIS images. In this paper we have discussed various techniques of fusion and compared them on

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:38:17 UTC from IEEE Xplore. Restrictions apply.

different parameters to conclude which methods are better for IR and VIS images.

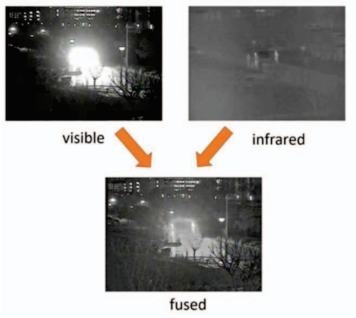


Fig. 2. Fig. 1: Fusion of IR and VIS image

Fig. 1.

B. Applications of Infrared and Visible Images

IR and VIS image fusion can be applied depending upon the requirement. Some of the applications are listed below:

1. Surveillance: For the surveillance [5,6] application we require a clear and expressive image. IR and VIS images have complementary properties which make them a good choice for fusion. The IR image detects the thermal radiation of the object and thus it is not affected by low visibility or bad weather conditions. However, the disadvantage of IR image is that it has low spatial resolution which causes less details in the image. On the contrary, the VIS images have high spatial resolution but easily affected by low visibility and weather disturbances. Thus fusion of these images improves visibility and help in surveillance application.

2. Recognition: Recognition is the method in which we specify the category of a particular target in an image which contains more than one objects. One of the major application in the recognition using IR and VIS image is Face Recognition[7].

3. Detection and Tracking: IR and VIS-based fusion is employed in a variety of real-world detection applications, object detection[8], fruit detection, and pedestrian *detection.*Apart from these, the fusion method is also useful in target tracking in unmanned aerial vehicles.

4. Color vision: Only visible light can be perceived by the human eye. Therefore the target image which is obtained by infrared radiation should be converted to images having pseudo colors.

5. Other applications: IR and VIS image fusion is also applied in other fields such as remote sensing[6], augmented reality and culture relic analysis re relic analysis.

III. IMAGE FUSION METHODOLOGY

In this section, details of the image fusion methodology is presented. As demonstrated in Figures 2 and 3, image fusion approaches can be broadly categorized into spatial and transform domain approaches [3].

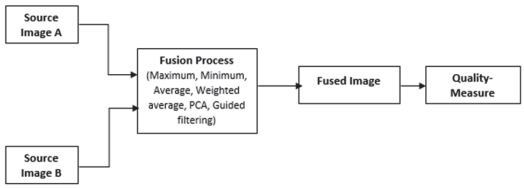


Fig. 3. Fig. 2: Spatial domain fusion

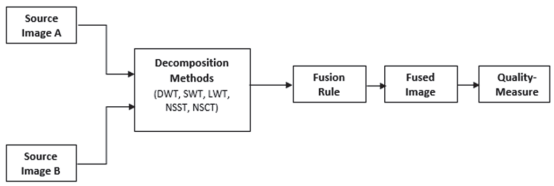


Fig. 4. Fig. 3:Transform domain fusion

A. Image Fusion Methods in Spatial Domain

1. Minimum Method

The input images are obtained and then compared in this technique. Pixel having lower value is then selected. This is a simple technique which is easy to implement and understand [9] but the fused image suffers from poor contrast and is blurred. Due to this, minimum method is not suitable for real life applications.

2. Maximum Method

This method is quite close to the IF minimum method. The difference is that the maximum pixel value is selected as the output pixel value. This method is also not suitable for IR and VIS image since poor contrast is present[9].

3. Simple average

As the name implies, we simply take the average of the pixel values of the two input images i.e., we add the pixel values of the two input images and divide by the number of input images [2]. The fusion's input images must all be from the same sensor and have a high contrast and brightness. This method is easy to comprehend and execute but noise is introduced in the fused image due to which image quality becomes low and contrast reduces.

4. Weighted average

The weighted average[2] intensity of corresponding pixels from both the input image is taken and the final fused image is derived using this method. This strategy aids in the improvement of detection accuracy.

5. Principal Component Analysis(PCA)

PCA is a technique that employs the orthogonal transformation technique [10] to create a new image with all of the properties of the previous images and two main dimensions referred to as principal components[10]. It's mostly used for feature extraction and dimensionality reduction. It reduces redundant information and has high spatial quality[11]. But the fused image has color distortion and spectral degradation.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:38:17 UTC from IEEE Xplore. Restrictions apply.

6. Guided Filtering

The guided filtering[12] method is translation variant, and has advantages of edge-preserving smoothing. Guided filters have been used in a variety of fields, including IR and VIS imaging. In computer vision applications such as noise reduction, haze removal, and smoothing, a guided filter is more effective.

B. Image Fusion Methods in Transform Domain

1. Discrete Wavelet Transform

The given source images are split into high frequency and low frequency coefficients through the use of the Discrete Wavelet Transform[13] (DWT). The lowest frequency band represent the approximation information and t The detailed information is represented by bands with high frequencies.. This method is better that the spatial domain techniques[2] because it gives the output image having high signal to noise ratio and low spectral distortion. Also it minimizes the color distortions and provide better resolution. However, DWT has the issue of shift variance and oscillations and also lacks directionality.

2. Lifting Wavelet Transform

In Lifting wavelet transform(LWT) wavelet transform the total number of samples at each stage is similar to the total number of samples at the start. It is quite alike to the DWT. To get approximation and detailed coefficients, the input samples are divided into even and odd sets and afterwards processed via the filters. Because the approximation coefficients can be obtained at a certain stage from the information coefficients previously computed and some of the input samples, the number of computations involved is also reduced. In terms of Fusion factor and Fusion symmetry, LWT is superior[14]. The Lifting scheme replaced the wavelet transform in terms of less memory and hardware requirements. However, it doesn't have shift-invariance and the output fused image has Gibbs Effect and distortion.

3. Stationary Wavelet Transform

The DWT approach suffers from translation invariance, which is overcome by the Stationary Wavelet Transform (SWT). It decomposes the source image's important feature by multi-resolution analysis. It is called a redundant technique since at each level of output same number of samples are present as the input. SWT proves to be better since it can be applied to any random sized image and the image might not necessarily be of size two's power. This method truncates zeros in filter while upsampling and suppresses while downsampling step in DWT.

4. Non subsampled Contourlet Transform

The non-subsampled contourlet transform (NSCT) [16] is a two dimentional tool for decomposition and analysis of the image derived from Contourlet transform (CT). The CT method of fusion had problems of shift variance, poor frequency selection and Gibbs phenomenon. The NSCT eliminates the upsampling and downsampling operation of CT therefore this method is said to have shift-invariance, also it ensures good frequency selectivity. The NSCT algorithm also reduces the effect of misregistration and Gibbs phenomenon. In IR and VIS image fusion, this approach is commonly utilized [17].

5. Non-Sampled Shearlet Transform

In order to overcome the demerits of Shearlet transform[16], NSST method was introduced. It has many decomposition directions due to which it possess multi-direction and translation invariance. In NSST method of fusion, non-subsampled pyramid filter bank is cascaded with shearing filter bank. It is shift invariant and can extract edge texture of the input images.

C. Fusion Rules

Following the process for fusion of source images, the fusion rule is chosen. Table I shows a few of the most commonly utilised fusion rules. It is crucial for the fusion of information from the input source image.

TABLE I. FUSION RULES

Fusion Rule	**Feature Extracted**
Maximum [9]	Texture and edge features introduced in output fused image.
Average [2]	Contrast reduction helps in distortion reduction.
PCA [10]	Reduction of dimensionality
Phase Congruency [19]	Improved localization of image features
Directive contrast [19]	Sensitive to noise, extract lines, edges and regional boundaries.

IV. EXPERIMENTAL ANALYSIS

For analysis of different fusion methods, subjective and objective analysis has been considered. For subjective analysis, Table II shows the comparison of typical spatial and transform domain techniques. For objective analysis the dataset used is TNO data set, from which two different types of IR and VIS images have been taken. The Dataset 1 in fig. 4 is the ''Bristol Queen's road[16]'' in which the person is visible in the IR image but lacks details. Whereas in VIS image the details are present but the person is not visible. The fusion of both these images has been done using 12 different methods. Another dataset 2 in fig. 5 of a market place has been taken in order to get more clarity on the Quality-metrics. The Table III and table V gives the values of Average Value (AV) and Standard Deviation[18] (SD) values of dataset 1 and dataset 2 respectively. In table IV and VI, the qualitative analysis[18] of both dataset is done. For more clarity of the variations in values, graphs have also been plotted for the AV and SD in fig. 7 and fig. 6 respectively for both the datasets

A. Performance Criterea[2,3,6]

The evaluation of the fusion method can be done with objective and subjective methods.

1. Subjective Evaluation

For this evaluation the output image of all the fusion algorithms is provided and the major areas of focus have been highlighted to get a clear idea about the difference in the quality of images. However, subjective evaluation should be done when there is a significant difference between images. Therefore, objective evaluation should also be done to get accurate judgement[16] of different algorithms.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:38:17 UTC from IEEE Xplore. Restrictions apply.

TABLE II. COMPARISON OF IMAGE FUSION TECHNIQUES

Fusion methods	Advantages	Disadvantages
Spatial Domain		
Simple Minimum [9]	Simple method, easy to implement and understand	Output is blurred, poor contrast, Not preferred for real life applications
Simple Maimum [9]		
Average method [2]		Noise is introduced, reduced contrast, low image quality
Weighted Average [2]	Detection reliability is improved	Poor contrast
PCA [10,11]	Reduces redundant information, high spatial quality	Spectral degradation and color distortion
Guided Filtering [12]	Edge preserving smoothening, translational variant	
Transform Domain		
Discrete Wavelet Transform[13]	Provides better resolution	Shift variance present, lacks directionality, problem of oscillation
Lifting Wavelet Transform[14]	Adaptive design, irregular sampling	Gibbs effect, shift variance present
Stationary Wavelet Transform[15]	Preserves edge information, and image texture	Weak in case of continuous region, time consuming, image represented only in three directions
Non-Subsampled Contourlet Transform[16,17]	Flexible, shift-invariant, faster implementation	Complicated structure, computation cost is high
Non-Sampled Shearlet Transform[16]	Multi-directionality, preserves edges and curves	Weak for subtle image features

TABLE III. QUANTITATIVE ANALYSIS OF IR AND VIS IMAGE FUSION OF DATASET 1

FUSION METHODS		QUALITY-METRICS	
		AV	SD
SPATIAL DOMAIN	Simple Minimum	69.7350	24.6568
	Simple Maximum	112.4457	29.8657
	Average	91.0903	22.6507
	Weighted Average	91.0903	22.6507
	PCA	81.3320	48.9606
	Guided Filtering	88.2701	29.7353
TRANSFORM DOMAIN	DWT	91.0900	22.7736
	LWT	91.1327	23.5047
	SWT-1	91.0903	23.0343
	SWT-2	91.0903	23.1597
	NSST	91.0916	24.3596
	NSCT	91.0888	24.4381

1. Objective Evaluation

(i) Average Value (AV)

Brightness of the output image is calculated using average value i.e the sum of all pixels is divided by total pixels in the image. If value of AV is high than mean image is brighter. Formula for AV is given below.

$$\mu = \frac{1}{M*N} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j)$$

where P(i, j) denotes pixel value of the fused image and N,M are the height and width of the image respectively.

(ii) Standard Deviation (SD)

SD is the deviation between a pixel and the average pixel value of the image. If SD is high, it means that the contrast is high, the pixel intensities are far from the mean and has a wider distribution of gray value. The formula for the SD is given below.

$$SD = \sqrt{\frac{1}{M*N} \sum_{I=1}^{M} \quad \sum_{i=1}^{N} (p(i,j) - \mu)^2}$$

where P(i, j) denotes gray value and μ represents the mean value of pixel at (i, j) and N,M are the height and width of the image respectively.



(a) (b)

Fig. 4: input (a) IR and (b) VIS image of dataset 1

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:38:17 UTC from IEEE Xplore. Restrictions apply.

TABLE IV. QUALITATIVE ANALYSIS OF FUSION METHODS OF IR AND VIS IMAGE FOR DATASET 1

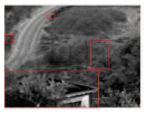
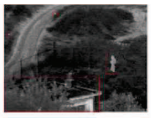
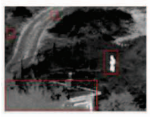
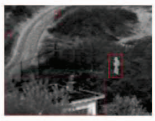
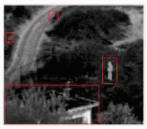
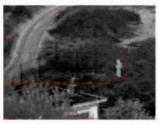
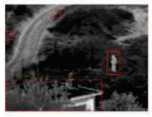
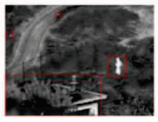
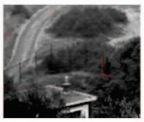
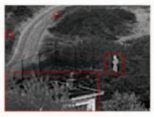
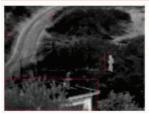
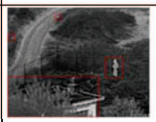
SPATIAL DOMAIN		TRANSFORM DOMAIN	
Minimum		DWT	
Maximum		LWT	
Average		SWT-1	
Weighted Average		SWT-2	
PCA		NSCT	
Guided Filtering		NSST	

TABLE V. QUANTITATIVE ANALYSIS OF IR AND VIS IMAGE FUSION OF DATASET 2

FUSION METHODS		QUALITY-METRICS	
		AV	SD
SPATIAL DOMAIN	Simple Minimum	19.8680	14.2797
	Simple Maximum	83.9372	35.7167
	Average	51.9026	20.8072
	Weighted Average	51.9026	20.8072
	PCA	92.9098	40.5680
	Guided Filtering	71.7971	35.5671
TRANSFORM DOMAIN	DWT	51.9028	20.9936
	LWT	51.9345	21.9035
	SWT-1	51.9026	21.2528
	SWT-2	51.9025	21.3777
	NSST	51.8985	22.0690
	NSCT	51.8982	23.0690

TABLE VI. QUALITATIVE ANALYSIS OF FUSION METHODS OF IR AND VIS IMAGE FOR DATASET 2

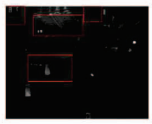
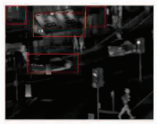
SPATIAL DOMAIN		TRANSFORM DOMAIN	
Minimum		DWT	
Maximum		LWT	
Average		SWT-1	
Weighted Average		SWT-2	
PCA		NSCT	
Guided Filtering		NSST	



Fig. 5: Input (a) IR and (b) VIS image of dataset 2

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:38:17 UTC from IEEE Xplore. Restrictions apply.

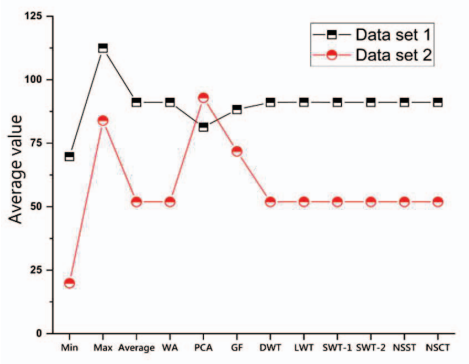


Fig. 6: Standard deviation of different fusion methods

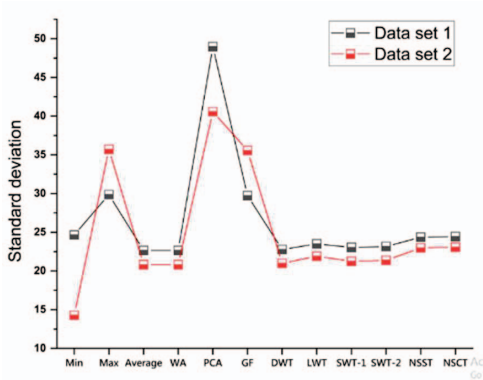


Fig.7: Variation of different fusion methods

V. CONCLUSION AND FUTURE SCOPE

In this paper, IR and VIS image fusion methodologies are described along with the various fusion rules. It also covers the applications and importance of the fusion for IR and VIS images. The resultant images are assessed using subjective and objective analysis, and several spatial and transform domain methodologies are described. The judgment of selection of the fusion method should be taken by considering combined results of subjective and objective parameters. The final conclusion of this paper is that for IR and VIS images, the transform domain techniques proves to be better in terms of combined subjective and objective parameters. This review aids in a better knowledge of contemporary image fusion methodologies and can serve as a testing ground for novel methods.

REFERENCES

[1] J. Wang, J. Peng, X. Feng, G. He, J. Fan, "Fusion method for infrared and visible images by using non-negative sparse representation," Infrared Physics & Technology 67 (2014) 477–489.

[2] Alex Pappachen James, and Belur V. Dasarathy, "Medical image fusion: A survey of the state of the art," Information Fusion 19 (2014): 4-19.

[3] Li, Shutao, et al., "Pixel-level image fusion: A survey of the state of the art," Information Fusion 33 (2017).

[4] J. Han, B. Bhanu, "Fusion of color and infrared video for moving human detection," Pattern Recognition 40 (6) (2007) 1771–1784.

[5] P.Kumar, A.Mittal, P. Kumar, "Fusion of thermal infrared and visible spectrum video for robust surveillance," in: Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing, 2006, pp. 528–539.

[6] G.Simone, A. Farina, F. C. Morabito, S. B. Serpico, L. Bruzzone, "Image fusion techniques for remote sensing applications," Information Fusion 3 (1) (2002) 3–15.

[7] M. Uzair, A. Mahmood, A. Mian, C. McDonald, "Periocular region-based person identification in the visible, infrared and hyperspectral imagery," Neurocomputing 149 (2015) 854–867.

[8] Yan, Y., Ren, J., Zhao, H. et al., "Cognitive Fusion of Thermal and Visible Imagery for Effective Detection and Tracking of Pedestrians in Videos," Cogn Compute 10, 94–104 (2018).

[9] Kaur, H., Koundal, D., & Kadyan, V. (2021), "Image Fusion Techniques: A Survey," Archives of Computational Methods in Engineering, 1 – 23.

[10] R. P. Desale and S. V. Verma, "Study and analysis of PCA, DCT & DWT based image fusion techniques," 2013 International Conference on Signal Processing , Image Processing & Pattern Recognition, 2013, pp. 66-69, doi: 10.1109/ICSIPR.2013.6497960.

[11] Arora, S., Abhinav, A. & D, A. (2019), "Image Fusion - A new hybrid technology," 2019 Innovations in Power and Advanced Computing Technologies (i-PACT). doi: 10.1109/i-pact44901.2019.8960192.

[12] D. P. Bavirisetti, R. Dhuli, "Two-scale image fusion of visible and infrared images using saliency detection," Infrared Physics & Technology 76 (2016) 52–64.

[13] Lingchao Zhan, Yi Zhuang, and Longda Huang, "Infrared and Visible Images Fusion Method Based on Discrete Wavelet Transform," Journal of Computers Vol. 28, No. 2, 2017, pp. 57-71.

[14] Y. Zou, X. Liang, T. Wang, "Visible and infrared image fusion using the lifting wavelet," Indonesian Journal of Electrical Engineering and Computer Science 11 (11) (2013) 6290–6295.

[15] Y. Deng, C. Li, Z. Zhang and D. Wang, "Image fusion method for infrared and visible light images based on SWT and regional gradient,"2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC), 2017, pp. 976-979, doi: 10.1109/ITOEC.2017.8122499.

[16] L. Junwu, B. Li and Y. Jiang, "An Infrared and Visible Image Fusion Algorithm Based on LSWT-NSST," in IEEE Access, vol. 8, pp. 179857-179880, 2020, doi: 10.1109/ACCESS.2020.3028088.

[17] J. Cai, Q. Cheng, M. Peng, Y. Song, "Fusion of infrared and visible images based on nonsubsampled contourlet transform and sparse k-svd dictionary learning," Infrared Physics & Technology 82 (2017) 85–95.

[18] S. Polinati and R. Dhuli, "A Review on Multi-Model Medical Image Fusion," 2019 International Conference on Communication and Signal Processing (ICCSP), 2019, pp. 0554-0558, doi: 10.1109/ICCSP.2019.8697906.

[19] Varalakshmi, Kolusu & Kumar, N.Raj. (2015), "Optimized Multimodal Medical Image Fusion Approach Based On Phase Congruency And Directive Contrast In Nsct Domain," International Journal Of Engineering And Computer Science. 10.18535/ijecs/v4i8.25.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:38:17 UTC from IEEE Xplore. Restrictions apply.

2021 Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 28-29, 2021

Comparison of IRPT, ANN and ANFIS based Control Algorithm for Shunt Compensation in Grid Connected Systems

Shreya Sinha
Department of Electrical Engineering
Delhi Technological University
Delhi, India
shreyasinha.53.ss@gmail.com

Ankita Arora
Department of Electrical Engineering
Delhi Technological University
Delhi, India
arora.ankita09@gmail.com

***Abstract*— These days the Power Quality has become the key issue in electrical distribution power systems. The power quality gets deteriorated due to the extensive use of non-linear loads and power electronic devices like solid-state switching devices which injects the current harmonics in the distribution systems. These harmonics leads to increase in power loss in transmission and distribution network. This paper shows a comparison of Adaptive Neuro Fuzzy Inference System (ANFIS) with Instantaneous Reactive Power Theory (IRPT) and Artificial Neural Network (ANN) for a three phase Distribution Static Compensator (DSTATCOM). This is used for compensation of power loss, current harmonics and reactive power compensation. These control methods are executed with MATLAB Simulink 2018a. The simulation results obtained demonstrates that ANFIS performs better than ANN and IRPT in terms of Total Harmonic Distortion, peak overshoot, settling time, convergence rate and the settling time of dc link voltage.**

Keywords—DSTATCOM, Fundamental current, IRPT, ANN, ANFIS, Total Harmonic Distortion.

I. INTRODUCTION

In the present scenario, power electronics plays an essential role in the distribution system. These power electronic devices like solid state switching devices and non-linear loads like refrigerators, television, conditioners etc. inject current harmonic in the distribution system and may lead to overheating of transformers, motors, compensation capacitors, cables [1] thus non-linear loads have a very deteriorating effect on the power quality. Power quality problems in the distribution system are harmonic distortion, voltage fluctuation, load unbalancing and poor power factor.

Earlier Passive filters [2-3] were used to overcome these power quality problems. But with the increase in harmonics active filters are used over passive filter as passive filters are large in size and have limitation of fixed compensation. Active filters provide adjustable power quality solution. Distribution Static Compensator (DSTATCOM) [4] is implemented for the compensation of reactive power, correction of power factor and balancing of load. DSTATCOM requires gate pulse for its operation. Control algorithm is employed to generate fundamental component of current for the required gate pulse. Shunt type active power filter [5] has been implemented to eliminate these harmonics.

PQ mitigation technique [6] is required to improve the power quality with the elimination of overshoot and oscillatory responses. Active filters [7], Passive filters, Unified power quality conditioners [8], Static var compensator, Pole zero cancellation, Synchronous d-q reference method [9], Zero voltage regulation method, Hybrid filters etc. are some of the control algorithms to mitigate oscillatory responses and overshoot. In this paper, Fundamental component of current is extracted using Instantaneous Reactive Power Theory (IRPT), Artificial Neural Network based (ANN), and Adaptive Neuro Fuzzy Inference System based (ANFIS) and a comparison of all the three controllers are shown. The simulation is done with the help of Simulink and sim power system (SPS).

II. SYSTEM CONFIGURATION

DSTATCOM is a shunt connected voltage source convertor with six IGBT (Insulated Gate Bipolar Transistor) and anti-parallel diodes connected to these IGBTs. A dc link capacitor is connected at one terminal of DSTATCOM to store energy and the other terminal is linked to the interfacing inductor. The dc link capacitor in DSTATCOM stores energy and supply reactive power as required. 3-phase non-linear load is connected and is implemented with a series resistance and inductor connected at the output terminal of the universal bridge. All the required load current and source voltage is fed to the controller and the fundamental component of current is generated as shown in Fig. 1. Hysteresis based Pulse Width Modulation method is employed to generate switching signal for DSTATCOM. Interfacing inductor is connected to minimize the ripples in the source current.

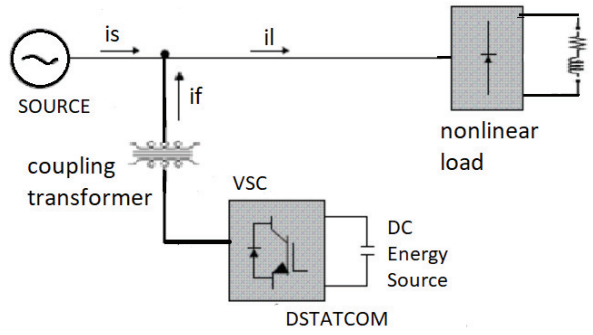


Fig. 1. Schematic Representation of DSTATCOM in Grid connected Systems

2021 Asian Conference on Innovation in Technology (ASIANCON) | 978-1-7281-8402-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ASIANCON51346.2021.9544740

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:39:08 UTC from IEEE Xplore. Restrictions apply.

III. CONTROL ALGORITHM

The performance of DSTATCOM is based on control algorithm for its gate pulse. The dc link capacitor in DSTATCOM stores energy and supplies the required reactive power.

A. Instantaneous Reactive Power Theory (IRPT)

IRPT (Instantaneous Reactive Power Theory) is applicable to all 3-phase power system, balanced, and unbalanced load and is formulated by the conversion of three phase quantities to two phase quantities. Flow diagram representation of IRPT is shown in Fig. 2. in which the three phase Voltage and current signals are Clark transformed and is converted to two orthogonal rotating vectors (α andβ). Instantaneous active and reactive power is calculated in $\alpha - \beta$ *frame* and the fundamental component of current is obtained by reverse Clark transformation of the obtained signal.

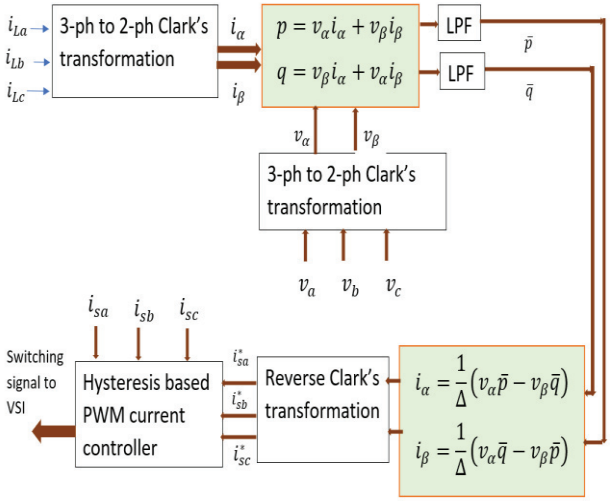


Fig. 2. Flow diagram Representation of IRPT

The three phase quantities are converted to alpha-beta frame by using Clark's transformation

$$\begin{bmatrix} V_\alpha \\ V_\beta \end{bmatrix} = \sqrt{\frac{2}{3}} \begin{bmatrix} 1 & -1/2 & -1/2 \\ 0 & \sqrt{3}/2 & -\sqrt{3}/2 \end{bmatrix} \begin{bmatrix} V_{Sa} \\ V_{Sb} \\ V_{Sc} \end{bmatrix} \quad (1)$$

$$\begin{bmatrix} i_\alpha \\ i_\beta \end{bmatrix} = \sqrt{\frac{2}{3}} \begin{bmatrix} 1 & -1/2 & -1/2 \\ 0 & \sqrt{3}/2 & -\sqrt{3}/2 \end{bmatrix} \begin{bmatrix} i_{La} \\ i_{Lb} \\ i_{Lc} \end{bmatrix} \quad (2)$$

Here α –β are the orthogonal component and a, b, c are the phase components.

Instantaneous active and reactive power is given by

$$q = -v_\beta i_\alpha + v_\alpha i_\beta \quad (3)$$

$$p = v_\alpha i_\alpha + v_\beta i_\beta \quad (4)$$

It is expressed in power and current matrix form as

$$\begin{bmatrix} p \\ q \end{bmatrix} = \begin{bmatrix} v_\alpha & v_\beta \\ -v_\beta & v_\alpha \end{bmatrix} \begin{bmatrix} i_\alpha \\ i_\beta \end{bmatrix} \quad (5)$$

$$\begin{bmatrix} i_\alpha \\ i_\beta \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} v_\alpha & -v_\beta \\ v_\beta & v_\alpha \end{bmatrix} \begin{bmatrix} p \\ q \end{bmatrix} \quad (6)$$

where,

$$\Delta = Determinant\ \ is\ expressed\ as\ {v_\alpha}^2 + {v_\beta}^2 \quad (7)$$

$$\begin{bmatrix} {i_{s\alpha}}^* \\ {i_{s\beta}}^* \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} v_\alpha & -v_\beta \\ v_\beta & v_\alpha \end{bmatrix} \begin{bmatrix} \bar{p} \\ 0 \end{bmatrix} \quad (8)$$

Where, $\bar{p}$ is the dc part

Reference source current is obtained in α-β coordinate

$$\begin{bmatrix} {i_{sa}}^* \\ {i_{sb}}^* \\ {i_{sc}}^* \end{bmatrix} = \sqrt{\frac{2}{3}} \begin{bmatrix} 1/\sqrt{2} & 1 & 0 \\ 1/\sqrt{2} & 1/2 & \sqrt{3}/2 \\ 1/\sqrt{2} & 1/2 & \sqrt{3}/2 \end{bmatrix} \begin{bmatrix} {i_0}^* \\ {i_{s\alpha}}^* \\ {i_{s\beta}}^* \end{bmatrix} \quad (9)$$

3-phase current is obtained from 2-phase current

The generated reference current is used in the hysteresis base pulse width modulation.

B. ANN

ANN is a machine learning based algorithm and it functions like the neurons of a human nervous system. ANN learns from the past data and gives output based on the predictions. ANNs exhibit a complex relation with input and output. There are three necessary layers (input layer, hidden layer and output layer) as illustrated in Fig. 3. Data sets are provided as input to the input layer. Mathematical calculations are computed in the hidden layer. Result is obtained from the output layer and it depends on parameters like weights, biases, learning rates etc. Each node in ANN has some weight and transfer function calculates the weighted sum of bias and the input. Activation function gives result based on the output of the transfer function.

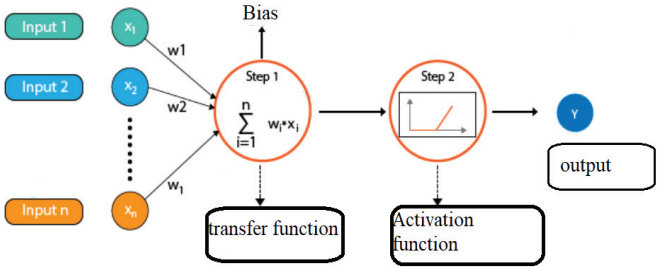


Fig. 3. Schematic Diagram of Neural Network

Levenberg Marquardt rule is applied here to train data and is based on nonlinear least squares minimization method. Three layers are present in the neural network. For the better performance of the system three input neurons with 10 hidden neurons and one output neuron is used as demonstrated in Fig. 4.

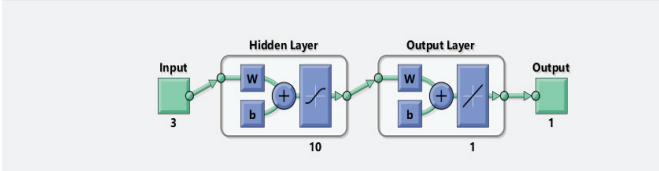


Fig. 4. Neural Network Configuration

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:39:08 UTC from IEEE Xplore. Restrictions apply.

ANN is trained using three phase load current and fundamental components of current and it gives the best result on the basis of prediction. Mean Square Error is utilized here to calculate the error and it specifies the deviation of predicted value from the correct value.

$$E_{mse} = \sum_{K=1}^{N} \frac{1}{2} \left[i_{fin} - i_{oin} \right]^2 \quad (10)$$

Here, i_{fin} is calculated fundamental component of current i_{oin} is the expected fundamental component of current and N denotes the number of samples required. IRPT controller is used to get the data. 200001 data is used and 70% data are used in training, 15% data in testing and 15% data are used for validation. MSE (Mean Square Error) obtained is $8.25712e^{-1}$. Error histogram of ANN is shown in Fig. 5.

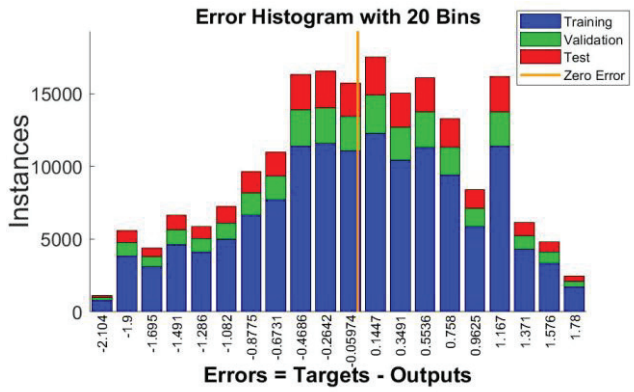


Fig. 5. Error histogram of ANN

C. ANFIS

ANFIS is a data learning technique that incorporates both neural and fuzzy logic. ANFIS is trained with a suitable training data and it uses hybrid learning, combining the least square and the gradient descent method as training learning rules. The Flow chart of ANFIS is illustrated in Fig. 6.

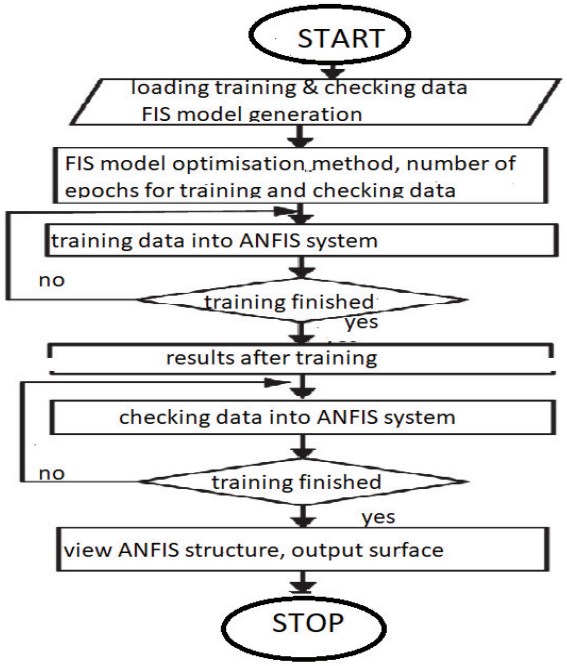


Fig. 6. Flow chart of ANFIS

ANFIS embodies five layers and are input node, rule nodes, average nodes, following nodes and output node layers. Layer1 is the adaptive node (input node) layer, three phase load current is fed as input to the adaptive node layer and it gives fuzzy membership grade of the input as output to the second layer. Layer2 is the rule node layer that generates the firing strength for rules. The AND operator is applied and its output is employed in the average node layer. Layer3 is the average node layer that normalizes the computed firing strength. Layer4 is the consequent node layer that takes normalized values as input and returns defuzzified values as output. Layer 5 gives the desired output direct current.

For training the ANFIS structure 10 epochs are used as shown in fig. 7. Grid partition is used to generate FIS. Triangular membership function is assigned at each input. Fig. 8. shows Optimized membership function for input1, Fig. 9. shows Optimized membership function for input2 and Fig. 10. shows Optimized membership function for input3. For training 200001 data are used. Fig. 11. shows the structure of ANFIS with 78 nodes, 27 linear parameter and 27 nonlinear parameters. For training 27 fuzzy rules are used and training MSE obtained as 0.607807.

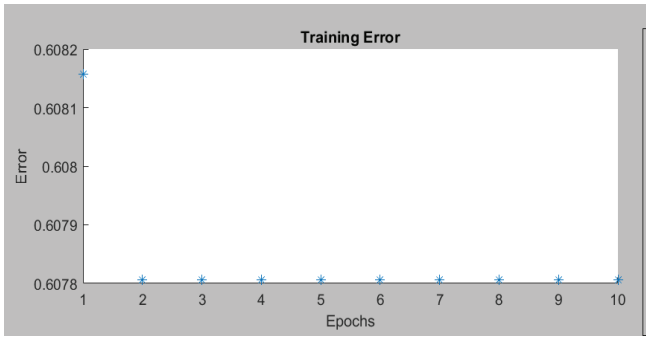


Fig. 7. Training error for 10 epoch

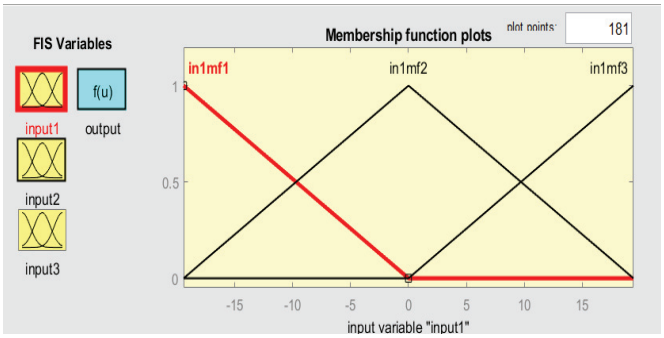


Fig. 8. Optimized membership function for input1

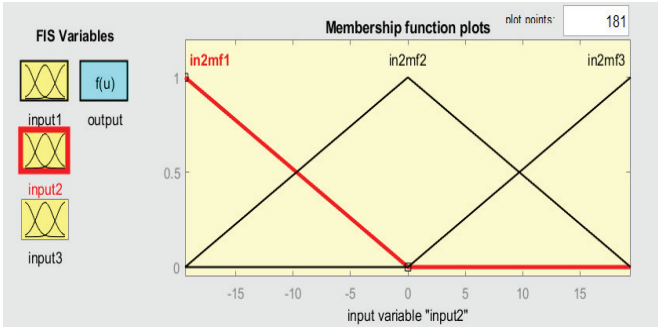


Fig. 9. Optimized membership function for input2

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:39:08 UTC from IEEE Xplore. Restrictions apply.

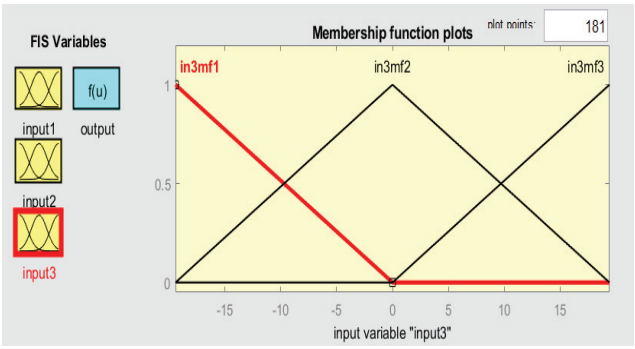


Fig. 10. Optimized membership function for input3

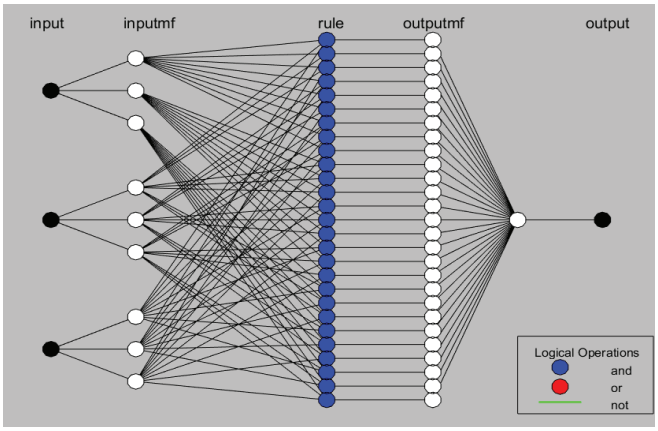


Fig. 11. Structure of ANFIS

IV. RESULTS AND DISCUSSION

Non-linear load is simulated and for this Resistance ($R = 15\Omega$) and inductance ($L = 3 * 10^{-3}H$) is used with a universal bridge. A circuit breaker is implemented in the circuit to increase the load. The Load current of 10A is obtained from t=0 sec to t=0.2 sec. The non-linear load is increased at t=0.2 sec and it is analyzed that from t=0.2 sec grid current also amplifies to fulfil the load demand.

The THD of load current is obtained as 21.01%. ANFIS control is implemented to reduce the THD to 1.06% thus fulfilling IEEE519 standards. It is analyzed that dc link capacitor voltage settles at t=0.03 sec and when load increases at t=0.2 sec it shows few disturbances but settles fast.

TABLE I. DESIGN PARAMETERS OF THE GRID INTEGRATED DSTATCOM MODEL

Source Inductance	0.002mH
Load Inductance	0.003H
Load Resistance	15Ω
Interfacing Inductance	0.002H
Source RMS Voltage	110V

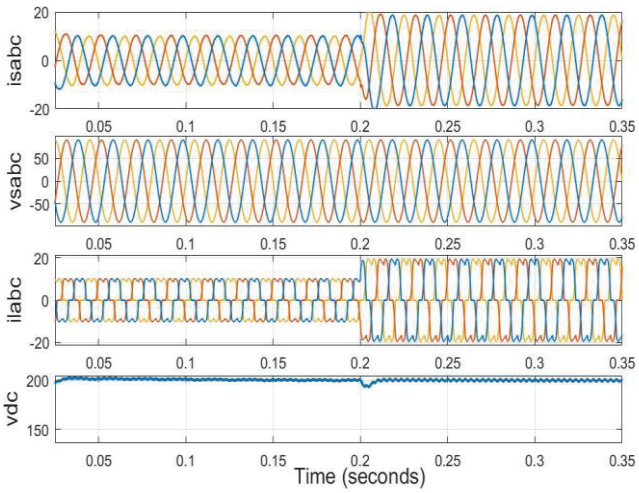


Fig. 12. Waveforms obtained using ANFIS

Fig. 12. illustrates the result of source current, source voltage, load current and dc link capacitor voltage. Fig. 13. illustrates THD of load current is 21.01%. Fig. 14. shows THD of source voltage is 0.10%, Fig. 15. illustrates THD of source current is 1.06%.

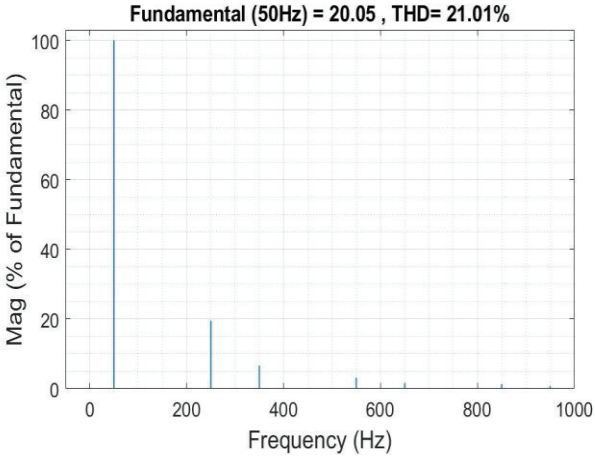


Fig. 13. THD of load current using ANFIS

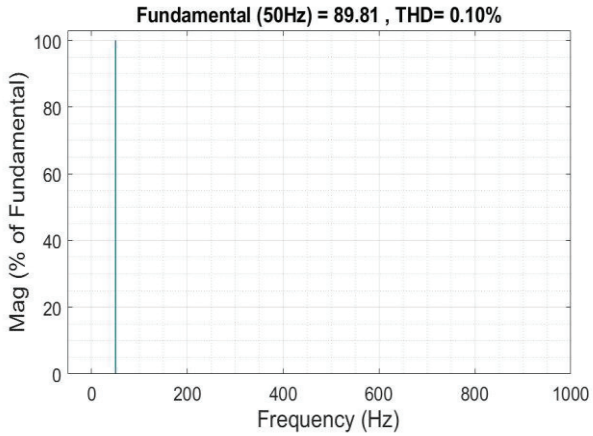


Fig. 14. THD of source voltage using ANFIS

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:39:08 UTC from IEEE Xplore. Restrictions apply.

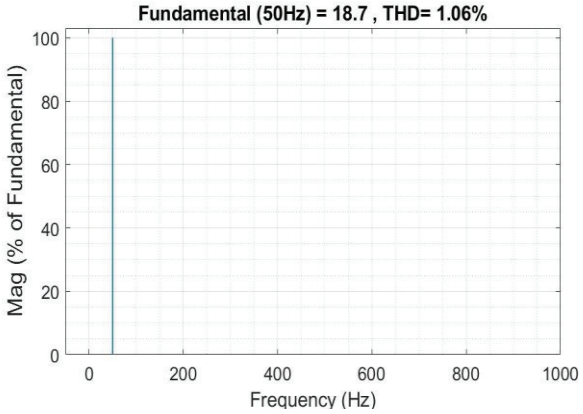


Fig. 15. THD of source current using ANFIS

V. COMPARISON ANALYSIS

The fundamental component of current of IRPT, ANN and ANFIS is extracted and is compared. Table 3 shows the obtained rise time, peak overshoot, THD and convergence rate. The rise time is 0.00076 sec in case of IRPT, 0.00332 sec in case of ANN and 0.00201 sec in case of ANFIS. Peak overshoot obtained is 3.984% in case of IRPT, 2.312% in case of ANN and 1.06% in case of ANFIS. THD of source current is 2.03% in case of IRPT, 1.78% in case of ANN and 1.06% in case of ANFIS. The convergence rate is high in IRPT, low in ANN and ANFIS. Fig. 16. shows comparison of fundamental component of current extracted using IRPT, ANN and ANFIS.

TABLE II. PERFORMANCE ANALYSIS COMPARISON OF ANFIS WITH IRPT AND ANN IN TRACKING FUNDAMENTAL COMPONENT OF LOAD CURRENT:

DSTATCOM Control Algorithm	Rise Time (second)	Peak Overshoot	%THD	Convergence Rate
IRPT	0.00076	3.984%	2.03%	High
ANN	0.00332	2.312%	1.78%	Low
ANFIS	0.00201	1.531%	1.06%	Low

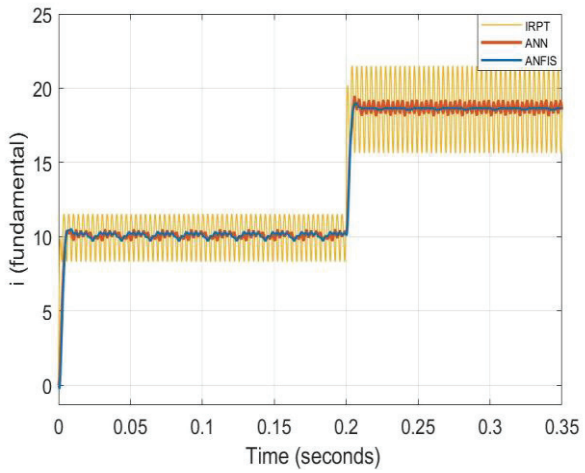


Fig. 16. Comparison results of performance analysis of fundamental component of current extracted using IRPT,ANN and ANFIS

The dc link capacitor voltage is obtained using IRPT, ANN and ANFIS and is compared. Table 4 shows the obtained dc link capacitor voltage in case of IRPT, ANN, ANFIS. The dc link capacitor voltage settles at 0.13 sec in case of IRPT, 0.07 sec in case of ANN and 0.04 sec in case of ANFIS. Fig. 17. shows the comparison result of dc link capacitor voltage of IRPT, ANN and ANFIS.

TABLE III. PERFORMANCE ANALYSIS COMPARISON OF ANFIS, ANN AND IRPT IN TRACKING DC LINK VOLTAGE:

DSTATCOM Control Algorithm	Settling Time (seconds)
IRPT	0.13
ANN	0.07
ANFIS	0.04

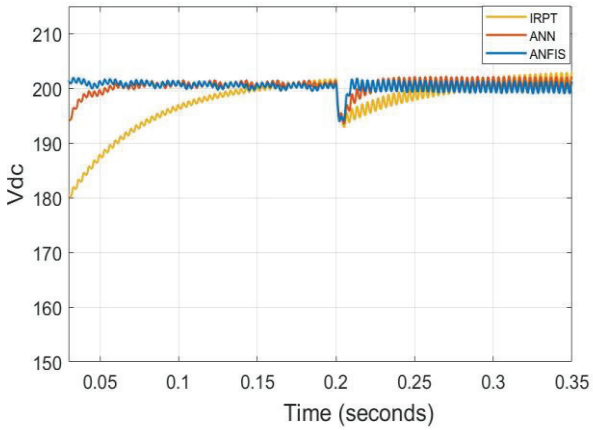


Fig. 17. Comparison results of performance analysis of DC Link voltage using IRPT, ANN and ANFIS

VI. CONCLUSION

This paper demonstrates the comparative analysis of IRPT (Instantaneous Reactive Power Theory), ANN (Artificial Neural Network) and Adaptive Neuro Fuzzy Inference System based (ANFIS) method and the demonstration is carried under dynamic load condition. The fundamental component of current of IRPT, ANN and ANFIS are extracted and compared. A comparative analysis of dc link capacitor voltage is also done. THD in case of IRPT, ANN and ANFIS is analyzed and the result demonstrates that THD in ANFIS is 1.06% which is better when compared to IRPT and ANN.

REFERENCES

[1] Tey LH, So PL, Chu YC. Improvement of power quality using adaptive shunt active filter. IEEE Trans Power Deliv 2005;20(April (2)):1558–68.

[2] J.e. Das, "Passive filters - potentialities and limitations," Industry Applications, IEEE Transactions on, vol. 40, pp. 232-241, 2004.

[3] N.G. Hingorani, "Introducing custom power," Spectrum, IEEE, vol. 32, pp. 41-48,1995.

[4] K. R. Padiyar, "FACTS Controllers in Power Transmission and Distribution," New Age International (P) Limited, Publishers, New Delhi, 2007.

[5] S. G. Seifossadat, R. Kianinezhad, A. Ghasemi and M. Monadi, "Quality improvement of shunt active power filter, using optimized tuned harmonic passive filters," 2008 International Symposium on

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:39:08 UTC from IEEE Xplore. Restrictions apply.

Power Electronics, Electrical Drives, Automation and Motion, Ischia, Italy, 2008, pp. 1388-1393, doi: 10.1109/SPEEDHAM.2008.4581063.

[6] A. Moreno-Munoz, Power Quality: Mitigation Technologies in a Distributed Environment. London, U.K.: Springer-Verlag, 2007.

[7] B. Singh, K. Al-Haddad, and A. Chandra. "A review of active filters for power quality improvement". Industrial Electronics, IEEE Transactions on, 46(5):960–971, 1999.

[8] Saha Shib Shankar, Suryavanshi Ramesh. Power system harmonic mitigation of an offshore oil rig using passive shunt filter. Presented in IEEE India Conference (INDICON), Kolkata, India; December 2010.

[9] B. Singh and J. Solanki, "A comparative study of control algorithms for DSTATCOM for load compensation," in Proc. IEEE ICIT, Dec. 15–17, 2006, pp. 1492–1497.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:39:08 UTC from IEEE Xplore. Restrictions apply.

Optik - International Journal for Light and Electron Optics 248 (2021) 168121

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.com/locate/ijleo

Defects assisted luminescence in *m*-HfO_2 nanocrystals: An experimental and theoretical study

Rajesh Kumar [a], Ankush Vij [b], Mukhtiyar Singh [a]

[a] Department of Applied Physics, Delhi Technological University (DTU), Delhi 110042, India
[b] Department of Physics, University of Petroleum and Energy Studies (UPES), Uttarakhand 248001, India

ARTICLE INFO

Keywords:
m-HfO_2
Photoluminescence
Thermoluminescence
First principles calculations

ABSTRACT

We report here the photoluminescence (PL) and thermoluminescence (TL) of single phase monoclinic (*m*)-HfO_2 nanocrystals. The *m*-HfO_2 shows broad PL in warm white region upon 325 nm excitation, which may be assigned to 4-fold (O_4) vacancies as predicted by first principles calculations. TL glow curve of gamma irradiated *m*-HfO_2 comprises of a main peak around 400 K along with a low intensity peak at 480 K, and detailed kinetic analysis confirms the presence of two kinds of metastable traps (0.86 and 1.60 eV). The carrier concentration in shallow traps (0.86 eV) were found relatively much higher than deep traps (1.60 eV), which augurs well for stronger TL signal at 400 K.

1. Introduction

From the prospective of energy saving, transition metal (TM) oxides phosphor based light emitting diodes (LEDs) are considered as useful devices for solid state lighting. Amongst TM oxides, hafnium oxide (HfO_2) has been extensively studied due to its versatile applications because of its high dielectric constant (K≈25), wide band gap (5.2–5.9 eV) and polymorphic phases [1–3]. The HfO_2 polymorphs show phase dependent optical and mechanical behavior such as high transmittance in wide spectroscopic range, mechanical hardness etc. making it suitable for optical coatings, bandpass filters, UV mirrors, efficient absorber for gamma, infrared optical devices etc. [4–8]. It has been shown that under different temperatures and pH conditions, doped HfO_2 nanocrystals exhibit change in size and phase dependent luminescent properties [9–11]. Metal oxides like HfO_2 and ZrO_2 have also been found to show intrinsic luminescence without any rare earth activators [12,13] which has been ascribed to the presence of oxygen vacancies (defects) in crystal lattice [14–16]. This intrinsic luminescence can further be tuned under different temperature and oxygen environment conditions. Along with the vacancies, there may also exist long lived radiative trapping states in materials which can be probed experimentally using thermoluminescence (TL) in pure and doped HfO_2 [17–20]. In addition to the defect characterization, TL is also used extensively in radiation dosimetry. However, to best of our knowledge only a few groups have synthesized nanocrystalline HfO_2 to particularly pay attention to thermoluminescence properties and analysis of actual trap distribution in the glow curve after gamma rays irradiation. This motivated us to carry out the detailed analysis of TL in undoped HfO_2 nanocrystals to probe the existence of trapping defects. Radiative defects study can further be understood by photoluminescence (PL) spectroscopy [21]. We performed density functional theory (DFT) based calculations to understand the defects induced modifications in electronic structure of HfO_2. Experimental and theoretical study of such defects is paramount for understanding intrinsic luminescence in HfO_2.

Here we present the experimental and theoretical study of chemically synthesized undoped HfO_2 nanophosphors to understand the

E-mail addresses: vij_anx@yahoo.com (A. Vij), mukhtiyarsingh@dtu.ac.in, msphysik09@gmail.com (M. Singh).

https://doi.org/10.1016/j.ijleo.2021.168121
Received 20 August 2021; Received in revised form 23 September 2021; Accepted 7 October 2021
Available online 10 October 2021
0030-4026/© 2021 Elsevier GmbH. All rights reserved.

defects induced PL and TL. The density functional theory was used to elucidate the band structure and role of defects in PL of HfO_2. Further trapping states were investigated experimentally using TL of gamma irradiated HfO_2 nanocrystals.

2. Experimental and theoretical details

2.1. Synthesis method

In this work we have synthesized undoped HfO_2 nanocrystals through low temperature chemical co-precipitation method [22]. To synthesis NP's of HfO_2, initially stoichiometric $HfCl_4$ was dissolved in the DI water upon magnetic stirring. NH_4OH was poured dropwise and pH value of solution was maintained at 7.2 during the synthesis. The temperature was maintained at 120 °C during the stirring to avoid intermediate Hf (OH) $_4$. After stirring for an hour on hot plate the resulting precipitates were washed with distilled water with the help of centrifuge. The excess salt was discarded and yellow colored precipitate was placed at the heating mantle overnight at 50 °C. The dried powder was crushed and placed in the muffle furnace for further annealing at 1100 °C.

Powder X-ray diffraction pattern was obtained using Rigaku Ultima-IV X-Ray diffractometer (Cu-Kα, λ = 1.5417Å) operated at 8.04 KeV. Photoluminescence and Raman spectra were performed on inVia confocal Raman microscope using 325 nm line of He-Cd laser and 514 nm line of Ar laser respectively in this work. UV-Vis absorption spectroscopy was performed on Indian make RIMS (RIII India) set-up. The samples (5 mg each) were irradiated with Co-60 gamma rays source (Dose rate = 3.29 kGy/h) for thermoluminescence response and data was recorded using 'Harshaw 3500' TLD Reader at a heating rate of 10 K/s.

2.2. Theoretical methodology

The electronic structure calculations of *m*-HfO_2 were performed using density functional theory (DFT) based Full Potential Linearized Augmented Plane-Wave (FPLAPW) method as implemented in WIEN2k program package [23]. Generalized gradient approximation (GGA) of Perdew-Burke-Ernzerhof [24] and Tran-Blaha modified Becke-Johnson potential (TB-mBJ) were used for exchange-correlation potentials [25]. The GGA approximation underestimate the band gap whereas TB-mBJ potential use semilocal quantities to reproduce well step structures. The muffin tin radii were chosen to be non-overlapping to get better convergence of energy eigen values. Plane wave cutoff was decided by $R_{mt}K_{max} = 7$ and convergence criteria for energy and charge was set to 0.0001 Ry/cell. A 17 × 17 × 16 k-points mesh was used which results in 518 k-points in the irreducible Brillouin Zone (IBZ). The oxygen vacancies were simulated using a 2 × 2 x 2 supercell of 92 atoms with 10 × 9 x 9 k-points mesh.

3. Results and discussion

Fig. 1 shows the XRD pattern of prepared HfO_2, which resembles well with standard reference JCPDS card No. 78-0049. Rietveld refinement of crystal structure was done using the FullProf suite using pseudo-Voigt shape function [26], which confirms the monoclinic phase of annealed HfO_2 with space group $P2_1/c$ and unit cell parameters as a = 5.06 Å, b = 5.11 Å, c = 5.23 Å, β = 99.19°. Reliability factor which confirms the quality of fit such as weighted profile R factor (Rwp), profile R factor (Rp), expected R factor (Rexp) are 14.0, 15.5, 13.5 respectively, convergence indicator like chi square 1.07 with goodness of fit 2.9 was observed. The average crystallite size was calculated to be 32.85 nm using Debye Scherer relation which confirms the nanometer size particles. The position and normalized occupation index of *m*-HfO_2 are mentioned in Table 1, indicating the presence of Hf and O vacancy.

Vibrational modes of *m*-HfO_2 nanoparticles were investigated using Raman spectrum as shown in Fig. 2. We observed Raman modes at wavenumbers 108, 136, 149, 164, 244, 258, 324, 335, 381, 398, 498, 522, 550, 582, 642, 673, and 779 cm^{-1} with most

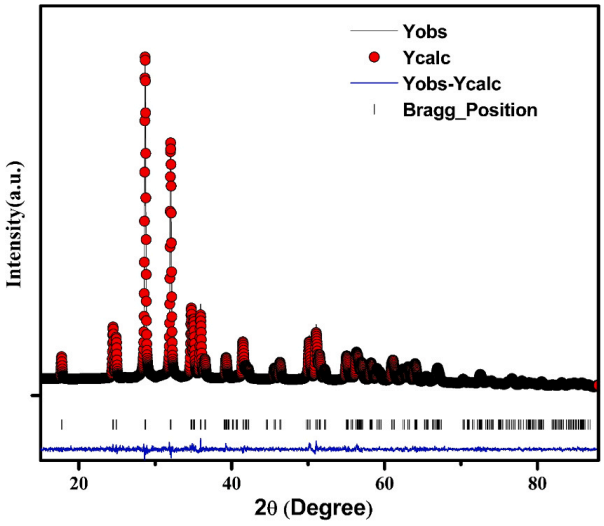


Fig. 1. XRD pattern and Rietveld refinement of *m*-HfO_2.

Table 1
Atomic positions and occupancy of atoms using rietveld refinement.

Atoms	X	Y	Z	Normalized Occupancy
O2	0.45123	0.75348	0.47909	1.00000
Hf	0.27525	0.03990	0.20769	0.92381
O1	0.07613	0.33631	0.34027	0.97980

intense peaks at 149 cm^{-1} and 498 cm^{-1}. Except the small variation in peak width all the modes have good resemblance with previous reports of *m*-HfO_2 phase [27–29]. The slight variation occurring in the peak intensity and peak position towards high wave number may be ascribed to the formation of nanostructures leading to phonon confinement.

The absorption spectrum of *m*-HfO_2 nanoparticles exhibits the fundamental absorption edge in the short UV region as shown in Fig. 3A. The indirect band gap in *m*-HfO_2 was calculated to be 5.71 eV using Tauc's plot as shown in Fig. 3B, which is in close agreement with previous studies [2–4]. In *m*-HfO_2 a small shoulder like structure also appears near the band edge, which may be ascribed to oxygen vacancy or shifting of Hf—O bond length. Without affecting the lattice structure, the small variation of Hf—O bond length may create a perturbed potential and influence the optical and electronic band properties, resulting in shoulder formation [30, 31].

To understand the band structure, we performed DFT calculations using the experimentally determined unit cell parameters. The band structures of the pristine *m*-HfO_2 calculated within GGA and TB-mBJ formalism are shown in Fig. 4. It is evident that the both the approximations predict maxima of valance band lying at the high symmetry Gamma point (Γ) of the Brillouin zone and conduction band minima lying at the high symmetry X point of Brillouin zone which results in an indirect band gap along Γ-X direction. The value of band gaps is listed in Table 2. As mentioned earlier, the value of band gap (4.1 eV) predicted by GGA approximation is quite smaller in comparison to the experimental determined value (5.71 eV), whilst the TB-mBJ predicted band gap (5.78 eV) is in excellent agreement with the experimental value, however with slight overestimation. The similar slight overestimation of band gap using TB-mBJ potential is also reported in literature [32].

To gain the more insight of electronic structure of *m*-HfO_2, we explored the partial density of state of *m*-HfO_2, using GGA and TB-mBJ potentials, as shown in Fig. 5. The qualitative features of partial density of states calculated using both the approaches are quite similar. Therefore, here we are presenting the details analysis using TB-mBJ only. The low-lying states are mainly composed of O-*p* states with a very feeble contribution from Hf-*d* states. Except a small peak at around − 10 eV, a high-density valence band gages from ∼ − 5.0 eV to 0 eV (which corresponds to Fermi energy here). The conduction band, on the other hand, is dominated by Hf-*d* states with a weak presence of O-*p* states which shows a hybridization between Hf-*d* states and O-*p* states.

Fig. 6 shows the photoluminescence emission spectra of *m*-HfO_2 taken at room temperature using 325 nm line of He-Cd laser source. In this phase the dominant emission appears at 450 nm (2.75 eV), 536 nm (2.31 eV) and 562 nm (2.20 eV) corresponding to blue, green and yellow regions respectively. Due to wide band gap, electrons in the valance band may not be directly excited to the conduction band at 325 nm. The electrons can only be excited via charge transfer band of the crystal which is located below conduction band. The relaxation of excited electrons to valance band through different defect levels show emission in blue, green and yellow regions, which is in good agreement with some previous study [33,34]. The colorimetric properties using Commission Internationale De I′Eclairage (CIE) (inset of Fig. 6) coordinates (0.35, 0.46) infer the warm white light emission in *m*-HfO_2.

Researchers have earlier experimentally shown the presence of oxygen vacancy related radiative defect states in HfO_2 thin film [35]. The electrons are captured in these oxygen vacancies and lead to the formation of ionized oxygen vacancy centers. Single and doubly charged vacancy leads to trap extra electron which is polaronic in nature. On the vacancy site, the trapped electron doesn't fit

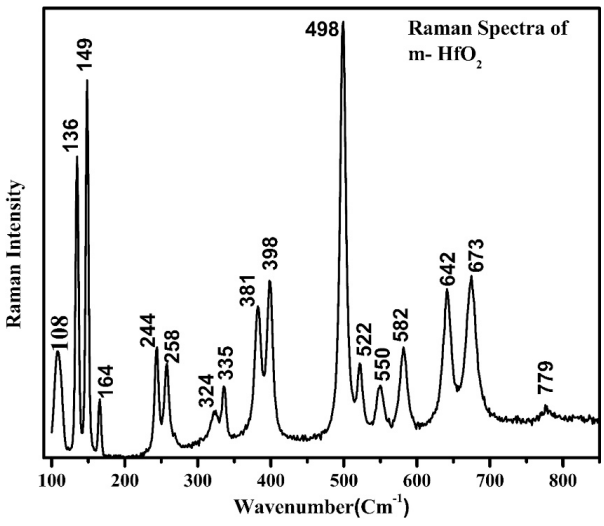


Fig. 2. Raman spectrum of *m*-HfO_2.

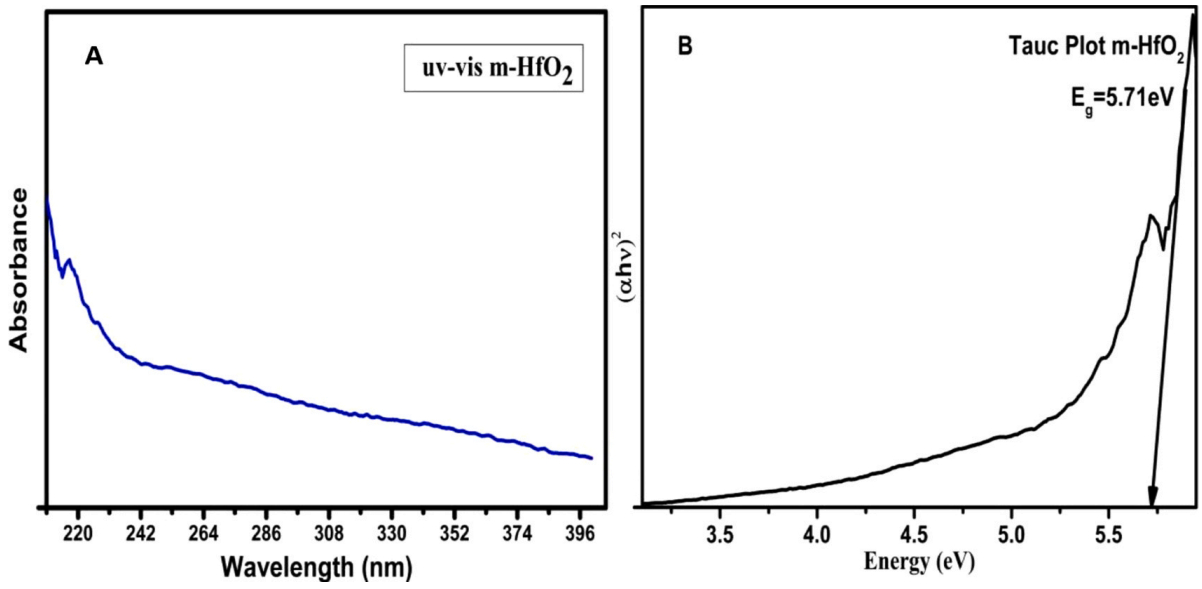


Fig. 3. (A) Absorbance spectrum and (B) Tauc plot for *m*-HfO_2.

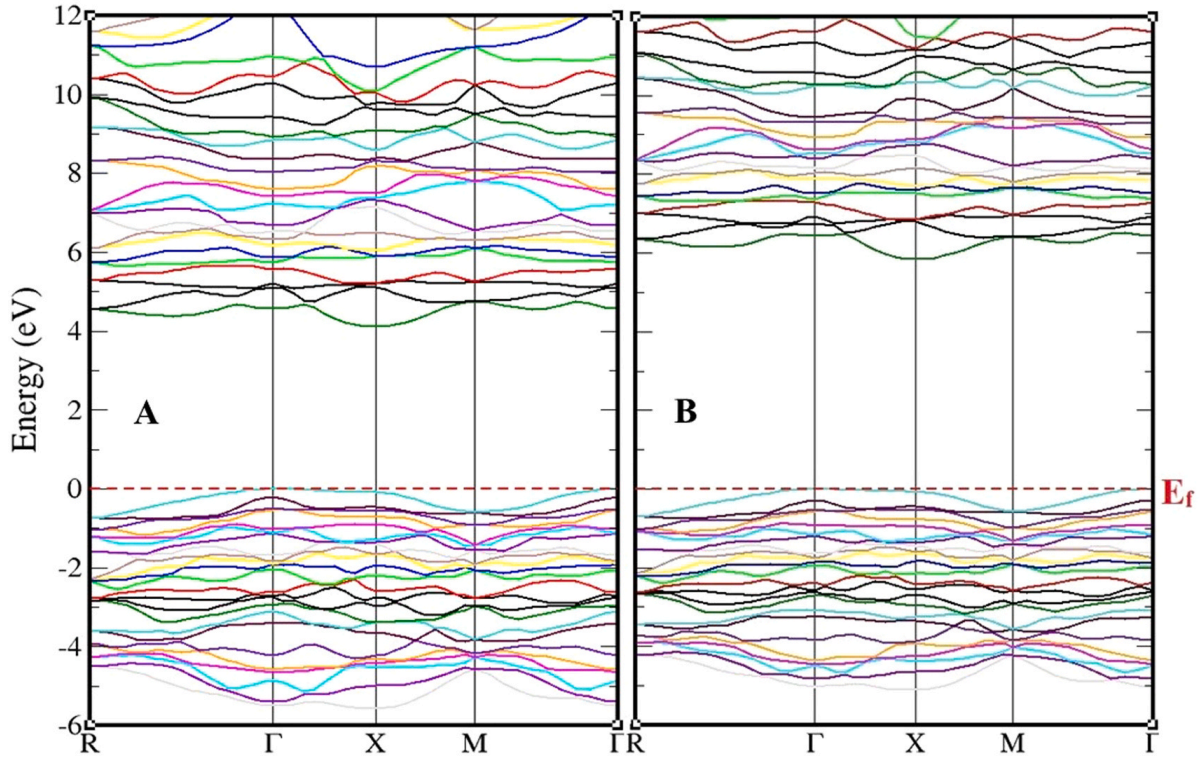


Fig. 4. Band structure of *m*-HfO_2 using (A) GGA and (B) TB-mBJ approach.

Table 2
Comparison of experimental and theoretical values of band gap.

Experimental band gap	5.71 eV (Present study)	
Theoretical band gap (Present study)	GGA	TB-mBJ
	4.11 eV	5.78 eV
[#]Other's result	4.08 eV	5.76 eV

#- Ref. [32].

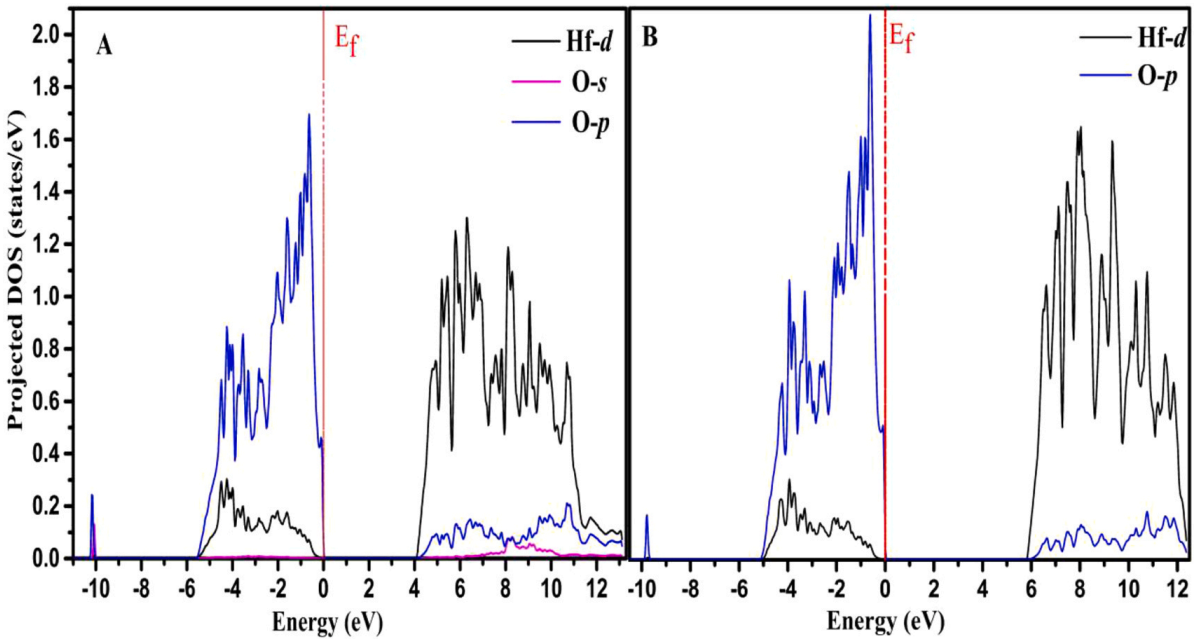


Fig. 5. (A) Projected density of states of *m*-HfO_2 using GGA approach and (B) TB-mBJ approach. (Fermi energy corresponds to 0 eV).

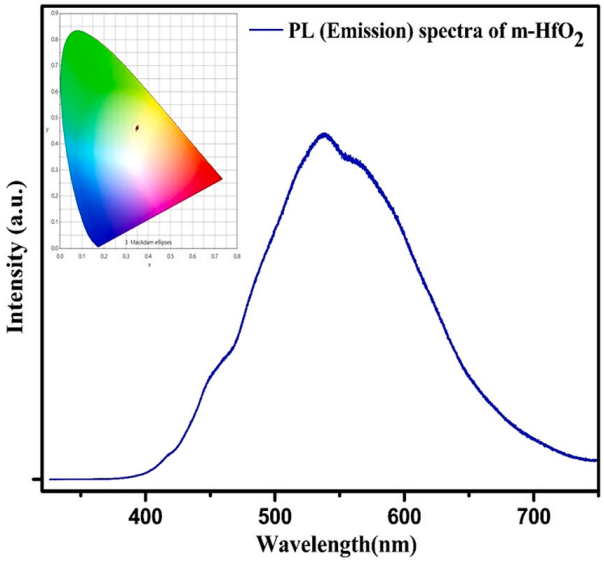


Fig. 6. PL emission spectrum and CIE plot (inset) of *m*-HfO_2.

in vacancy space rather they distributed over nearest Hf atom [36], leading to the broad PL spectrum. It is seen that two types of oxygen vacancies exist in *m*-HfO_2 named as 3-fold (O_3) and 4-fold (O_4) because each oxygen atom is surrounded by 3 Hf atoms and 4 Hf atom, respectively [36,37]. The removal of oxygen atoms leads to the structural relaxations of the nearest-neighbor Hf ions surrounding the vacancy which leads to the lower energy configuration corresponding to defect states which is verified by our calculations. As the doubly O-vacancy lowers the band gap up to 0.60 eV and 0.39 eV in case of O_3 and O_4, respectively, the observed PL spectra in *m*-HfO_2 cannot be explained using doubly O-vacancy as discussed in following text. So, here we only concentrate on single O-vacancies. It has been found that the defect states corresponding to O_4 vacancy is more stable than the O_3 vacancy. The total density of states corresponding to these vacancies is depicted in Fig. 7. The conduction band minima, in case of O_3 single vacancy, is red-shifted at 2.27 eV and at 2.18 eV in case of O_4 single vacancy which results in decrease of energy band gap. As the TB-mBJ approach predicts the accurate band gap, near to experimental value, this probably increases the accuracy with which it envisages the defect states. According to our calculations the PL peaks (shown in Fig. 6) corresponding to 562 nm, 536 nm and 450 nm wavelength are attributed to O_4 single vacancy. These peaks are correlated with total density of states as shown in Fig. 7B. First peak arises at 563 nm, second at 534 nm, and third at 476 nm which are in close agreement with our experimental observation.

Fig. 8 shows the TL glow curve of *m*-HfO_2 after exposure with gamma rays. A strong TL peak at 400 K was observed for monoclinic phase. The TL behavior of *m*-HfO_2 were investigated as function of gamma dose which shows increase in TL intensity with increasing

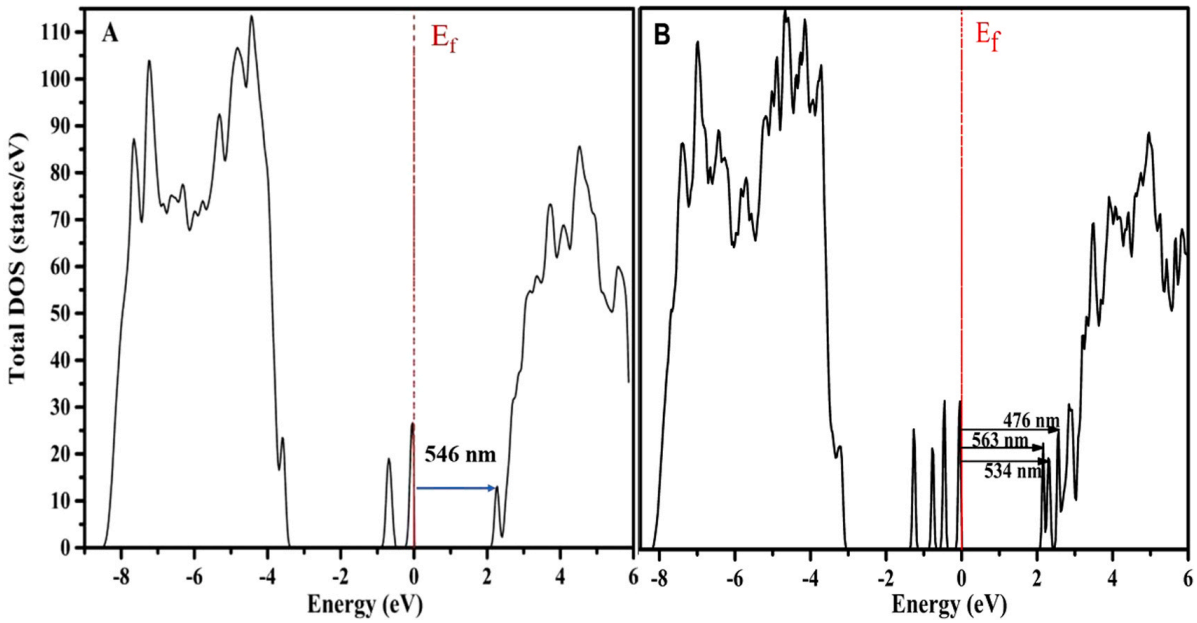


Fig. 7. Total density of states of *m*-HfO_2 with three and four-fold single oxygen vacancy using TB-mBJ. (Fermi energy corresponds to 0 eV).

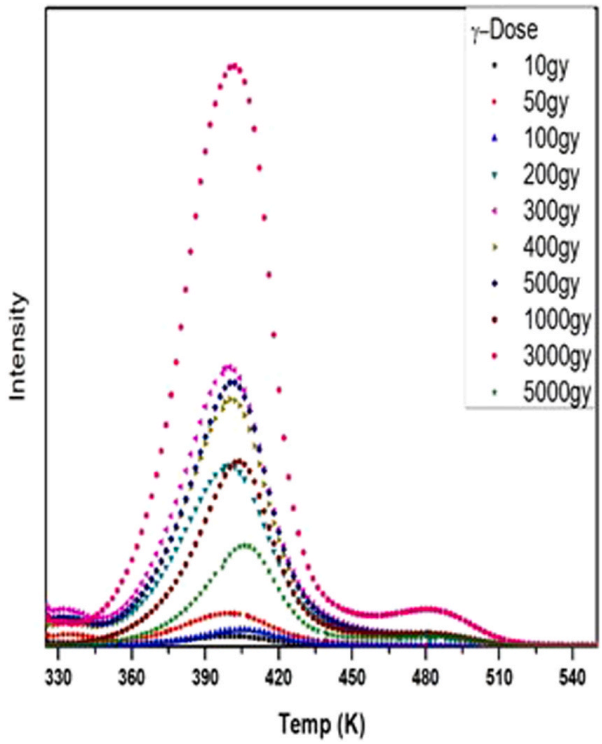


Fig. 8. TL glow curve of *m*-HfO_2 irradiated with different doses of gamma rays.

the exposure dose. Highest TL maxima was observed at 3000 Gy while at 5000 Gy the TL intensity decreases. TL intensity shows almost linear dependency with the dose, and become supralinear at 1000 Gy. The appearance of both the peaks at 400 K and 480 K respectively can predominantly be correlated to the O-vacancies and partly to of intrinsic defect states of *m*-HfO_2. These defect traps the electrons during gamma irradiation which gets de-trapped upon heating giving rise to TL glow curve. Two peaks in the glow curve signifies the presence of at least two kinds of traps with different activation energies, order of kinetics and frequency factor. Therefore, in order to probe the trapping states and their distribution in gamma irradiated *m*-HfO_2, we evaluated the kinetic parameters such as activation energy, order of kinetics and frequency factor by deconvoluting TL glow curves using TLanal software based on Kiti's general order equation [38–40]. Fig. 9 shows the deconvoluted TL curves of gamma (500 Gy) irradiated HfO_2 nanophosphors. The best fit corresponds to 2 trap states at 400 K, 480 K and lies in the activation energies 0.86 eV and 1.60 eV, respectively. The carriers

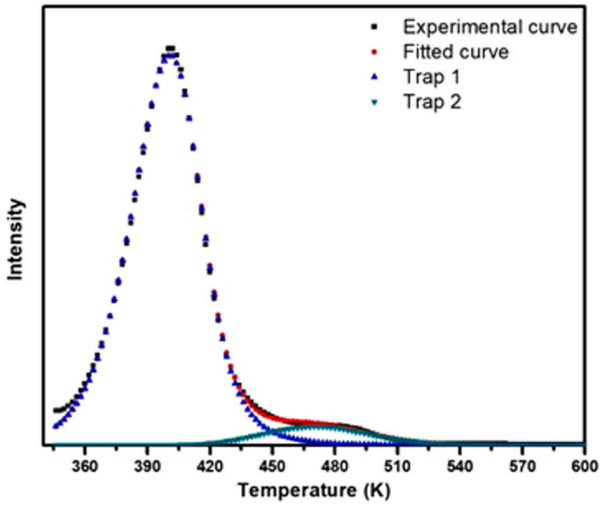


Fig. 9. Deconvoluted TL glow curve of 500 Gy gamma rays irradiated *m*-HfO_2 using TLAnaL software.

trapped at 0.86 eV activation energy appeared to be little shallow and decays through first order kinetic. While the electron trap state at 480 K having activation energy 1.60 eV lies comparatively deeper thus probability of re-trapping of electrons in trap state lies above this energy level become significant enough for mixed order (b = 1.85) decay kinetics. Also, the number of electron traps ($n_\circ$) in shallow trap states (E = 0.86 eV) is relatively high than those of deeper trap state hence thermal agitation equivalent to 0.86 eV give an intense TL peak. From the TL glow curve deconvolution results, it can be inferred for HfO_2 that there might be two kinds of trap states present, one being close to conduction band require less activation energy to thermally de-trapped and the other exist relatively deeper. For other gamma doses, TL glow curves shape remain same except the variation in intensity, which shows that increasing gamma dose does not induce new trapping states however affects the carrier density in both the traps. The present study establishes that *m*-HfO_2 have intrinsic trapping states giving rise highly gamma responsive TL behavior which can be optimized for TL based gamma dosimetry.

4. Conclusions

We synthesized monoclinic HfO_2 nanocrystals using chemical co-precipitation method. The XRD and rietveld analysis shows the presence of single monoclinic phase, which is also confirmed using Raman spectrum. The *m*-HfO_2 nanocrystals shows experimental indirect band gap of 5.71 eV calculated using Tauc's method, which is in close agreement with theoretical results. The projected density of states shows the predominant presence of O-*p* states and Hf-*d* states in valence and conduction band, respectively. The PL spectrum of *m*-HfO_2 shows a broad emission in warm white region which has been ascribed to 4-fold (O_4) single vacancies as predicted by our first-principles calculation. Thermoluminescence of gamma irradiated *m*-HfO_2 nanocrystals shows the presence of shallow (0.86 eV) and deep trap (1.60 eV) states estimated using TLanal software based on Kiti's general order equations.

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.

Acknowledgement

The authors wish to express their gratitude to the director of IUAC New Delhi for allowing and Mr Birender to help in performing TL experiments in the Health Physics Lab. One of the authors (Mukhtiyar Singh) thanks the University Grants Commission (UGC) of India for funding the computational facilities at the Department of Applied Physics, Delhi Technological University, Delhi, India, under the START-UP GRANT (No. F.30-416/2018 (BSR), dated 18/09/2018).

References

[1] G.D. Wilk, R.M. Wallace, J.M. Anthony, High-k gate dielectrics: current status and materials properties considerations, J. Appl. Phys. 89 (2001) 5243.
[2] H.P. Kumar, S. Vidhya, S.S. Kumar, C.V. Kumar, Sam Solomon, J.K. Thomas, Optical properties of nanocrystalline HfO2 synthesised by an auto igniting synthesis, J. Asian Ceram. Soc. 3 (2015) 64–69.
[3] J. Manikantan, H.B. Ramalingam, Chandra Shekar, B. Murugan, R. Ranjith Kumar, J. Sai Santhoshi, Physical and optical properties of HfO2 NPs- synthesis and characterization in finding its feasibility in finding its feasibility in opto-electronic devices, Adv. Powder Technol. 28 (2017) 1636–1646.
[4] M. Berdova, X. Liu, C. Wiemer, A. Lamperti, G. Tallarida, E. Cianci, M. Fanciulli, S. Frenssila, Hardness, elastic modulus of wear resistance of hafnium oxide-based films growing by atomic layer deposition, J. Vac. Sci. Technol. A 34 (2016), 051510.
[5] M. Aldrigo, M. Dragoman, S. Iordanescu, F. Nastase, S. Vulpe, Tunable microwave filters using HfO2-based ferroelectrics, Nanomaterials 10 (2020) 2057.

[6] P. Torchio, A. Gatto, M. Alvisi, G. Albrand, N. Kaiser, C. Amra, High-reflectivity HfO2/SiO2 ultraviolet mirrors, Appl. Opt. 41 (2002) 3256–3261.
[7] M. Kirm, J. Aarik, M. Jurgens, I. Sildos, Thin films of HfO2 and ZrO2 as potential scintillators, Nucl. Instrum. Methods Phys. Res. A 5337 (2004) 251–255.
[8] O. Dominguez, T.L. McGinnity, R.K. Roeder, A.J. Hoffman, Optical characterization of polar HfO2 nanoparticles in the mid- and far-infrared, Appl. Phys. Lett. 111 (2010), 011101.
[9] 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 HfO2: Ln+3, synthesized by hydrothermal route, J. Lumin. 175 (2016) 243–248.
[10] C.L. Ordóñez-Romero, C. Flores J, J. Hernández A, E. Camarillo G, E. Cabrera B, Manuel Garcia-Hipólito, H. Murrieta S, Effect of the HfO2 sinterization temperature on the erbium luminescence, J. Lumin. 145 (2014) 713–716.
[11] A. Lauria, I. Villa, M. Fasoli, M. Niederberger, A. Vedda, Multifunctional role of rare earth doping in optical materials: nonaqueous sol-gel synthesis of stabilized cubic HfO2 luminescent nanoparticles, J. Am. Chem. Soc. 7 (2013) 7041–7052.
[12] N.C. George, Kristin A. Denault, Ram Seshadri, Phosphors for solid-state white lighting, Annu. Rev. Mater. Res. 43 (2013) 481–501.
[13] R. Martinez-Martinez, E. Alvarez, A. Speghini, C. Falcony, U. Caldiño, Cold white light generation from hafnium oxide films activated with Ce3+, Tb3+, and Mn2+ ions, J. Mater. Res. 25 (2010) 484–490.
[14] S.H. Chaung, H.C. Lin, C.H. Chen, Oxygen vacancy relationship to photoluminescence and heat treatment methods in Hafnium oxide powder, J. Alloy. Compd. 534 (2012) 42–46.
[15] V. Kiisk, S. Lange, K. Utt, T. Tatte, H. Mandar, I. Sildos, Photoluminescence of sol-gel-prepared hafnia, Physica B 405 (2010) 758–762.
[16] R.T. Wen, L.S. Wang, H.Z. Guo, Y. Chen, G.H. Yue, D.L. Peng, T. Hihara, K. Sumiyama, K, Blue-luminescent hafnia nanoclusters synthesized by plasma gas-phase method, Mater. Chem. Phys. 130 (2011) 823–826.
[17] E. Montes, P. Cerón, T.R. Montalvo, J. Guzmán, M.G. Hipólito, A.B. Soto-Guzmán, R.G. Salcedo, C. Falcony, Thermoluminescent characterization of HfO2: Tb3+ synthesized by hydrothermal route, Appl. Radiat. Isot. 83 (2014) 196–199.
[18] K. Fiaczyk, A.J. Wojtowicz, W. Drozdowski, K. Brylew, E. Zych, Thermoluminescent properties of HfO2: Ti after exposure to X-rays, Radiat. Meas. 90 (2016) 140–144.
[19] E. Montes, P. Ceron, J. Guzmán-Mendoza, C. Falcony, M.A. Vallejo, M.A. Sosa, Effect of europium concentration on the photoluminescent and thermoluminescent properties of HfO2:Eu3+ nanocrystals, Ceram. Int. 44 (2018) 8081–8086.
[20] R.R. Manríquez, J.A.I. DíazGóngora, J. Guzmán-Mendoza, T.R. Montalvo, J.C. GuzmánOlguín, P.V. CerónRamírez, M. García-Hipólito, C. Falcony, Photo-, cathodo- and thermoluminescent properties of dysprosium-doped HfO2 films deposited by ultrasonic spray pyrolysis, Appl. Radiat. Isot. 92 (2014) 91–95.
[21] S.T. Chang, S.H. Liao, Light emission and photoluminescence from the high-dielectrics containing Ge nanocrystals, J. Vac. Sci. Technol. B 27 (2009) 535.
[22] A. Sharma, M. Varshney, H.J. Shin, K.H. Chae, S.O. Won, XANES, EXAFS and photoluminescence investigation on the amorphous Eu: HfO2, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 173 (2017) 549–555.
[23] P. Blaha, K. Schwarz, G.K.H. Madsen, D. Kvasnicka, J. Luitz, An Augmented Plane Wave Local Orbitals Program for Calculating Crystal Properties, ISBN 3-9501031-1-2, Karlheinz Schwarz, Techn. Universitat, Wien, Austria, 2001.
[24] P. Perdew, S. Burke, M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. 77 (1996) 3865.
[25] F. Tran, P. Blaha, Accurate band gaps of semiconductors and insulators with a semilocal exchange correlation potential, Phys. Rev. Lett. 102 (2009), 226401.
[26] J. Rodriguez-Carvaial, Full proof: A Program Rietveld Refinement Pattern Matching Analysis, 127, 1990.
[27] P.E. Quintard, P. Barberis, A.P. Mirgorodsky, T.M. Mejean, Comparative lattice-dynamical study of raman spectra of monoclinic and tetragonal phases of Zirconia and Hafnia, J. Am. Ceram. Soc. 85 (2002) 1745–1749.
[28] B.K. Kim, H. Hamaguchi, Raman spectrum of 18O-labelled hafnia, Mater. Res. Bull. 32 (1997) 1367–1370.
[29] C.W. Li, M.M. Mckerns, B. Fultz, Raman spectroscopy study of phonon anharmonicity of hafnia at elevated temperatures, Phys. Rev. B 80 (2009) 54304.
[30] J. Li, S. Meng, L. Qin, H. Lu, T. Tohyama, A small shoulder of optical absorption in polycrystalline HfO2 by LDA+U approach, Solid State Commun. 244 (2016) 28–32.
[31] J.W. Park, D.K. Lee, D. Lim, H. Lee, S.H. Choi, Optical properties of thermally annealed hafnium oxide and their correlation with structural change, J. Appl. Phys. 104 (2008), 033521.
[32] P. Ondracka, D. Holec, D. Necas, L. Zajickaova, Accurate prediction of band gaps and optical properties of HfO2, J. Phys. D: Appl. Phys. 49 (39) (2016).
[33] S. Kumar, S.B. Rai, C. Rath, Monoclinic to cubic phase transformation and photoluminescence properties in Hf1-xSmxO2(x=0-0.12) nanoparticles, J. Appl. Phys. 123 (2008), 055108.
[34] P. Rauwel, E. Rauwel, C. Persson, M.F. Sunding, A. Galeckas, One step synthesis of pure cubic and monoclinic HfO2 nanoparticles: correlating the structure to the electronic properties of the two polymorphs, J. Appl. Phys. 112 (2012), 102107.
[35] T.V. Perevalov, V.Sh Aliev, V.A. Gritsenko, A.A. Saraev, V.V. Kaichev, Electronic structure of oxygen vacancies in hafnium oxide, Microelectron. Eng. 109 (2013) 21–23.
[36] Z. Gangqiang, T. Sanqi, T. Tingting, Electronic structure and optical properties of monoclinic HfO2 with oxygen vacancy, Rare Metal Mater. Eng. 42 (2013) 1576–1580.
[37] D.L. Dexter, J.H. Schulman, Theory of concentration quenching in inorganic phosphors, J. Chem. Phys. 22 (1954) 1063–1070.
[38] G. Kitis, J. Gomez-Ros, J. Tuyn, Thermoluminescence glow-curve deconvolution functions for first, second and general orders of kinetics, J. Phys. D: Appl. Phys. 31 (1998) 2636–2641.
[39] A. Vij, S.P. Lochab, R. Kumar, N. Singh, Thermoluminescence response and trap parameters determination of gamma exposed Ce doped SrS nanostructures, J. Alloy. Compd. 490 (2010) 33–36.
[40] M. Jain, Manju, Savita, A. Vij, A. Thakur, Activation of multiple defects in zinc aluminate through gamma and UV irradiation, Vacuum 180 (2020), 109600.

2021 Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 28-29, 2021

Design and Modeling of a Filter Assisted 6-Pulse VSC-IGBT Based D-Statcom for Reactive Power Compensation

Biswabijet Panigrahy
Dept. of Electrical Engineering
Delhi Technological University
Delhi, India
biswabijet20@gmail.com

Bushra Arif
Dept. of Electrical Engineering
Delhi Technological University
Delhi, India
bushra75arif@gmail.com

Ritul Panchal
Dept. of Electrical Engineering
Delhi Technological University
Delhi, India
ritulpanchal2709@gmail.com

Radheshyam Saha
Dept. of Electrical Engineering
Delhi Technological University
Delhi, India
rshahacno@yahoo.com

Krishna Dutt
Dept. of Electrical Engineering
Delhi Technological University
Delhi, India
krishna@dtu.ac.in

2021 Asian Conference on Innovation in Technology (ASIANCON) | 978-1-7281-8402-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ASIANCON51346.2021.9544618

Abstract— **Utilities are looking for cost-effective solutions to improve the grid's power quality and stability. Applications of FACTS devices provide a popular and proven solution to reactive power compensation, mitigating grid instability, improving power quality and voltage stability. Amongst a plethora of FACTS devices, shunt controllers are the most sought-after power controllers in which the power electronics-based Static Synchronous Compensators or STATCOM is the popular state-of-the-art technology. The increasing penetration of distributed RE (DRE) based electricity generation plays a crucial role in supplying power to loads in remote areas, especially where the main grid system is inaccessible. Such areas are frequently suffering the problems of reliable power supply, power quality, and voltage instability. The Utilization of smart inverters based on dynamic reactive power compensation by Distributed-STATCOM (D-STATCOM) has enabled to qualitatively improve the end-user requirements in the electricity distribution system. The design and behavior of a simple filter assisted D-STATCOM for reactive power compensation during variation of inductive loads is the topic of this paper. A cost-effective ±200kVAr D-STATCOM armed with a novel output filter and improved PI controller employing the pulse width modulated (PWM) technique to IGBT based Voltage Source Converter is proposed to compensate reactive power requirement of loads and to improve transient and dynamic stability, limiting harmonic distortion within standard IEEE-519 limits. The proposed STATCOM design is simulated in Simulink/MATLAB, which confirms its satisfying dynamic reactive power compensation capabilities and low level of harmonic distortion.**

Keywords—FACTS, D-STATCOM, VSC, PI, reactive power compensation, PWM, DRE, TDD

I. INTRODUCTION

The majority of transmission networks today are AC networks, and some main characteristics of ideal AC grid networks include nominal system voltage, unity power factor, and smooth sinusoids of voltage and current[1]. Spikes, harmonics, notches, noise, voltage sags or swelling and flickers are the main power quality problems that are borne because of rush load, fast switching, lighting, etc. which in turn lead to perturbations in the AC grid. Therefore there is a profound necessity for power flow control in power systems[2]–[4].

FACTS controllers maintain voltage profile through reactive power compensation which leads to voltage regulation. The need for more transmission lines is minimized to exchange power economically and reliably in neighboring utilities [1]–[5].The primary objectives in distribution are voltage regulation for loads and power quality [7], [8].

STATCOM is an active and very fast acting shunt controller similar to a controllable reactive current source; its reactive current will rapidly follow changes in the load reactive current and automatically compensate the reactive power required by the grid system, meeting specific grid harmonic requirements.[1], [3], [8], [10]. Due to its DC-link capacitor, real power can be exchanged with the AC system by switching power converters with the support of an adequate DC power source. During a power outage, the STATCOM reacts rapidly by surging output that improves the real power transmitted to the consumer thus minimizing the post-fault effects. On fault clearance, reactive loads come back online and cause voltage sag which it responds rapidly by supplying the necessary reactive power to boost voltage and prevent under-voltage conditions. With STATCOM systems, grid operators gain fast accurate control of the reactive power all with minimal infrastructure investment and low environmental impact[8].The use of VSC opens the avenue for multilevel topologies, and increasing pulse order improves its features [11], [13]. For effective power controllability, near to the ideal fundamental frequency voltage is obtained by low pass filtering. It allows to manage both reactive and active power independently[13].

II. ARCHITECTURE AND OPERATING PRINCIPLE OF STATCOM

The STATCOMS are utilized both in power transmission and distribution grids and categorized into T-STATCOMS and D-STATCOMS respectively. The D-STATCOM is rated up to 5MVAr [8], [10] and IGBTs are the matured technology being operated at PWM mode.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:35:17 UTC from IEEE Xplore. Restrictions apply.

Generally, in a STATCOM as shown in Fig 1; Voltage Source Converter, DC-link capacitor, filter, and coupling transformer are the main components. As indicated in the single line diagram. STATCOM is connected as a shunt controller to a transmission system, where V_S is the voltage produced by a three-phase power supply, V_t is the voltage of the bus connected as an interface between STATCOM, Source and Load where V_{Sh} is the voltage generated by (STATCOM) generating as current I_{Sh} flowing through reactance X_t to supply or absorb reactive power as per Eq (2.1). The three voltage and current phasors are shown in Eq (2.2). The active power and reactive power to be generated by the compensator are defined in Eq (2.3) and Eq (2.4); the reactive power generated shall be meeting the load demand and active power should be almost zero (neglecting losses due to switching)[2], [3], [14].

$$I_{sh} = \frac{V_{Sh} - V_t}{jX_t} \tag{2.1}$$

$$\overline{V} = \begin{bmatrix} V_a \\ V_b \\ V_c \end{bmatrix}; \ \overline{V}_{Sh} = \begin{bmatrix} V_{aSh} \\ V_{bSh} \\ V_{cSh} \end{bmatrix}; \ \overline{I}_{Sh} = \begin{bmatrix} I_{Sha} \\ I_{Shb} \\ I_{Shc} \end{bmatrix} \tag{2.2}$$

$$\mathrm{P_{Sh}} = -V_t V_{Sh} \cdot sin(\theta_t - \theta_{Sh}) / X_t \tag{2.3}$$

$$Q_{Sh} = V_t \ (V_{Sh} \cdot cos(\theta_t - \theta_{Sh}) - V_t) / X_t \tag{2.4}$$

$$\overline{V} - \overline{V}_{Sh} = R_{Sh}\overline{I}_{Sh} + L_{Sh}\frac{d\overline{I}_{Sh}}{dt} \tag{2.5}$$

$$V_d - V_{Sh\,d} = R_{Sh} \cdot I_{Sh\,d} + L_{Sh}\frac{dI_{Sh\,d}}{dt} - L_{Sh} \cdot \omega \cdot I_{Sh\,q} \tag{2.6}$$

$$V_q - V_{Sh\,q} = R_{Sh} \cdot I_{Sh\,q} + L_{Sh}\frac{dI_{Sh\,q}}{dt} - L_{Sh} \cdot \omega \cdot I_{Sh\,d} \tag{2.7}$$

The STATCOM design is based on controlling the power compensation ability of STATCOM by controlling the d-axis voltages and q-axis voltages [14] as depicted in Eq (2.6) and Eq (2.7).

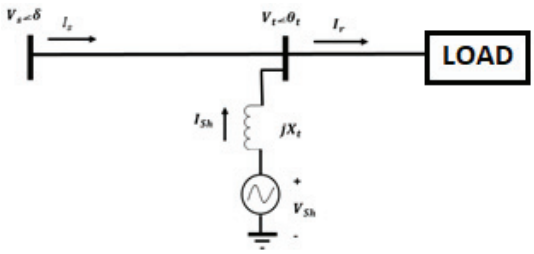


Fig. 1. Single line diagram of STATCOM in the transmission system

The STATCOM's V-I characteristics as shown in Fig 2 elucidates the degrees of compensation independence (in capacitive and inductive mode); which primarily depends on the AC-system voltage and the converter voltage. Over the rated maximum capacitive or inductive range, it can regulate its output current. Since it is capable of delivering a reactive current regardless of the network voltage, it can provide the highest inductive or capacitive current even at low voltage during sudden voltage collapse. The permissible converter's switches junction temperature regulates the device's transient rating in the inductive and area [1], [8], [14].

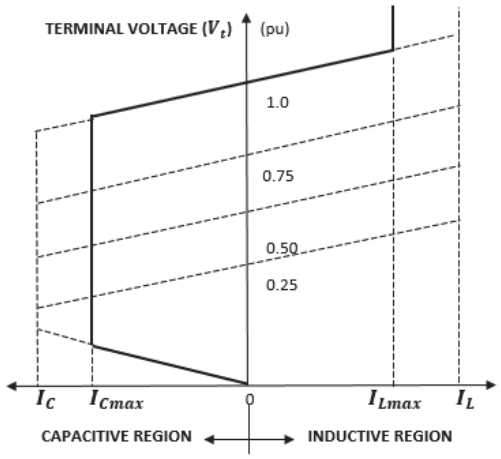


Fig. 2. V-I characteristics of STATCOM

The converter switches provide an interface for directly connecting the AC-output circuit to the DC-input circuit, thereby equalizing their net instantaneous power. The converter generates a circulating reactive-power that is exchanged among the phases. The power exchange occurs between the converter and systems' AC-terminals by utilizing the reactive power absorbed /or supplied by the capacitor[2], [3], [15].The DC capacitor not only provides circulating current paths but also provides a stiff voltage source. [2], [8]. The voltage difference between the voltage bus and the STATCOM determines reactive power exchange. The capacitor voltage is varied in order to change the amplitude of the output voltages as per the basic principles of STATCOM operation [2]. Thus, the magnitude and phase angle of the output voltage can be regulated.

PWM switching applied on IGBT converters can reduce harmonics. The output voltage can be regulated by varying voltage pulse width. The modulation frequency pushes the resulting harmonic content into the kHz range thereby making filter construction cost-effective. PWM can prevent fault current from entering the VSCs during transient conditions triggered by line faults. Therefore, the STATCOM can withstand AC transients without being blocked [5], [13].The fast switching action by supplying appropriate PWM gate signals and switching off through zero or negative gate signals makes IGBT easy to control and therefore used in D-STATCOMS [5], [8], [14]. The PWM signals fed into the convertor are generated by the control circuit which is driven by error signals which are controlled using PI controllers [17].

Fuzzy logic controllers (FLC) in combination with conventional PI-controllers are also used in STATCOM to obtain improved performance of the compensator. Such FLC does not require any mathematical formulation and is formulated based on rule based algorithm, linguistic variable and inference system guided by a set of membership functions [2], [18].

Among the various control strategies adopted in STATCOM [8] , the indirect current control theory (using instantaneous PQ theory) is chosen as the base algorithm to design the control circuit for D-STATCOM which involves direct-axis and quadrature-axis load current and voltage values for generating error signals under Phase Locked Loop control. The PLL is used for tracking the line frequency based on the PI controller and the Park's transformation. The

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:35:17 UTC from IEEE Xplore. Restrictions apply.

stationary reference frame is used to obtain an accurate synchronization to the grid [5], [15].

III. DESIGN

IGBT is a combination of MOSFET topology with an insulated gate leading high voltage and current carrying capability. It is a majority carrier device which turns on by applying a positive base voltage to the channel. They are commutated by anti-parallel diodes, a bilayer fabrication of p-type and n-type semiconductors. It is a bi-terminal device that allows unidirectional current flow and blocks reverse currents. These are primarily used to rectify alternating current, suppress voltage spikes and have fast operation. Voltage ratings range from 50V-5kV and current ratings range from 1 A to 3500A [1]

The block model of the proposed filter assisted 6-pulse PWM VSC IGBT based ±200kVAr D-STATCOM design using indirect current control technique is presented in Fig. 3. The corresponding SIMULINK model for operating the device at grid frequency of 50 Hz is shown in Fig 4. The circuit design is comprising of four sections as depicted below.

A. Power Circuit:

This circuit consists of a 415V three phase AC supply connected with variable R-L load at the voltage bus which is connected to the STATCOM shunted by the LCL filter.

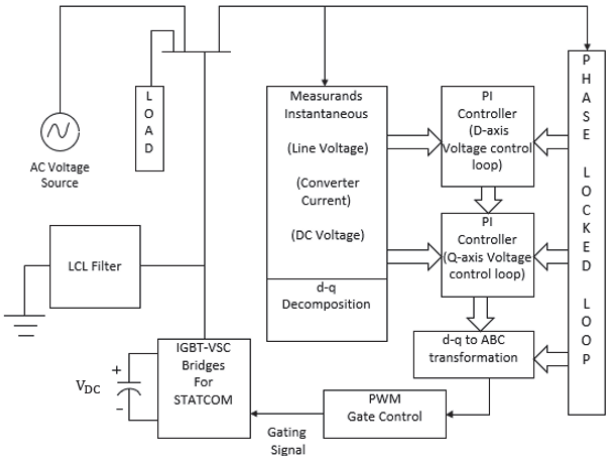


Fig. 3. D-STATCOM diagram using Indirect Current Control technique

1) Capacitor Design:

The capacitor voltage rating (V_{dc}) is calculated on the basis of maximum DC-bus voltage and PWM switching strategy to ensure low ripple currents using the following equations.

$$V_{DC} = \frac{2 \cdot V_{ll} \cdot \sqrt{2}}{m \cdot \sqrt{3}} \quad (3.1)$$

$$I_{RMS} = \frac{kVAr \times 10^3}{\sqrt{3} \times V_{ll}} \quad (3.2)$$

$$C = 0.6124 \times \frac{I_{RMS}}{f_{SW} \times \Delta V} F \quad (3.3)$$

where, V_{ll} (supply line voltage) = 415V, m=modulation index (84.7%), V_{dc}=800V, ΔV = ripple voltage (1%)

Using Eq (3.2), the rms current is calculated as 278 A for the proposed 200kVAr rated D-STATCOM.

Considering the safety margin, 1000V is chosen as the voltage rating of the DC bus. Using Eq. (3.3), where ΔV = ripple voltage= 1% of 800V=8V and f_{sw} = 10kHz (PWM switching frequency), and capacitor rating= 2130μF. Two 4260μF, 500V capacitors being readily available are considered in series [15].

2) LCL Filter Design:

The designing a low pass filter as LCL filter filters out high frequency harmonics, assuming current fluctuation of 8.5% on account of very low ripples produced by the D-STATCOM and a base capacitance ($C_{b)}$ of 2000 μF. The design parameters are worked out by using the following equations:

$$I_{max} = I_{rms}.\sqrt{2} \quad (3.3)$$

$$\Delta I_{Lmax} = 8.5\% \cdot I_{max} \quad (3.4)$$

$$L_1 = \frac{V_{DC}}{6 \cdot f_{sw} \cdot \Delta I_{Lmax}} \quad (3.5)$$

$$C_f = 0.05 \cdot C_b \quad (3.6)$$

As per the calculation I_{max}=393.15A; ΔI_{Lmax}= 33.418A; With these values, inductance connected to the converter (L_1 = 500 micro henry) and filter capacitance (C_f = 100 μF) are calculated.

$$\omega_{res} = \sqrt{\frac{L_1 + L_2}{L_1 \cdot L_2 \cdot C_f}}; \quad f_{res} = \frac{\omega_{res}}{2\pi}; \quad (3.7)$$

$$10 f_g < f_{res} < 0.5 f_{sw} \quad (3.8)$$

The value of L_2 (inductor connected to voltage bus) is calculated to be 272μF using Eq 3.7 and frequency limitations as presented in Eq 3.8 [19].

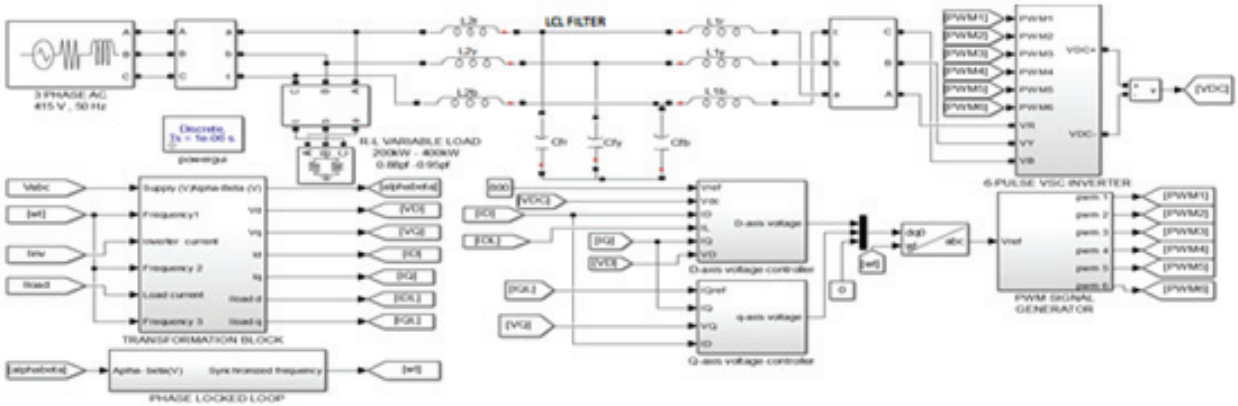


Fig. 4. Simulink model of 6-pulse STATCOM

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:35:17 UTC from IEEE Xplore. Restrictions apply.

B. *Transformation block:*

This transformation block shown in Fig. 4 receives voltage and current values of AC supply and load, and using the synchronous frequency generated by the PL loop, it produces instantaneous quadrature and direct axis voltages and currents based on the Park's and Clarke's transformation.

C. *Control Circuit:*

The control circuit is designed based on an indirect current control approach and instantaneous PQ theory principle. Revised d-axis and q-axis voltages are calculated using two controller circuits viz. direct and quadrature axis voltage controllers as shown in Fig. 5 and Fig. 6. The voltage controller compares the instantaneous voltage of the DC – link capacitor with a reference value set at 800V. The error is amplified by a tuned PI controller which is then fed into the (ID) direct axis current controller and compared with source and load ID current; this error is amplified by PI block and further undergoes signal conditioning to produce revised d-axis voltage. The current measurement block compares the instantaneous q-axis current with the reference load q-axis current which is modified and amplified to generate the revised q- axis voltage.

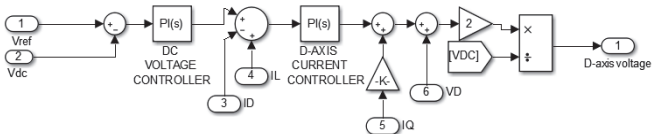


Fig. 5. D-axis voltage controller circuit design

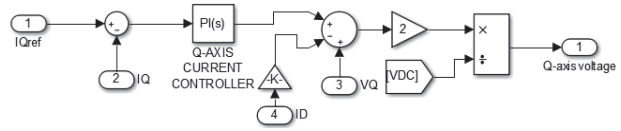


Fig. 6. Q-axis voltage controller circuit design

D. *Pulse Width Modulation Block:*

The 'sine-triangle' method, in which a triangle waveform is superimposed on reference three-phase sinusoidal waveforms, is used. The switches are triggered once the carrier triangle waveform crosses the reference sinusoidal waveform, this block uses the revised d-q voltages to produce the revised 3- phase reference voltages which are used to generate PWM gate pulses fed to the IGBT switches of D-STATCOM [5].

IV. Result And Discussion

Fig. 7 exhibits the quantum of reactive power compensation provided by the proposed D-STATCOM corresponding to R-L loads of 200kW, 300 kW and 400kW with its power factors ranging from 0.9 to 0.95.

It is noted that the D-STATCOM compensates effectively the load reactive power requirement at different load power factors and the quantum of compensation is given in Table- II.

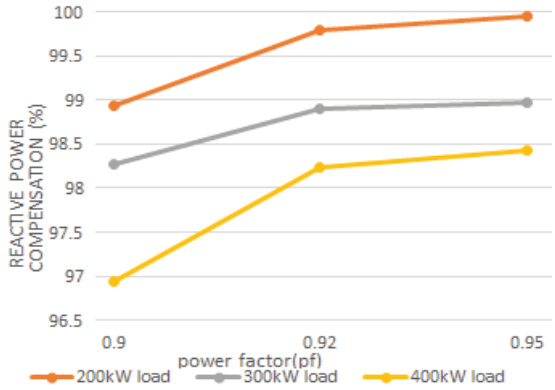


Fig. 7. Reactive power compensation at various R-L loads

Fig 8 explains the TDD characteristics at various R-L loads and it is observed that the system design fairly eliminates current harmonics at high loads and low power factors with TDD values ranging within acceptable level [20] from 2.99% to 4.6%.

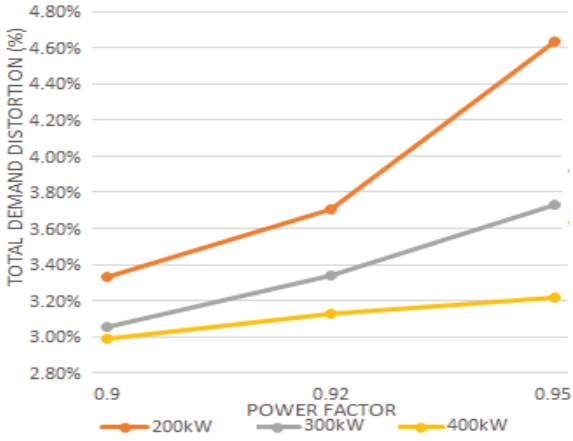


Fig. 8. Current THD at various R-L loads

The harmonics reduction happens to be due to the LCL filter which enables to reduce the dominant 5th and 7th harmonics as shown in Fig 9. The TDD is 2.99% at 400kW (0.9 pf) whereas without the filter, TDD surges to 115.71% . The LCL therefore reduces TDD by approximately 97.4%. The reduction in 5th and 7th order harmonics are mentioned in Table-III at various loads .

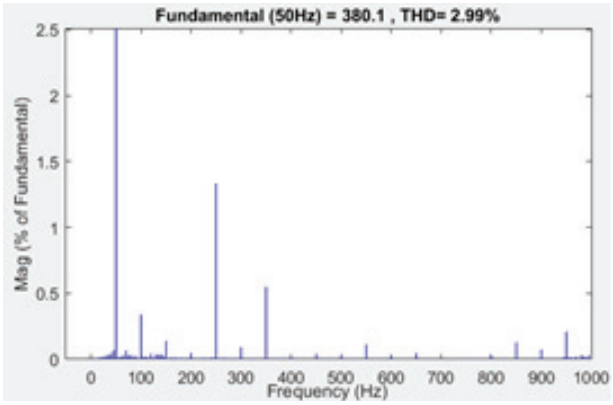


Fig. 9. TDD of inverter current using LCL filter

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:35:17 UTC from IEEE Xplore. Restrictions apply.

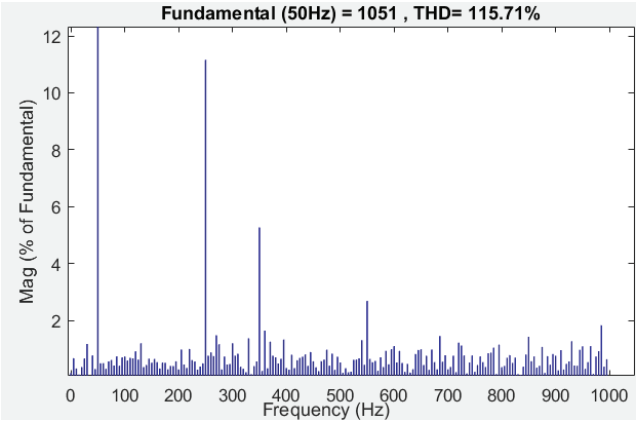


Fig. 10. TDD of inverter current without using LCL filter

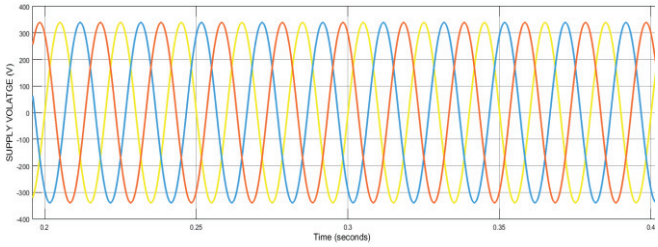


Fig. 11. Three phase supply voltage waveforms (V_{abc})

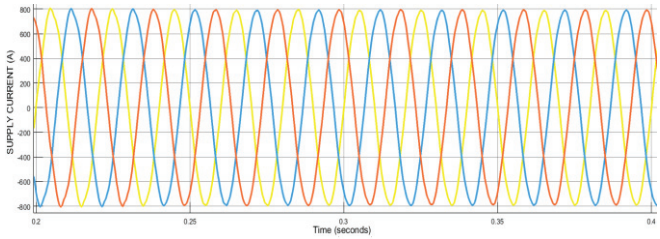


Fig. 12. Three phase supply current waveform (I_{abc})

Fig 11 is the three-phase voltage waveform of supply voltage which attains a peak value at 338V. Fig 12 explains the three phase current waveforms which attains peak value of 800A and is nearly in phase with supply voltage.

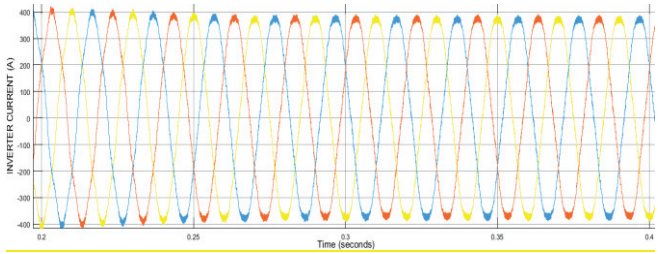


Fig. 13. Three phase converter current waveform(I_{inv})

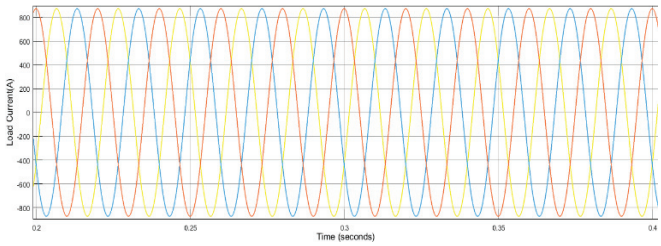


Fig. 14. Three phase load current waveform (I_{Load})

The converter current waveform shown in Fig 13 attains steady state at approximately 0.27s with steady state value of 390A. From Fig 14, it is observed the proposed D-STATCOM operates at 0.9 pf; I_{inv} lags I_{Load} by a very small margin as observed at time intervals 0.2s and 0.25s.

The DC voltage waveform shown in Fig 15 attains an overshoot of 1620V at 12ms and settles at approximately 0.23s at 800V. The ripple factor of this waveform is [(804-796)/800] *100 = 1%.

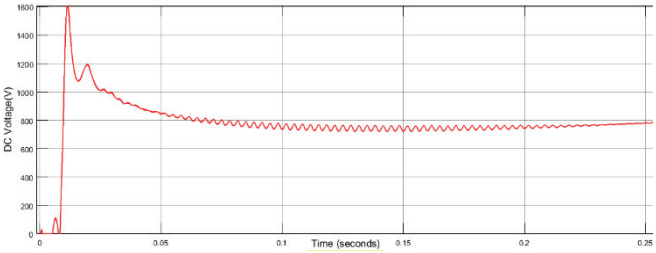


Fig. 15. DC Voltage waveform (V_{DC})

Fig 16 displays exclusively the three-phase voltage waveform of the 6-pulse IGBT-VSC based D-STATCOM operating at switching frequency of 10kHz. The gate signals are generated from Sinusoidal Pulse Signal Generator, and the voltage waveform settles at approximately 0.23s at 521V. Due to almost negligible reactance between converter and the source, the waveform is nearly in phase with the supply voltage.

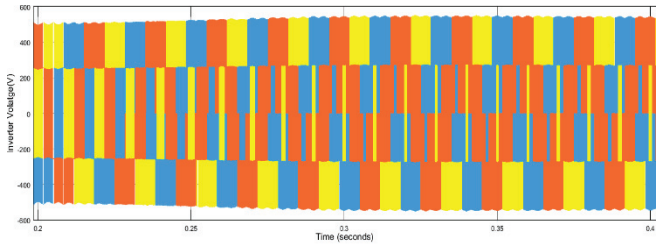


Fig. 16. Three phase converter voltage waveforms (V_{inv})

The active (P) and reactive (Q) power components in respect of d- axis and q-axis voltage and current are defined by the following equations:

$$P = \frac{3}{2}(V_d I_d + V_q I_q) \tag{4.1}$$

$$Q = \frac{3}{2}(V_q I_d - V_d I_q) \tag{4.2}$$

Fig 17 shows the d-axis current component which is approximately zero at steady state. The q-axis current waveforms as depicted in Fig 18, attains steady state value of 400A at 0.21s. The d-axis voltage waveform attains a steady state value of -338V in 2.5ms as indicated in Fig 19 (since the load is inductive and D-STATCOM shall be supplying reactive power). The q-axis voltage waveform settles to zero in 3ms as displayed in Fig 20 indicating thereby the fast response of the circuit to attain steady state in about one cycle (20ms) [15].

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:35:17 UTC from IEEE Xplore. Restrictions apply.

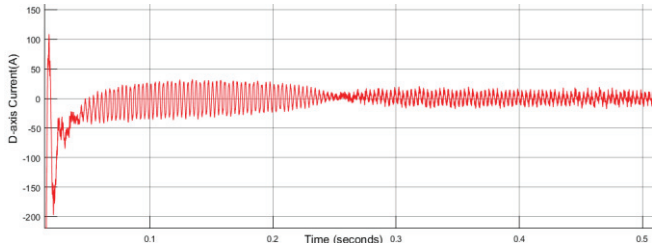


Fig. 17. D-axis current waveform

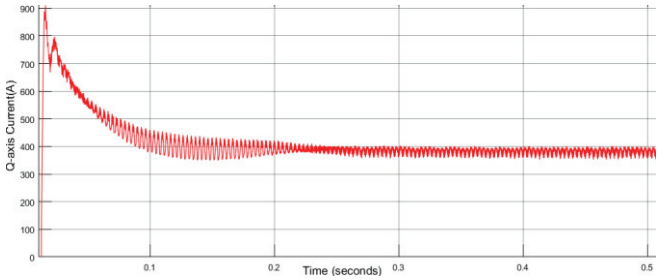


Fig. 18. Q-axis current waveform

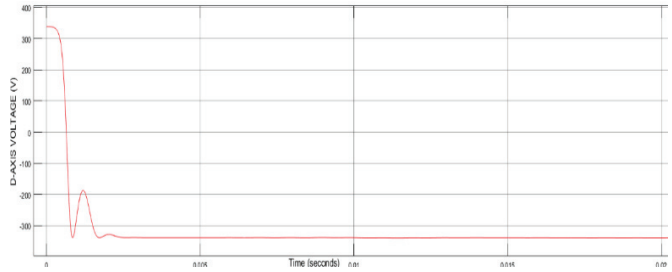


Fig. 19. D-axis voltage waveform

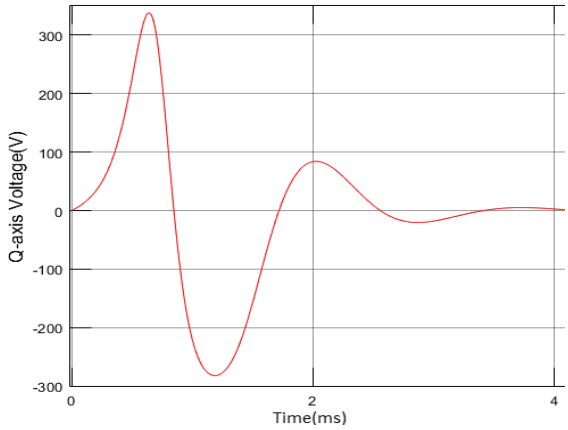


Fig. 20. Q-axis voltage waveform

TABLE I. DESIGN PARAMETERS OF A STATCOM

DEVICE	PARAMETERS	VALUES
SOURCE	Source volatge	415V
	Configuration	Y type
	Source resistance	1 μΩ
	Source inductance	1 μH
	Phase angle	0 degrees
	System frequency	50 Hz
CONVERTER	Power switch internal resistance	1mΩ
	Snubber resistance	1MΩ
	Switching frequency	10kHz
	Capacitor rating	Two 4260μF in series
	Capacitor resistance	1mΩ
FILTER	L_1	500 μH
	L_2	272 μH
	C_f	100 μF
	f_{res}(Hz)	1200Hz
PI CONTROLLER GAINS	DC VOLTAGE CONTROLLER (K_P,K_I)	(5,200)
	OUTPUT CURRENT CONTROLLER (K_P,K_I)	(25,500)
	INPUT Q-AXIS CURRENT CONTROLLER (K_P,K_I)	(25,500)

TABLE II. TABULAR PRESENTATION OF SIMULATION RESULTS

LOAD DEMAND (kW)	POWER FACTOR (p.f)	Q SUPPLIED BY STATCOM (kVAR)	PERCENTAGE COMPENSATION (%)	CURRENT TDD (%)
200kW	0.90	95.83	98.932	3.33%
	0.92	85.02	99.788	3.71%
	0.95	65.70	99.944	4.64%
300kW	0.90	142.80	98.280	3.05%
	0.92	126.39	98.897	3.34%
	0.95	97.59	98.971	3.73%
400kW	0.90	187.79	96.934	2.99%
	0.92	167.40	98.240	3.13%
	0.95	129.40	98.422	3.22%

TABLE III. TABULAR PRESENTATION OF HARMONICS WITH AND WITHOUT LCL FILTER AT VARIOUS LOADS

LOAD (kW)	POWER FACTOR (p.f)	WITH LCL FILTER			WITHOUT LCL FILTER		
		TDD	5^{TH} Harm	7^{th} Harm.	TDD	5^{TH} Harm.	7^{TH} Harm.
200kW	0.9	3.33 %	0.07 %	0.145 %	164.55 %	21%	11%
	0.95	4.64 %	0.4%	0.25 %	169.68	22%	12%
300kW	0.9	3.05 %	0.8%	0.5%	131.33	15%	7%
	0.95	3.73 %	1.5%	0.7%	135.63	16%	8%
400kW	0.9	2.99 %	1.4%	0.55 %	115.71 %	11%	5%
	0.95	3.22 %	1.8%	0.8%	119.86 %	12%	6%

V. CONCLUSION

This paper has provided a simple approach to design and modeling of ± 200 kVAr D-Stacom using 6-pulse IGBT based VSC triggered with 10kHz PWM gate signals, and use of a low pass LCL filter. The device has been simulated in MATLAB/Simulink platform for reactive power compensation to the varying inductive loads. The results validate the fast response of the compensator for load reactive power balancing in electrical network. With the proposed design of the LCL filter and the compensator, the current TDD level has also been maintained as stipulated in IEEE-519 standards [20]. A prototype development of the proposed model can be undertaken and tested for validation as next phase of work. Further scope of improvement in the performance of STATCOM can be studied and modeled through multi-staging converter topologies and self-adaptive controllers like AI controller for versatile applications in transmission and distribution networks.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:35:17 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] L. G. N.G. HINGORANI, UNDERSTANDING FACTS CONCEPTS AND TECHNOLOGIES OF FLEXIBLE AC TRANSMISSION SYSTEMS, vol. 6, no. 11. 1967.

[2] G. Gupta, W. Fritz, and M. T. E. Kahn, "A comprehensive review of DSTATCOM: Control and compensation strategies," Int. J. Appl. Eng. Res., vol. 12, no. 12, pp. 3387–3393, 2017.

[3] H. A. Hassan, Z. H. Osman, and A. E.-A. Lasheen, "Sizing of STATCOM to Enhance Voltage Stability of Power Systems for Normal and Contingency Cases," Smart Grid Renew. Energy, vol. 05, no. 01, pp. 8–18, 2014, doi: 10.4236/sgre.2014.51002.

[4] Aishvarya Narain and S. K. Srivastava, "An Overview of Facts Devices used for Reactive Power Compensation Techniques," Int. J. Eng. Res., vol. V4, no. 12, pp. 81–85, 2015, doi: 10.17577/ijertv4is120071.

[5] K. Wang and M. L. Crow, Modern flexible AC transmission system (FACTS) devices. Woodhead Publishing Limited, 2013.

[6] J. Liu, J. Chen, and Z. Qian, "Comparative analysis of FACTS devices based on the comprehensive evaluation index system," MATEC Web Conf., vol. 95, pp. 1–5, 2017, doi: 10.1051/matecconf/20179515002.

[7] B. Oum, E. Fadhel, and M. Fellah, "Theory , Modelling and Control of FACTS devices THEORY , Modelling and Control of FACTS devices . Laboratoire ICEPS (Intelligent Control & Electrical Power Systems)," no. February, 2015.

[8] A. ÇETİN, "Design and implementation of a voltage source converter based statcom for reactive power compensation and harmonic filtering," Pravoslavie.ru, no. April, 2007.

[9] S. Kumar, "controllers using Evolutionary Techniques," pp. 962–965, 2020.

[10] B. S. Pali, S. Bhowmick, and N. Kumar, "Newton-Raphson power flow models of static VAR compensator," India Int. Conf. Power Electron. IICPE, pp. 9–13, 2012, doi: 10.1109/IICPE.2012.6450443.

[11] T. Navpreet et al., "Voltage Source Converters as the building block of HVDC and FACTS Technology in Power Transmission System: A Simulation based Approach," Pelagia Res. Libr. Adv. Appl. Sci. Res., vol. 3, no. 5, pp. 3263–3278, 2012, [Online]. Available: www.pelagiaresearchlibrary.com.

[12] S. Bernet, R. Teichmann, S. Member, A. Zuckerberger, and P. K. Steimer, "Comparison of High-Power IGBT ' s and Hard-Driven GTO ' s for High-Power Inverters," vol. 35, no. 2, pp. 487–495, 1999.

[13] T. Mittal, N. S. Tung, A. D. Diwan, and K. Janardhana, "STATCOM Analysis for improving Power Quality Constraints - A Literature Survey," vol. 1, no. 1, pp. 1–10, 2012.

[14] A. Mokhtari, F. Z. Gherbi, C. Mokhtar, and D. E. Kamel, "Study , analysis and simulation of a static compensator D-STATCOM for distribution systems of electric power," Leonardo J. Sci., no. 25, pp. 117–130, 2014.

[15] J. K. Bhutto and R. Bana, "Design of the Dstatcom and Analysis of the Controllers for Power Quality Improvement," pp. 24–37.

[16] C. Wang, X. Yin, M. Wen, J. Liu, Q. Xiong, and B. Zhang, "Structure and parameters design of output LC filter in D-STATCOM," 2010 Int. Conf. Power Syst. Technol. Technol. Innov. Mak. Power Grid Smarter, POWERCON2010, pp. 1–6, 2010, doi: 10.1109/POWERCON.2010.5666674.

[17] K. S. Rao and R. Mishra, "Comparative study of P, PI and PID controller for speed control of VSI-fed induction motor," vol. 2, no. 2, pp. 2740–2744, 2014, [Online]. Available: http://www.ijedr.org/papers/IJEDR1402230.pdf.

[18] A. Oudalov, R. Cherkaoui, and A. J. Germond, "Power Flow Control By Multiple Series Facts Devices," System.

[19] M. Dursun and M. K. Dosoglu, "LCL Filter Design for Grid Connected Three-Phase Inverter," ISMSIT 2018 - 2nd Int. Symp. Multidiscip. Stud. Innov. Technol. Proc., no. 3, pp. 2018–2021, 2018, doi: 10.1109/ISMSIT.2018.8567054.

[20] "IEEE Recommended Practice and Requirements for Harmonic Control in Electric Power Systems," in IEEE Std 519-2014 (Revision of IEEE Std 519-1992) , vol., no., pp.1-29, 11 June 2014, doi: 10.1109/IEEESTD.2014.6826459.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:35:17 UTC from IEEE Xplore. Restrictions apply.

2021 Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 28-29, 2021

Detection, Localization and Retrieval of tampering/Alteration in Microsoft Word Documents for Authentication and Copyright Protection using Multiple Techniques

Manmohan
Department of Electronics and Communication Engineering.
Delhi Technological University
Delhi, India
mathurmanmohan99@gmail.com

Lavanisha Prasad
Department of Electronics and Communication Engineering.
Delhi Technological University
Delhi, India
lavanishaprasad@gmail.com

Nitin
Department of Electronics and Communication Engineering.
Delhi Technological University
Delhi, India
dhemiwal.nitin@gmail.com

Jeebananda Panda
Department of Electronics and Communication Engineering.
Delhi Technological University
Delhi, India
jpanda@dce.ac.in

2021 Asian Conference on Innovation in Technology (ASIANCON) | 978-1-7281-8402-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ASIANCON51346.2021.9545046

***Abstract*— The Covid-19 Pandemic has changed the nature of work culture and made everything digitized. While this has many advantages, making everything online brings with it security threats such as ensuring authenticity of the documents. It is often seen that the documents transferred from one source to another may undergo changes which tampers the original work of the author. We have taken this challenge as an opportunity to develop techniques to Detect, Localize and Retrieve any forgery that may be attempted to an original document wherein there is removal, replacement or addition of a word in the document without the need of original document. We have incorporated the method of hashing along with encryption and decryption to accomplish the techniques to its fullest efficiency.**

Keywords*— Detection, Forgery, Hashing, Localization, Retrieval.

I. INTRODUCTION

The Internet has changed our life drastically. From the basic copying of data to sharing files and digital media on the internet has become a common practice in every sphere of life. Most resources on the internet are in the form of text which is often used for articles, websites, e-books, research papers etc[1]. Cyber attacks are increasing and becoming very dangerous as our reliance on information technology grows[2].In addition to this, the pandemic has further increased sharing of documents online. This brings to limelight the problem of copyright infringement and authenticity of documents for the authors putting up their work on the internet. We have channeled our focus on the most common forgeries such as the removal of content, replacement of words and addition of words to the original document that could tamper the original work of the author online.

To overcome these challenges, we have developed a technique to make online transfer of documents more secure and authentic.

Main points covering contribution to this research are:

1. The proposed work promises a method where forgery in a document can be detected if a word or special character is removed, replaced or new one is added. The objective is to Detect, Localize and Retrieve the words that tampered with the document and display it.
2. The technique uses fernet encryption for encryption and decryption to attain the objective.
3. The technique is robust up to 99.99% and is secured by security lock i.e. key
4. This technique can be used to check authenticity of monocoloured as well as colored MS Word Document.

Our work is presented as follows: Section II contains related work in this field done by other researchers. Section III explains the proposed method in detail which is followed by Section IV showing the Algorithms and working. Section V gives the results obtained. Lastly Section VI concludes the paper and paves way for future work.

II. RELATED WORK

Copyright protection in text documents is an active area of research. There is a lot of work that has been done in copyright protection of audio, video and image files[3]. In case of text documents copyright protection and authentication some interesting work has been done by many scholars using various approaches. Drishti Agarwal et al. [3] has proposed a

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:36:55 UTC from IEEE Xplore. Restrictions apply.

technique in which they implemented and Integrated multiple watermarking techniques to detect any Alteration happened after watermark was embedded in the document hidden and imperceptible to viewer and can detect if text has been tampered with good robustness and imperceptible to humans eyes. However, they were unable to localize alterations and retrieve the original content. This paper extends the work to implement localisation i.e. getting the location of the alterations and also to retrieve the original content.

N. Nikolaidis and I.Pitas [4] in their paper about digital picture copyright protection implemented a technique of embedding an invisible, digital signature for protection of copyright infringement.

In the paper by J.T Brassil et al.[5] the authors have described and compared several methods of marking texts in other ways after the documents have been altered in the most common forms of distortion. Marks are designed to protect documents of limited value held by individuals who may wish to have an official copy rather than an unlicensed copy if they have not separated. For attacks on text documents, the author describes measures to combat the attacks which remove the protection marks or signals embedded in documents.

Sukhpreet kaur and Geetanjali babbar [6] introduced a zero watermarking process that uses the frequency of characters occuring in the name of watermark production.

Zunera Jalil et al.[7] introduced an open text documents algorithm based on watermarking technique to protect it from alteration in its content. Embedding the watermark logically hidden within the text was the baseline of the introduced algorithm and the performance was tested in 12 text samples where the algorithm dispersed the attack and the extracted image can prove the ownership.

Suganya Ranganathan et al.[8] in her paper, Combined Text Watermarking introduced an efficient approach on combining image-based and semantic watermarking for copyright protection.

In Zunera Jalil and Anwar M. Mirza [9] paper presented a review of specific research conducted on watermarking techniques to find text texts in a clear scan. The revised methods are divided into three sections, the image analysis method, the open space opening method and the semantic analysis method. The paper also discussed important issues, contributions, benefits and barriers of the various types that were used to record watermark in the past.

Omar Tayan et al.[10] presented a new zero-based method based on zero-watermarking of sensitive documents to achieve content verification and authentication. Suggested Algorithm worked to embed the watermark logo of the first publisher in the same part of the cover text and feature feature. The algorithm can be used to protect all digital text content from fraud and deception of illegal content.

Afify [11] has proposed a hybrid approach in which they implemented the concept of natural language processing along with hashing and structural technique. They have used the most occurring word in the document for their watermark embedded in the word document. They have altered the spaces towards the left of the word to insert a bit in that word along with the watermark. They have also implemented MD5 hashing technique to hash the complete document to check later if the document has been tampered or not.

Milad Taleby [1] presented a comparative analysis of data encryption techniques, especially those that focus on changing the structure and content of digital texts. Here, various aspects of writing and steganography are shown along with their various programs. Various aspects of writing and steganography are highlighted along with their applications. In addition, various types of content attacks are described and their effects are also analyzed to reflect the pros and cons of current strategies over other options.

III. THE PROPOSED METHODS

In this section, the detailed process of Detection, Localisation and Retrieval of text documents is presented below. To begin the process, we parse the original document word by word and create two parallel files called original hashed doc and encrypted file by SHA3 hashing[12] and fernet encryption algorithm. We save the encrypted and original hash file to check the authenticity of our work, which is protected and locked with a key. Further, we create another file, called forged hash doc, by iterating content of the forged document, hashing and embedding each word. This file will be used to compare with the original hash file to detect any forgery in the document.

A. Detection

To detect the forgery of the document, we considered 3 cases separately where forgery can happen:

- Removal of a word or special character.
- Replacement of a word or special character.
- Addition of a word or special character.
- Alteration of word.

To detect forgery in a document, we compare the two files original hash doc and forged hash doc. On comparing the two hashed files word by word using algorithm 3, we can detect whether the document is forged or not. Further to find the location of the word tampered with and the word itself, we use localisation and retrieval.

B. Localisation

Once we detect that the file has been tampered with, we need to localise the alteration i.e. word removed, replaced, concatenated, spacing changes, upper case, lower case or added. For this we use localisation and in this method, we compare the two hashed files original hash doc and forged hash doc. If a word has been deleted, replaced or added, then this method will return the index value of the word before which the word was deleted in case of deletion. In case of replacement, the exact index value of the word replaced is returned. When the forgery is done in terms of addition of word or special character, then the index value of the word before the added word would be returned. Further we are interested in what exact word has been tampered with, for this we use retrieval method.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:36:55 UTC from IEEE Xplore. Restrictions apply.

C. Retrieval

In order to retrieve the word that has been deleted, replaced or added, we use the index number obtained on comparing original hash doc and forged hash doc in the localization method. Further we use the encrypted file and decrypt the tampered/altered word using the key. We can obtain the word from the index values obtained earlier and get the exact word that was deleted, replaced or if a new word was added before it.

This method completes the process cycle of detection, localization and retrieval which together can help us determine if the document has been tampered with and also gives what words have been deleted, replaced or if any new added at that index.

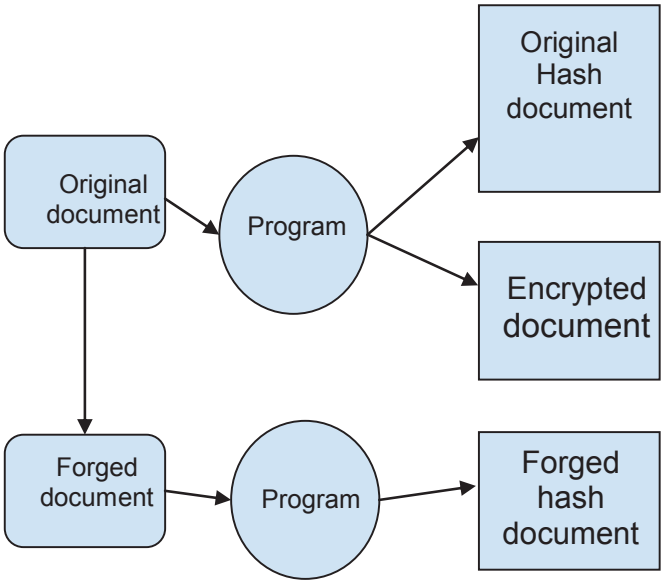


Fig. 1. Architecture of generating the files.

IV. Algorithms and Working

Algorithm 1: Original hash file generation and encrypted file.

Input: Original document file
Data: Secret Key

Step 1: Iterate through each word of the original document file.
Step 2: Generate hash and encrypt the word using key.
Step 3: Embed generated hash of each word into a new word file.
Step 4: Embed each encrypted word into a new text file.

Output : Original hash document and encrypted file.

After the generation of encrypted and hashed documents. The encrypted words can be decrypted only if the encrypted file and key is available. Whereas the hashed words cannot be decoded and is used here only to detect and localize the forgery. Following are the images of original, hashed and encrypted documents.

Video provides a powerful way to help you prove your point. When you click Online Video, you can paste in the embed code for the video you want to add. You can also type a keyword to search online for the video that best fits your document.

Fig. 2. Content of original document.

1a023cb0fad312495c692fd8f2a03ecc9b38a72b0f8034bd8be553093bae44f4
237eaf31a9063d6278e710c4b7fd09920a8037c994c36dbd45c701e4c5d4a32f
c614d7757e7c429c19b43cb6e878bf0ac847e5bad85ed72708629c9692e5100e
afcfa78c8072ee60fbb7dca0f27f99da27327096e34ca8a6090a4039811135c3
a5c4146cff3bdf10d71bb94d827f7c6fab14436c4904775f4e28c35f73c7316a
56a4d43b5eb7d56483b3aad01d65be880b117bcaed28d4b1fa4e0bd3b3f0a7d0
0c0958cc4cecbad13977aa1cc361ffed08c754e065fab4a4e186a597389d8153
2a91037fbd3d1c1016eb630dc47cd758cb1f8ceda13e91d752f2ae7ee0b35d61
c973737e07e7b202c7df447e1c9affcea95074275cc3b5896f474472459280fa
19a24fda394075a0a8140df0c82eee1bc9057e62df8e666934dbea3a656e58b5
8f970ad66f2446cde7237abffebf929065b4a5d6fc313e02940e611493419c36
b4e5766ffe13f751775e0e0633534f618cf2c0b0de70744f42582a4d56ba1556
54a630d2d12538bacee1117c3d65a9f404418e597169578ff0b8681b68d2c9d4
2a91037fbd3d1c1016eb630dc47cd758cb1f8ceda13e91d752f2ae7ee0b35d61
355047701c814b3a7e788da03f0292113735fdb339bd8af1896a4a3de4bbaf2d
ed2965fbdd06510d978f6d624e345bf9aa64ab47d0b906d2985496646b792362
1e3bdb60f243140a6ca8d8adf0c7217ea24333baae6c5ceb28a823ee7e5e70be

Fig. 3. . Contents of original hashed document.

File Edit Format View Help

gAAAAABgiAUK4Dg5XZE6tt8Im9sHhKNeiOjK4nlvsLtj9Y
Aivsri4Is7twsBm1L-JwMme8wwUMx8q4XP-0XCyU_7dY2MD
3yqQA== gAAAAABgiAUK3E_coc70TfLXcwTyqKqTzz5r0c9
45Eu6GMRzhq96ySb4YiVikQPran1sKexSzkrlb3bJV6IFX7
Je8cdU0YSrOQ== gAAAAABgiAUKsS33wjNzpefJnTisJrR_
bL4xLgzuzF5s9aFgaViPXcfFVXLcCcVKsbGHmTT3N9vsPnt
ifCXCg8lb2KcyBv0A4w== gAAAAABgiAUKOIlE2Whqebt5w
bcRyd7dEEPQS-GPHkB1IqP2lOI2mJUBTX4NFLVQ5r_UK6xI
zenr7ItyBNmlZ3_5nUjuhMWhPQ== gAAAAABgiAUK-Imyqr
l5P0O-hxPZVFCeos_UR-JHGMl0FzvVvIbY44cpdv3WXH79k
lHLVz5phxsuzODppCOFyi2rME_nIRt4ZQ== gAAAAABgiAU
KiJ6Ys3ixODRqpD4Hkya00jCXXlzc1zEeF2GfTaUdqPRAoJ
Qpsq68jgX831k4kbKBN9WGNNn5eIIXe8q038Y5sA== gAAA
AABgiAUKtQsplRHe3aIuudTuA_tritzjSThPhKzoAcPjHgW
0UiNFGx2Xn2nVmRD7DiIIXoCKx0scQ9zllXrJyVnwbejriw
== gAAAAABgiAUKD4yFnUxddmBwRNawPPX-JBFYT2YdxE3v
MKCFnhVkXqE844-obieYoi3ov_bHPDLnT4CVSzUd8YejVCd
ezjIbPA== gAAAAABgiAUKfHAwfwAxZNO9AlcigGaiNoGLS
cO7xmravS8xKMQj2qhdbN-0Wn5CJYAm-DmPziTrwrkeyC2c
y7VXNecZiLBcCA== gAAAAABgiAUK_t2DAIw0c6o_4nqPzZ
ScKFfomm0dMG5_9qzYsXqPm8se-UPnJVz_7PN9NrpVI3vV8
VzH6exP2Zg1KYivy9_Bbw== gAAAAABgiAUKeK_uxTOV5AF
cLOjZyhlIZSg5SIUqunqKyyMmteTEoXGvhNnox9Rw1ee7g_
MbTmhQz6a0lqmT_If7j7Hq_wrxPw== gAAAAABgiAUK3EAU
XtzNo48dmEMVMjCVZWI2Tk4MXp_Cx8rj-5oW4xK7RQo21yx
Z5daE1qgsgbu3nB2kl25tY iGkytxGVAxmw== gAAAAABgi

Ln 1, Col 2 | 100% | Windows (CRLF) | UTF-8

Fig. 4. Contents of encrypted document

If the original document is forged i.e. if any deletion, addition and replacement of text has occurred that file will be considered and referred as the forged document for further process.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:36:55 UTC from IEEE Xplore. Restrictions apply.

Algorithm 2: forged hash file generation and encrypted file.

Input: Forged document file

Step 1: Iterate through each word of the forged document file.
Step 2: Generate the hash of each word.
Step 3: Embed generated hash into a new word file.

Output : Forged hash document.

After generation of forged hash document. The original hash document is compared with the forged hash document to detect and localize the changes, once the changes is located then index at with the change has occurred is noted and the word on the same index is fetched from the encrypted document which is decrypted using the highly secured key and returned by the program.

Algorithm 3: Detection, localization and retrieval in case of deletion.

Input: Original hash document, forged hash document and encrypted file.

Step 1: Compare total number of words in original hash document and forged hash document.
Step 2: If the number of words are unequal this means the formatting attack has occurred.
Step 3: Open original hash document and forged hash document and compare each hash word from both the files.
Step 4: If the hash is unequal this means at the particular location forgery has occurred**(Detection).**
Step 5: Note the index of unmatched hash and return the index**(Localisation).**
Step 6: Open the encrypted file and go to the identified index and decrypt the specific word using the same key as in algorithm 1.**(Retrieval).**
Step 7: Return the index and original word which was deleted.

In a different case if additional words have been added in the original file, algorithm 4 will take reference of forged hash document with original hash document and separate out and print the added words along with the index.

Algorithm 4: Detection, localization and retrieval in case of addition.

Input: Original hash document, forged hash document and forged document.

Step 1: Compare total number of words in original hash document and forged hash document.
Step 2: If the number of words are unequal this means the formatting attack has occurred.
Step 3: Open original hash document and forged hash document and compare each hash word from both the files.
Step 4: If the hash is unequal this means at the particular location forgery has occurred**(Detection).**
Step 5: Note the index of unmatched hash and return the index**(Localisation).**
Step 6: Open the forged file, go to the identified index and return the word present at the specified index. The words returned are the added word.**(Retrieval).**
Step 7: Return the index and identified words.

Algorithm 5 detects the replaced words, in this case no extra words are added or deleted. The total no. of words are the same in original and forged document.

Algorithm 5: Detection, localization and retrieval in case of replacement.

Input: Original hash document, forged hash document and encrypted file.

Step 1: Compare total number of words in original hash document and forged hash document.
Step 2: Open original hash document and forged hash document and compare each hash word from both the files.
Step 4: If the hash is unequal this means at the particular location forgery has occurred**(Detection).**
Step 5: Note the index of unmatched hash and return the index**(Localisation).**
Step 6: Open the encrypted file and go to the identified index and decrypt the specific word using the same key as in algorithm 1.**(Retrieval).**
Step 7: Return the index and original word which was replaced.

V. RESULTS AND DISCUSSIONS

A. Deletion

In this case, the highlighted words in the original document which can be seen in Fig 5 have been deleted. This deletion of words has been shown in the forged document in Fig 6.

Video provides a powerful way to help you prove your point. When you click Online Video, you can paste in the embed code for the video you want to add. You can also type a keyword to search online for the video that best fits your document.

Fig. 5. Words before being deleted from the original file.

Video provides a way to help you prove your point. When click Online Video, you can paste in the embed code the video you want to add. You can also type a keyword to search online for the video that fits your document.

Fig. 6. Content of the forged document

In order to detect the forgery, Algorithm 3 is used wherein the original hash document is compared with the forged hash document to gain the altered index and fetching the encrypted word from the encrypted file at the same index which is then decrypted using the key to show the original word as shown in Fig 7.

```
C:\Users\Manmohan\PycharmProject
words in original document: 54
words in forged document: 50
Formatting attack occured
index: 4
original word: powerful
index: 14
original word: you
index: 26
original word: for
index: 47
original word: best
```

Fig. 7. Deleted words output

B. Addition

In this case, the original document can be seen in Fig 8 and new highlighted words have been added to it as shown in Fig 9 which is the forged document.

Video provides a powerful way to help you prove your point. When you click Online Video, you can paste in the embed code for the video you want to add. You can also type a keyword to search online for the video that best fits your document.

Fig. 8. Content of original document.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:36:55 UTC from IEEE Xplore. Restrictions apply.

Video provides a low powerful way to help you prove your point. When you will click Online Video, you can paste in the embed for code for the video you want to add. You can also type a keyword to search online for the video that best fits your document.

Fig. 9. Added words in the forged document.

As shown in Fig 10, the formatting attack was detected by comparing the original hash document with the forged hash document and the index value was obtained along with the added word as shown below.

```
C:\Users\Manmohan\PycharmProjects
words in original document: 54
words in forged document: 57
Formatting attack occured
index: 4
Added word: low
index: 16
Added word: will
index: 27
Added word: for
```

Fig. 10. Added words output

C. Replacement

If a word is replaced in the original document there will be the same number of words as shown in the fig 11. in which the highlighted word is replaced with the highlighted words in fig 12. The forged document is then parsed through algorithm 2 followed by algorithm 5 which identifies the altered word and fetch the encrypted word to decrypt and show the original word seen in fig 13.

Video provides a powerful way to help you prove your point. When you click Online Video, you can paste in the embed code for the video you want to add. You can also type a keyword to search online for the video that best fits your document.

Fig. 11. Content of original document

Video provides a weak way to help you prove your point. When you click Online Video, you cannot paste in the embed code for the video you want to add. You can also type a keyword to search offline for the video that best fits your document.

Fig. 12. Content of forged document.

```
C:\Users\Manmohan\PycharmPro
index:4
original word: powerful
index:20
original word: can
index:42
original word: online
```

Fig. 13. Replaced words output

VI. CONCLUSIONS AND FUTURE WORK

This paper proposes a fast and secure method to check and verify the authenticity of a Word Document. Several techniques have been developed in this field to protect and detect any changes made to the original document. Multiple Watermarking Techniques such as-"A Novel MS-Word Text Document Multiple Watermarking Technique for Authentication and Copyright Protection." proposed a robust and fast implementation of copyright, authenticity protection of work but was unable to localize "add, delete, replaced" alterations and hence unable to retrieve the original content but with this proposed method alteration were localized using hashing and original content retrieved using fernet encryption and decryption algorithm. This results in the highest possible accuracy and a fast system to detect and localize the alteration if any found further retrieve original content using encryption and decryption. In the future, the work can be extended to implement other type attack detection and retrieval algorithms in integrating all functionality under one cycle. Apart from that, work can be done on hiding and securing hashed and encryption data possibly with watermarks to secure data in the growing age of digitization on other types of data/documents, images, audio and other mediums so that it can be shared securely.

REFERENCES

[1] Milad Taleby Ahvanooey, Qianmu Li, Hiuk Jae Shim, Yanyan Huang, "A Comparative Analysis of Information Hiding Techniques for Copyright Protection of Text Documents", Security and Communication Networks, vol. 2018, Article ID 5325040, 22 pages, 2018.

[2] Julian Jang-Jaccard, Surya Nepal, A survey of emerging threats in cybersecurity, Journal of Computer and System Sciences, Volume 80, Issue 5, 2014, 973-993.

[3] Arora, Ayush & Agarwal, Drishti & Panda, Jeebananda. "A Novel MS-Word Text Document Multiple Watermarking Technique for Authentication and Copyright Protection" International Journal of Engineering Trends and Technology.(68)(2020). 1-6. 10.14445/22315381/IJETT-V68I5P201S.

[4] N. Nikolaidis and I. Pitas, "Copyright protection of images using robust digital signatures," 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Atlanta, GA, USA, 1996, pp. 2168-2171 vol. 4, doi: 10.1109/ICASSP.1996.545849.

[5] J. T. Brassil, S. Low and N. F. Maxemchuk, "Copyright protection for the electronic distribution of text documents," in Proceedings of the IEEE, vol. 87, no. 7, pp. 1181-1196, (July 1999), doi: 10.1109/5.771071.

[6] Sukhpreet Kaur and Geetanjali Babbar."A ZeroWatermarking algorithm on multiple occurrences of letters for text tampering detection". International Journal on Computer Science and Engineering.(2013).Vol. 5, No.05.

[7] Z. Jalil, A. M. Mirza and T. Iqbal, "A zero-watermarking algorithm for text documents based on structural components," 2010 International Conference on Information and Emerging Technologies, Karachi, Pakistan, 2010, pp. 1-5, doi: 10.1109/ICIET.2010.5625705.

[8] Suganya Ranganathan , Ahamed Johnsha Ali, Kathirvel.K & Mohan Kumar.M,".Combined Text Watermarking". International Journal of Computer Science and Information Technologies.(2010). Volume 1, No. 5.

[9] Z. ali and A. M. Mirza, "A Review of Digital Watermarking Techniques for Text Documents," 2009 International Conference on Information and Multimedia Technology, Jeju, Korea (South), 2009, pp. 230-234

[10] Omar Tayan, Yasser M. Alginahi and Muhammed N. Kabir."An Adaptive Zero-Watermarking Approach for Authentication and Protection of Sensitive Text Documents". International conference on advances in computer and information technology – ACIT.Vol.1.(2013)

[11] Afify, Amira Eid, Ahmed Emran, and Ahmed Yahya. "A Tamper proofing text watermarking shift algorithm for copyright protection." Arab Journal of Nuclear Sciences and Applications 52, no. 3 (2019): 126-133.

[12] Sahu, Aradhana & Ghosh, Samarendra. "Review Paper on Secure Hash Algorithm With Its Variants." International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES) (Volume 3, Issue 05) (2017)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:36:55 UTC from IEEE Xplore. Restrictions apply.

2021 Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 28-29, 2021

Drowsiness Detection using Light Framework

Amit Pratap Singh
Information Technology
Delhi Technological University
New Delhi, India
amitpratapsingh0000@gmail.com

Ritu Agarwal
Information Technology
Delhi Technological University
New Delhi, India
ritu.jeea@gmail.com

Chirag Gupta
Information Technology
Delhi Technological University
New Delhi, India
chirag7145@gmail.com

Himanshu
Information Technology
Delhi Technological University
New Delhi, India
himanshuchahar007@gmail.com

2021 Asian Conference on Innovation in Technology (ASIANCON) | 978-1-7281-8402-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ASIANCON51346.2021.9544929

***Abstract*—People must be aware that to be at their best, they have to be at least awake. However, if one operates a vehicle despite being tired or out of concentration, they are not at their best—moreover, they are endangering themselves. Human fatigue could, at times, become a severe problem that can hamper the protection of travelers. Nearly every fifth case in more than 200 significant investigations completed between year 2000 to 2010 claimed driver fatigue as a ground, decisive issue, or a finding as part of the investigation. In this research, we compare and propose a machine learning-based solution that can detect drowsiness in drivers efficiently and in real-time.**

Keywords—Drowsiness Detection, Random Forest, SVM, Cross-Validation, Scikit-learn

I. INTRODUCTION

For the past few years, road accidents are one of the major problems all around the world. And according to various surveys or reports, most of them are caused due to drowsiness or the tiredness of the driver, which results in some serious injuries or even worse it may result in death. In India according to a survey, more than 40% of accidents are caused due to driver's fatigue. Lack of sleep, continuous driving for long hours, monotonous driving pattern, drinking while driving all these can result in drowsiness state of the driver. Even drowsiness driving is also compared with drunk driving.

Many safety-related or intelligent driving systems have been proposed to decrease the risks of road accidents. Many companies like Audi, Volkswagen, Mercedes-Benz, Ford, etc. proposed certain methods [2] and many of them are still working on improving those methods.

Many researchers research around finding the pattern or the state to identify the proper state of the driver. So, detection of driver drowsiness is generally concerned with various behavioral measurements, vehicle-based measurements, or physiological measurements [1]. Behavioral measurements consist of driver's eye/face movements using a cam, vehicle-based measurements consist of various driving patterns such as driving speed, steering wheel movements, brake patterns, whereas physiological measurements consist of heart rate measurement, breathing rate pattern, electromyogram (EMG), electroencephalogram (EEG), electrocardiogram (ECG) and electrooculogram (EOG) all can be used to monitor drowsiness. Due to number of methods, it is generally difficult to say which one works better and there are pros and cons with every particular method.

This paper emphasize the use of behavioral or physical movement measurements for drowsiness detection. Our drowsiness detection app cleverly keeps track of the driver's fatigue by monitoring the eyes as well as any yawns. If the eyes of the driver are being closed for more than a few milliseconds (something called microsleep in sleep science), an alarm is sounded immediately to alert the driver. This element has been added because microsleep can be incredibly dangerous and is often responsible for a majority of the accidents happening due to long-haul rides. In addition, if the app sees the driver yawning, a warning message is displayed indicating the driver's fatigue. Our model expects the camera to be placed above the steering wheel on top of the speedometer, directly facing the driver. Our application is optimized for such an angle and this setup would be recommended.

II. RELATED WORK

Drowsy detection is a car safety algorithm that can alert a snoozing driver in hopes of preventing an accident. One of the methods is to build what is known as a deep drowsiness (DDD) network to learn useful features and detect drowsiness provided by the driver's RGB input video. Here one can use the latest CNN models to detect drowsiness and alert the drivers at the time of need but it has certain limitations such as i) lack of benchmark performance on the datasets, due to which other classification algorithms such as variants of CNNs and LRCN are used ii) DDD-FFA showed poor results as compared with DDA-IAA. iii) Accuracy is limited up to 73% only. iv) Limited datasets and more layers of neural networks can be added to achieve better results. Some of the systems that existed then were either bio-electrical signal processing based or video processing based to have driver's behavioral and physiological info. But the complexity in terms of algorithm and the required device for processing is way too high. Using phones can reduce the complexity of everything, i.e. getting video, storage, or any transmission. But because of the highly dynamic environment of driving mobiles can't afford high-end algorithms requiring both storage and computational processing; thus, the algorithm

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:38:35 UTC from IEEE Xplore. Restrictions apply.

needs improvement. Treating each information as mutually independent, SOBI with its high potential provided a balance of acceptable performance under low computational complexity. Usual signals like yawn can be detected from visuals. Still, for blinks, the best one is an electrocardiogram(Bio-electricity in eyes). Yet, then this requires using contact electrodes which make this inconvenient to be used in real driving scenarios.

Most of the existing algorithms with good accuracy were a kind of fusion were overlapping various methods one after another thus multiplying the complexity of computation. These people proposed a single method for combining and acquiring information more naturally into one algorithm, reducing the complexity and maintaining the accuracy.

ROI with help from the Meanshift tracker was used to have stable face detection. Then SOBI estimated the sources for various physiological activities. Then the algorithm distinguished yawn, BVP, and blink signals in terms of duration and frequency together. Then if anyone shows positive, it triggers the alarm.

1. But some of the shortcomings for this is that, while testing they got to know that if the rotation of face (or the camera) is beyond ±30.0 that can result in degradation of results because of drastic change in pixel values in ROI.
2. Similarly, in case the driver has small eyes, it becomes relatively difficult for the system to get the difference in the pixel values in case of blinks.
3. Some common items such as sunglasses when used while driving can lead this system to fail because of the non-availability of any visual clues as eyes are apparently invisible to the camera.
4. Cars going across a curve or a rough road in speed causes vibration or inertial movement that's too severe that the complete area of the face might not be present in the view of the camera. The structure of the overlapping (observation) data gets altered drastically. Signals like blink or yawn are out of the question.
5. The idea used here is very naive to merge multiple physiological information. For better robustness and adaptability to the real-life driving environment, more advanced methods could have been used.
6. Using some smart wearables (almost all are wireless) like bands and all we might be able to have much more accurate data being captured that could be used for more precise and reliable analysis.

Eye Detection is also proposed by various papers where they just monitor the state of the driver's eyes which can be either open or closed. Drowsiness is detected when the state of the eye remains closed for a longer duration than the normal eye blink time of 100-400 ms. The eyes of the driver are tracked using a camera that records the real-time video.

The location of eyes from the video is detected by facial landmark detectors already in use. The current landmark detectors are very precise in locating the eye including eye boundaries and contours. The state of the eye for each frame of the video is detected using an algorithm that monitors a single scalar quantity called the eye aspect ratio [3]. The eye aspect ratio is a single quantity representing the level of eye-opening(higher for an opened eye). This eye aspect ratio for each frame of the video is classified using an SVM classifier. This SVM classifier will be trained on sample datasets of blinking and non-blinking eyes. The eye aspect ratio significantly drops when a person closes his eyes(for blinking) and then raises quickly to the original level when the eye is reopened.

If the eye remains closed for a greater than expected time, then it is the red light for drowsiness detection. This method can be implemented in real-time as the computational time for landmark detection and testing on pre-trained SVM classifiers is negligible.

There are no major hardware requirements in this idea as a single camera is sufficient enough for tracking the state of the eyes.

One of the shortcomings of this idea is that there can be cases for false positives in classification. This idea is also not able to make use of facial features other than eye blink which can be useful for drowsiness detection.

We placed a video camera in the vehicle of 30 individual volunteers to record their activities and faces during driving. We used a device with 30 frames per second (fps) camera. The images from these video cameras were carefully chosen to create the dataset, where each image was marked as the state of the driver's eyes either being closed or not and the state of the driver's mouth either yawning or not.

III. Proposed Method

Our aim is to present a simple and efficient method to detect a driver's drowsy state. The contributions of our work can be briefly summarized as follows:

A. Data Collection

A setup for real time capturing of video from the camera mounted some where near the steering wheel.

B. Feature Extraction

Analyzing video frame by frame, landmark coordinates were extracted from each frame using DLIB. These landmark coordinates were then used as features for the dataset.

C. State Detection

States for mouth and eyes were then predicted using the supervised learning model from the features extracted from the last step.

D. Raise Alert

Using the trained model prediction, a visual trigger like LED can be triggered, which will inform the driver in case of yawning (open mouth).

Similarly, an audio alarm can be triggered if the driver goes into drowsy state (state when model predicts the eye closed, continuously for a predefined number of frames).

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:38:35 UTC from IEEE Xplore. Restrictions apply.

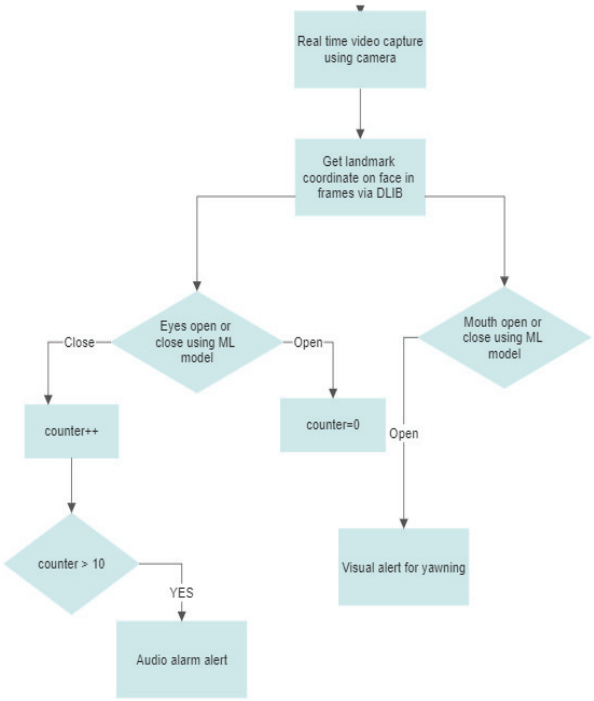


Fig. 1. Flow Chart of Model Proposed

IV. Feature Engineering

For each image, facial landmarks were detected using DLIB.

Landmarks using DLIB: Pre-built landmark systems are used which show pretty good results in the detection of landmark features and localizing facial expressions such as eye, ear & nose. We used a pre-trained facial landmark recognition algorithm present in a python library called dlib. The facial landmark detector is primarily based on linear SVM and histogram of oriented gradients (HOG). There is a pre-trained landmark detector present in the DLIB library, through which we can predict the location of 68 coordinates on the human face. These are 68 landmark coordinates, which can be mapped to human facial structure. The coordinates used in this paper are of left eye, right eye and mouth.

The six landmark coordinate points of each eye were used as features in training and testing the machine learning models for drowsiness detection.

Similarly, the 8 coordinate points of mouth were used as features in training and testing the machine learning models for yawn detection.

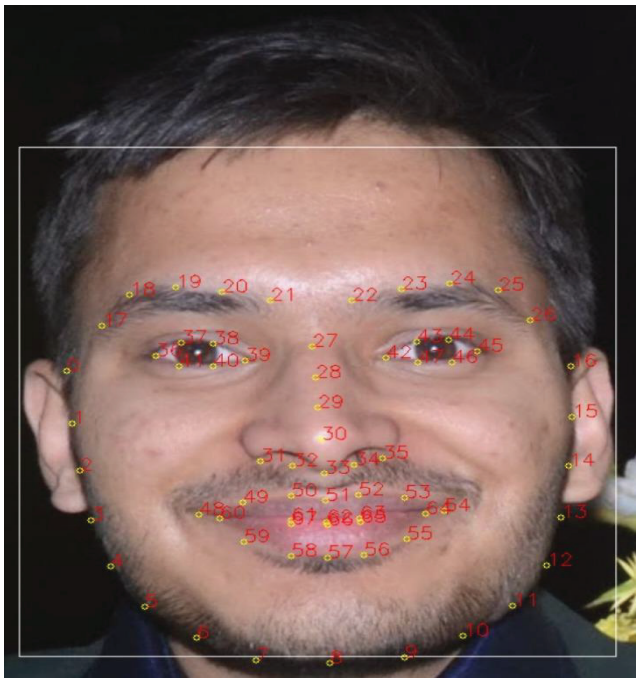


Fig. 2. Landmarks using DLIB

V. Model Selection

A. Linear Regression –

Line regression is the most basic and widely used form of prediction analysis. Linear Regression is a machine-aligned machine algorithm where the predicted result is continuous and has a constant slope. Instead of classifying our output into categories we divide them into the continuous range. It is generally divided into two broad categories linear regression and multivariable regression. Here we are using multivariable regression. These regression rates are used to build relationships between variables. To analyze the performance of the algorithm we can use a variety of error methods such as total error, mean square error.



Fig. 3. Learning curves for a Linear Regression model

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:38:35 UTC from IEEE Xplore. Restrictions apply.

B. Ridge

This model falls under the scikit read. It solves a retrospective model in which the loss function is a function of small squares. This rating has built-in support for various reversals. It has a divisive variant known as the RidgeClassifier. This separator first converts the binary target to one by one and then takes the function as a regression function, using the same purpose as above. The predicted category corresponds to the regressor predictive indicator.

It works in part because it does not require impartial measurements; While small squares produce an unbalanced balance, the variance can be so large that it may not exist at all. Ridge retreat can add enough bias to make a reliable and reasonable balance to true human values.



Fig. 4. Learning curves for a Ridge Regression model

C. SVM

We used 5 different kernels in SVM (Support Vector Machine) for identifying whether the person is sleeping or not. SVMs are a set of supervised machine learning models, which were used here for classifying the images into whether the person has eyelids closed or open, which further tells whether the person is sleeping or not. We used linear, nonlinear, polynomial, Gaussian, sigmoid, and Radial basis function (RBF) kernels. Out of these RBF performed the best for our dataset. In the given fig 3, we can clearly see how increasing the training set SVM with RBF kernel has significantly lower Mean Squared Error than other models, and hence it performs much better than other models we have seen so far. Also, the curve drops much sharper than in other models.



Fig. 5. Learning curves for a SVM (RBF) model

D. AdaBoostRegressor

It is a meta-algorithm, it can be used together with other algorithms for performance improvement. The concept of boosting is a type of linear regression. This model is also given the data of people with eyelids closed and open to predict if the person is sleeping or not. For our dataset AdaBoost performed rather poorly. Looking at fig 2, MSE started to decrease initially, but very soon the rate of fall decreased sharply, and when the training set size increased from 600 the error increased again. The model fit poorly, Changing different parameters did not improve the performance of this model, and hence this model is a bad choice for classification for our dataset.



Fig. 6. Learning curves for a Ada Boost Regressor model

As machine learning algorithms' substantial part is related to classification i.e. we try to find which class (i.e. group) data point comes from. How accurately and precisely we predict the class is what is rewarded the most in this kind of set of problems. Overlooking various other classification algorithms discussed above like logistic regression, SVM, or for that matter naive Bayes and decision trees, Among the top levels is Random Forest Classifier.

Decision Tree works on the basic idea of splitting observations based on features in some order resulting into groups that are as different from each other as possible.

E. Random Forest

Random Forest very much evident from the name, basically it, in turn, relies on some more significant number of decision trees which then works as a supporting player. Every tree alone suggests a class prediction, and then by the law of voting, the majority class is the proposed class from the forest. The basis of this idea working is more than one relatively uncorrelated model (trees) deciding as a committee will perform much better than any of the models working individually. Thus the remarkable effect is that the trees protect each other from their errors. Our Suggested model is also based on Random forest as it rigorously tested the data collected, and it outperformed almost all the classifiers.

It is very quick and is the most accurate. Also, as the data set's size increases, its performance is expected to scale

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:38:35 UTC from IEEE Xplore. Restrictions apply.

accordingly from the validation error graph shown below from our dataset.



Fig. 7. Learning curves for a Random Forest model

Comparison -

Using Cross-Validation and Learning curves for the various models that we tried for classification using the data points the most promising is Random Forest. As far as our small dataset of some 1000+ data points following Table gives the ACCURACY for the various models that did better than average.

TABLE I. ACHIEVED VALUES FOR DIFFERENT MODEL TESTED ON OUR DATASET

Model Used	Mean Abs. Error	Mean Sq. Error	Root Mean Sq. Error	R2 score
Linear Regression	0.125	0.026	0.16	0.896
Ridge	0.125	0.026	0.16	0.896
SVM(RBF)	0.048	0.0139	0.11	0.93
ADA boost Regressor	0.0957	0.0200	0.1417	0.919
Random Forest	0.0225	0.0089	0.0948	0.964

Of the models tested, two of them namely SVM with RBF Kernel and Random Forest shows a promising reduction in errors as the size of the data increases. Above we have also plotted the learning curves of the above-mentioned models for reference. And from these two as well We propose Random Forest to be used as the model for classifying the data points because of its nature of the curve and R2 score attained on our dataset.

VI. RESULT AND CONCLUSION

We were able to show that machine learning-based models, trained on eye landmarks can be successfully and efficiently used to detect drowsiness state in human eyes. Although much work is to be done in this field, we were able to show that using landmark features can be the way forward in tackling many road accidents happening due to drowsy driving. Once this idea is fully utilized, it will be a boon to humanity and can be a great help in saving millions of priceless lives lost on roads.

VII. FUTURE SCOPE OF WORK

In this paper, we used machine learning-based algorithms, which are random forest, Support Vector Machine, Lasso, Ridge, and Regression-based techniques to train and test the state of eyes and use it further to detect drowsiness. For analyzing 60 frames per second video for drowsiness detection, the speed of the algorithm in providing the output becomes crucial. This is the main reason for not using neural network-based algorithms. A fast and effective neural network algorithm can be a way forward, for this kind of project. Currently, our research is dependent on dlib python library to capture landmark points. If in the future, some other techniques arrive to capture data more efficiently, then it will be a boon for the researchers in this field. Also, having more data will definitely help in giving our model a more accurate standing.

The signals of drowsiness in this project are only the facial landmarks. In the real world, many factors can be indicators of drowsiness. These factors can be -

1. Steering Pattern Monitoring - The input from the electric power steering system can be useful in drowsiness detection as long as the vehicle is steered by the driver and is not using an automatic line monitoring system.
2. Physiological Measurement - Other body behavior measurements such as brain activity measurement, heart rate measurement, skin conductance measurement, and muscle activity measurement using respective sensors can be helpful in drowsiness detection.
3. Using some smart wireless gadgets like bands we might be able to capture things like heart rate or stress levels, whose patterns could definitely be must more robust in the detection of any kind of inactivity or micro-sleep.

If these factors can also be incorporated into the current model, then this will increase the robustness of the current research.

REFERENCES

[1] A. Sahayadhas, K. Sundaraj, and M. Murugappan, ‘‘Detecting driver drowsiness based on sensors: A review,’’ Sensors, vol. 12, no. 12, pp. 16937–16953, Dec. 2012. doi: 10.3390/s121216937

[2] D. Dawson, A. K. Searle, and J. L. Paterson, ‘‘Look before you (s)leep: Evaluating the use of fatigue detection technologies within a fatigue risk management system for the road transport industry,’’ Sleep Med. Rev., vol. 18, no. 2, pp. 141–152, Apr. 2014, doi: 10.1016/j.smrv.2013.03.003.

[3] Tereza Soukupova and Jan ´ Cech, "Real-Time Eye Blink Detection using Facial Landmarks" 21st Computer Vision Winter Workshop Luka Cehovin, Rok Mandeljc, Vitomir Struc (eds.)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:38:35 UTC from IEEE Xplore. Restrictions apply.

[4] Chao Zhang, Xiaopei Wu, Xi Zheng, & Shui Yu, "Driver Drowsiness Detection Using Multi-Channel Second Order Blind Identifications" in IEEE Access, vol. 7, 2019, doi: 10.1109/ACCESS.2019.2891971

[5] M. Hajinoroozi, Z. Mao, and Y. Huang, ''Prediction of driver's drowsy and alert states from EEG signals with deep learning,'' in Proc. CAMSAP, Cancún, Mexico, 2016, pp. 493–496.

[6] S.-Y. Shi, W.-Z. Tang, and Y.-Y. Wang, ''A review on fatigue driving detection,'' in Proc. ITA, Guangzhou, China, 2017, p. 01019.

[7] L. B. Leng, L. B. Giin, and W.-Y. Chung, ''Wearable driver drowsiness detection system based on biomedical and motion sensors,'' in Proc. IEEE SENSORS, Busan, South Korea, Nov. 2015, pp. 1–4.

[8] W.-C. Li, W.-L. Ou, C.-P. Fan, and C.-H. Huang, ''Near-infrared-ray and side-view video based drowsy driver detection system: Whether or not wearing glasses,'' in Proc. APCCAS, Jeju, South Korea, 2017, pp. 429–432.

[9] K. Churiwala, R. Lopes, A. Shah, and N. Shah, ''Drowsiness detection based on eye movement, yawn detection and head rotation,'' Int. J. Appl. Inf. Syst., vol. 2, no. 6, pp. 0868–2249, May 2012.

[10] V. B. Hemadri and U. P. Kulkarni, ''Detection of drowsiness using fusion of yawning and eyelid movements,'' in Advances in Computing, Communication, and Control, vol. 361. Berlin, Germany: Springer, 2013, pp. 583–594.

[11] S. Samiee, S. Azadi, A. Nahvi, A. Eichberger, and R. Kazemi, ''Data fusion to develop a driver drowsiness detection system with robustness to signal loss,'' Sensors, vol. 14, no. 9, pp. 17832–17847, Sep. 2014, doi: 10.3390/s140917832.

[12] A. Noda et al., ''Simultaneous measurement of heart rate variability and blinking duration to predict sleep onset and drowsiness in drivers,'' J. Sleep Disorders Therapy, vol. 38, no. 1, pp. 12–16, 2015, doi: 10.4172/2167- 0277.1000213.

[13] A. Tsuchida, M. S. Bhuiyan, and K. Oguri, ''Estimation of drowsiness level based on eyelid closure and heart rate variability,'' in Proc. Int. Conf. IEEE Eng. Med. Biol. Soc., Minneapolis, MN, USA, Sep. 2009, pp. 2543–2546.

[14] N. Alioua, A. Amine, and M. Rziza, ''Driver's fatigue detection based on yawning extraction,'' Int. J. Veh. Technol., vol. 2014, no. 1, Aug. 2014, At. no. 678786, doi: 10.1155/2014/678786.

[15] S. Abtahi, B. Hariri, and S. Shirmohammadi, ''Driver drowsiness monitoring based on yawning detection,'' in Proc. I2MTC, Binjiang, China, 2011, pp. 1–4.

[16] D. Comaniciu, V. Ramesh, and P. Meer, ''Real-time tracking of non-rigid objects using mean shift,'' in Proc. CVPR, Hilton Head Island, SC, USA, Jun. 2000, pp. 142–149.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:38:35 UTC from IEEE Xplore. Restrictions apply.

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-021-11440-8

EDGly: detection of influential nodes using game theory

Minni Jain[1] · Aman Jaswani[1] · Ankita Mehra[1] · Laqshay Mudgal[1]

Received: 24 May 2020 / Revised: 9 February 2021 / Accepted: 17 August 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract
Identifying those nodes that play a critical role within a network is of great importance. Many applications such as gossip spreading, disease spreading, news dispersion, identifying prominent individuals in a social network, etc. may take advantage of this knowledge in a complex network. The basic concept is generally to identify the nodes with the highest criticality in a network. As a result, the centrality principle has been studied extensively and in great detail, focusing on creating a consistent and accurate location of nodes within a network in terms of their importance. Both single centrality measures and group centrality measures, although, have their certain drawbacks. Other solutions to this problem include the game-theoretic Shapley Value (SV) calculations measuring the effect of a collection of nodes in complex networks via dynamic network data propagation process. Our novel proposed algorithm aims to find the most significant communities in a graph with community structure and then employs the SV-based games to find the most influential node from each community. A Susceptible-Infected-Recovered (SIR) model has been employed to distinctly determine each powerful node's capacity to spread. The results of the SIR simulation have also been used to show the contrast between the spreading capacity of nodes found through our proposed algorithm and that of nodes found using SV-algorithm and centrality measures alone.

Keywords Centrality · Game-theory · Shapely Value (SV) · Susceptible-Infected-Recovered (SIR) model · EdMot · Ensemble Clustering for Graph (ECG)

✉ Minni Jain
minnijain@dtu.ac.in

Aman Jaswani
ajaswani11@gmail.com

Ankita Mehra
mehra.ankita99@gmail.com

Laqshay Mudgal
laqshay.mudgal@gmail.com

[1] Delhi Technological University, Delhi, India

Published online: 09 October 2021

1 Introduction

The identification of influential nodes in complex networks has attracted considerable attention within the research community [36]. Social networking has become a modern tool for people to connect and spread the news with the development of science and technology. Thus, the critical issue of detecting a collection of the most prominent individuals in a network can help monitor the dissemination of social networking messages and rumours, discovering social influencers, etc. Particularly in the case of rumour propagation, as rumours spread quickly and widely and they have great power of destruction. It is therefore of great theoretical and practical importance to decide whether there is an influential spreader and to recognize who is the influential spreader in the process of spreading or controlling the spread of information [10, 14, 20].

Various approaches to the identification of influential nodes in complex networks including single centrality and group centrality measures have been explored in the past [21, 23]. A brief understanding of these concepts is discussed in Sect. 4.2. Other than the basic centrality measures, SV-based centrality algorithms have also been employed for similar problems [1, 22]. This is based on the main concept of a game-theoretic network which means defining a cooperative game across a network where agents are nodes, coalitions are node groups, and coalition payoffs are specified to meet the requirements of a given application.

Although many methods have been proposed, most of them fail to take into account the community structure of graphs while identifying influential nodes. Community structure organizes the nodes into sets such that a dense internal connection exists among the nodes in each set. In large and complex real-world networks, it is rarely the case that a single node could be able to dissipate a piece of information to the entire network. Therefore, we aim at identifying a group of nodes that might be able to scatter information to the rest of the network. Community structures refer to a subset of nodes in which all nodes are densely connected to all the other nodes present within the same community, but loosely connected to all the nodes present in other communities [27]. Higher-order community structure deals with small dense subgraph patterns and also the density of connections between two different communities. The benefit of identifying these is that we are able to find such a set of communities (rather than an individual nodes or just one community), that can dissipate information to majority of the network.

Various approaches have been used in the past for leveraging higer-order community structure such as Harmonic Motif Modularity [12], HBNmining [15], Attribute Homogeneous Motif [16], etc. The major drawback with all of them, however, is that they do not optimally deal with the lower-order topological structure. They do not take into consideration the smaller, isolated, and fragmented components that may be present in a large real-world network, and focus only on the larger connected components. EdMot [17] approach overcomes this by rewiring the network and by taking into account the higher-order community structure instead of only the lower-order one.

Graph clustering aims at partitioning the vertices of a graph into good clusters (based on different objective functions). Yang et al. [34] have studied and compared eight state-of-the-art clustering algorithms. They have also metioned the various merits and demerits of each of the approaches. Poulin et al. [26] proposes the ECG algorithm which outperforms all the algorithms studied by Yang et al. [34]. The ECG algorithm has been used to detect communities in a graph and is based on Multi-level Louvain algorithm [5]. Furthermore, the ECG algorithm also provides the information about strength between different communities present.

Our proposed algorithm initially aims to employ the EdMot approach that leverages the higher-order characteristics at the level of dense sub-graph patterns to identify the community structure in a graph. We propose an enhancement in this approach by using the Ensemble Clustering for Graph (ECG) algorithm for network partitioning, thus the rewired network is then fed to the ECG algorithm. After this, the most influential node from each partitioned community is identified according to the Shapely centrality value using EDGly algorithm. The ECG algorithm is based on the Louvain algorithm but produces more accurate results as a consequence of the ensembles of clusters used. Thus, we obtain the most significant partitioned communities from the original network and then employ SV based games to find the most significant nodes from each significant community identified. This gives us a list of all the influential nodes in the graph.

We have used the measures of prime centrality, the game-theoretical approach and our proposed algorithm to identify the most prominent top-k nodes and demonstrated a distinct and detailed contrast between all the approaches. We have finally used the SIR model to evaluate the performance of the proposed algorithm in comparison to the centrality-based measures and the SV games. The SIR model simulates an information spreading process in which our most significant nodes, identified via any of the algorithms, are considered as infected and tries to spread the information in the entire network. This helped to portray the different aspects and accuracy of the varied approaches vividly. More details on the working of SIR model and the evaluation criteria are mentioned in Section 7.3.

The main contribution of our work is to detect influential nodes in a graph by incorporating the community structure. EDGly algorithm uses and leverages the EdMot approach—an Edge enhancement approach for Motif-aware community detection, and ECG algorithm—a graph clustering method. The SV of each node is then calculated and the most influential nodes are obtained.

The remainder of this paper is organized as follows. Section 2 gives a detailed study of the various works done in the related field. Section 3 explains the datasets used. Section 4.1 explains the preliminaries while Sect. 6 explains the algorithmic flow. Section 7.1 illustrates the working of the proposed algorithm with the help of a diagram. Section 7.3 describes the evaluations performed and the results obtained. Section 9. concludes the research work with detailed insight into its future scope.

2 Literature review

One of the fundamental research discussions in the literature on network analysis is the topic of connectivity. The first to experiment to detect the primary top-k nodes were Domingos and Richardson [8]. They developed an algorithmic model to address this problem by modelling social media network as Markov random fields which mathematically characterized the probability of occurrence of an event. Chen and Teng [7] explained that single node centrality measures are suitable for assessing individual influence in isolation while Shapley centrality assesses individuals' performance in group influence settings.

Wei et al. [32] explored the need to learn distributed representation of the vectors for each vertex in a network. They laid emphasis on node classification and link prediction. A unique approach to identify influential nodes is to consider the k-shell value of each vertex as the mass and shortest path distance between two vertices as the distance, and then apply the concept of gravity centrality index [20].

We further investigated various kinds of centrality measures used for finding the most influential nodes in a network. Degree Centrality (DC), discussed by Gao et al. [10], is used to efficiently measure the significance of nodes. However, it does not take into consideration the overall, detailed anatomy of the network. Eigen-Vector Centrality (EC), according to Stephenson and Zelen [29], overcomes the defects associated with DC. It takes into account the influence of neighbours of the node in consideration. Betweenness Centrality (BC), as explored by Freeman [9], learns topology-related data of networks in advance. Al-Garadi et al. [2] describes how Closeness Centrality (CC) can be efficiently used to identify more than one prominent spreader.

Previous research by Tan et al. [30] on spreading rumours focused primarily on communities' viral epidemics. The normal (and somewhat standard) model for viral epidemics is called the restored or SIR model that is susceptible-infected- recovered. Various methods have been defined for the same, including weighted k-core decomposition method [2] and rumour centrality with a mass centre technique [30]. Algorithms with low complexity have been devised that can be applied to large scale networks by introducing entropy into a complex network and using the SIR model for evaluation. Such a model [37] can be used for rumour management and targeted advertisements.

We studied the disadvantages associated with using centrality algorithms to find the most significant node in [7, 19, 24]. An attempt has been made to find the most influential node in a network using Mapping Entropy (ME) that reflects the correlation between a node and its neighbours [24]. We particularly inspected the application of ME using ENRON email dataset which is commonly used for the study of social networks [28].

Game-theory is a significant paradigm that finds its applications in various fields. It is used in statistics and business analytics for prototyping the interactivity among participating agents [4]. One of its most prominent applications is finding the most influential node within a network. Typical social network analysis cannot capture the dynamics of strategic interactions among the individuals in the network. Our proposed model is based on cooperative game theory that solves this issue [33]. The elemental constituents of intricate interactivities in a network can be efficiently processed using a rich class of games, called influence games, as has been demonstrated in [13]. Game-theoretic network centrality measures along with their computational complexity and axiomatic properties have been neatly explained in [31].

We studied about Explosion-Trust (ET) Game Model by referring to [18]. It remarkably explains how a rumour spreading model can be constructed using game-theory by considering two very significant factors—rumour explosion degree and trust degree of the source node. In complex networks, dense clusters [14] can be formed by detecting influential nodes using degree centrality and selecting the most influential nodes as the initial spreaders. This ensures that not just the spreader nodes are prominent themselves, but also that the distance among the nodes is relatively scattered. In [36], the probability of transfer of information was used between any pair of nodes and the clustering method, called k-medoid. Such an efficient algorithm helps to classify large scale powerful nodes in complex networks with community structures.

Amati et al. [3] aimed to find the most influential nodes from various Twitter scraped datasets. Primitive centrality measures like degree centrality, closeness centrality, eigenvector centrality, and page-rank centrality were employed for the same. They also studied how the influence of spreaders evolved over time with the help of Dynamic Retweet Graph (DRG) that focused on how retweets of a tweet were made over time. This helped them better analyze the relationship between users. They concluded that closeness centrality gave

too many nodes with high centrality and was, therefore, not a good indicator. Furthermore, the remaining measures gave almost similar results in terms of the most influential nodes.

3 Data sets

This section gives an elaborate description of the datasets that have been used for our implementation. We have used two datasets which have been described below.

The first dataset is the USAir97 dataset [6] which has been transformed into an undirected network, created by 332 nodes, where one airport represents a node, and 2126 edges, with each edge reflecting a direct air-line between two American airports, if any. Here, weights represent the normalized distance between two airports.
The second dataset is a tiny Facebook network [25] made up of 899 users serving as nodes and having 7089 edges. An edge occurs between two vertices if a few messages have been exchanged between the two users and are weighted by the number of messages exchanged.

3.1 Unweighted graph

An unweighted graph can be technically defined as a graph G(N, E) having 'n' nodes represented by set N and 'e' edges represented by set E consisting of unordered pairs, such that $(n_1, n_2) = (n_2, n_1)$ and $(n_1, n_2) \in E$ and $n_1, n_2 \in N$. Games 1 and 2 are played by creating an unweighted network from both the datasets.

3.2 Weighted graph

A weighted graph can be technically defined as a graph G(N, E) having 'n' nodes represented by set N and 'e' edges represented by set E consisting of ordered pairs, such that $(n_1, n_2) \neq (n_2, n_1)$ and $(n_1, n_2) \in E$ and $n_1, n_2 \in N$. Games 3–5 are played by creating a weighted network from both the datasets.

4 Preliminaries

Section 4.2 explains the centrality measures used while Sect. 4.3 describes the process used for community detection in detail.

4.1 Centrality measures

We have used certain centrality measures in our implementation to compare the results our proposed algorithm with these already existing measures. For more insight into these centrality measures along with their mathematical derivations, refer to the work of Zhan et al. [35]. The centrality measures used in our study are as follows:

(i) Eigen-vector Centrality (EC): EC [29] considers the relative power or significance of the nodes. Here, each node is assigned a value representing its relative significance considering the fact that nodes which are connected to high-power nodes have high influence over the network in comparison to those who are connected to low-power nodes.

$$C_e(v_i) = \frac{1}{\lambda} \sum_{j \in M(v_i)} C_e(j) \tag{1}$$

In Eq. (1), $M(v_i)$ is set of neighbours of v_i, and λ is a constant. $C_e(v_i)$ is the measure of EC for vertex v_i.

(ii) Betweenness Centrality (BC): BC [9] measures how strongly two nodes are connected via a given node. It is estimated as the ratio of the aggregate of shortest distances between any two nodes in the network, on which the node lies, to the shortest path between the two nodes considered.

$$C_b(v_i) = \sum_{j=1, k \neq 1}^{N} \frac{g_{jik}}{g_{jk}} \tag{2}$$

In Eq. (2), g_{jk} is the total number of shortest paths from v_j to v_k, and g_{jik} is the number of paths that pass through v_i. $C_b(v_i)$ is the measure of betweenness centrality for vertex v_i.

(iii) Closeness Centrality (CC): CC [2] measures how quickly information can be spread from one node to all the other nodes in a network. It is measured as the inverse of the total sum of all shortest path distances between a given node to all other nodes in a network.

$$C_c(v_i) = \sum_{j=1}^{N} \frac{1}{d(v_i, v_j)} \tag{3}$$

In Eq. (3), $d(v_i, v_j)$ is the distance between vertices $v_i$andv_j, and $C_c(v_i)$ is the measure of closeness centrality for vertex v_i.

Single centrality measures suffer from certain disadvantages due to the failure to recognize the effects when considered in groups on node functionality. Group centrality measures place great focus on operating in groups of nodes and not on their individual functionalities. Nonetheless, group centrality also suffers from a drawback by focusing on a- priori-determined node groups and contributing to confusion when prioritizing individual nodes within the network.

4.2 Community detection

The algorithm divides the task of detecting influential nodes into two parts:

(i) Detecting significant communities in a graph

We used the ECG algorithm to detect the communities present in the graph, primarily because of its efficiency in accuracy and the number of communities found. The latter metric enables us to experiment with the number of communities and see the effect of that on the infected scale in our SIR Model. The ECG algorithm works

with consensus clustering. It has two main steps—generation and integration. The generation process uses the Multilevel Louvain algorithm [5]. The ML algorithm creates a hierarchy of partitions, and ECG uses the randomized partitions at level-1. The number of the randomised partitions used determines the size of the ensemble of clusters. The integration step involves applying ML on the weighted version of a graph, G, where the weight of each edge is given by:

$$W_P(u,v) = \begin{cases} w_* + (1-w_*).\left(\frac{\sum_{i=1}^{k} v_{P_i}(u,v)}{k}\right), \text{if } (u,v) \text{is in the 2-core of } G \\ w_*, \text{otherwise} \end{cases} \tag{4}$$

In Eq. (4), $0 < w_* < 1$ is the minimum ECG weight and $v_{P_i}(u,v) = \sum_{j=1}^{l_i} 1_{C_i^j}(u).1_{C_i^j}(v)$ indicates if the vertices u and v co-occur in the same cluster of P_i or not. The k-core of a graph is its maximal subgraph such that every vertex has degree of atleast k. In this case, the 2—core of G is its maximal subgraph such that each vertex has degree of atleast 2.

For this, we proposed an enhanced version of the EdMot algorithm by replacing the Louvain algorithm for network partitioning by the ECG algorithm. The EdMot approach for network rewiring has been used to increase connectivity within connected components in the graph. This approach takes into account the higher-order community structure instead of only lower-order structures. The algorithm aims to unravel peculiar characteristics at the level of network motifs. These are basic interaction patterns in a graph that recur throughout graph networks, much more than in random networks. This is crucial to inculcate while considering the influence of nodes in a graph.

(ii) Finding the most influential node from each community

For finding the most influential node per community in a graph, we use *The Shapley Value*. On the basis of the concept of marginal (or borderline) contribution, an important solution concept was proposed by Shapely [1, 22].

Player $i^{'}s$ SV, denoted by $SV_i(v)$ is equal to the weighted mean of $i^{'}s$ borderline contributions to each coalition C, to which the player may belong.

$$SV_i(v) = \frac{1}{n!} \sum_{\pi \in \pi(I)} \{v\left(C_\pi(i) \bigcup \{i\}\right) - v(C_\pi(i))\} \tag{5}$$

In Eq. (5), the aggregate count of players is given by 'n' while $\pi(I)$ gives the set of all permutations with 'n' players. More insight into the SV and its derivation can be found in [1, 22].

This idea is based on the theory of cooperative games—an element of game theory that allows players to form coalitions to maximize their game yields. Cooperative games theory provides a high-level approach as it describes only the coalitions' structure, strategies and benefits.

4.3 Coalition games

Coalitions are gatherings of players that form the essential or fundamental elements of decision making. These are assumed to uphold cooperative conduct which makes it reasonable to view these games as a contest between alliances of participants and not between separate players. The core assumption here is that as the game proceeds, an eminent alliance or coalition comprising all participants will manifest eventually.

5 Proposed EDGly algorithm for influential nodes detection

Focusing on Game Theory's Shapley algorithm, we referred to the algorithms described in Michalak and Szczepański's work [1, 22]. In both weighted and unweighted networks, the exact analytical formulae for SV-based centrality were established. The SV-based centrality polynomial-time algorithms have been developed.

Step 1: Creation of weighted and un-weighted network graphs — Graphs are created by using the networkx library in Python for both USAir97 and Facebook datasets. Games 1 and 2 require unweighted graphs whereas the remaining games require weighted graphs. Please refer Sect. 3.

Step 2: Detection of communities — Connectivity within the network is enhanced by network rewiring as explained in Algorithm 1. After this, communities are detected using the ECG algorithm for the detection of network communities. In line with EDGly algorithm, from each community, we picked the most influential node (as described in Step 3). It was observed that the number of communities identified does not change results significantly. We have shown the seven communities and their corresponding influential nodes.

Algorithm 1: The **EDGly** approach

Input: Original network $A \in R^{nXn}$, parameter K
Output: The top influential nodes from A. (One from each community detected.)

1. Construct motif adjacency matrix W_M from A via (2).
2. Obtain connected components Φ via (5) and select top K largest connected components.
3. Apply ECG to partition each $\phi_l \in \Phi_K, \forall l \in \{1, ..., K\}$ and aggregate the results to form the module set $\{M_1, \ldots, M_{\overline{m}}\}$.
4. Construct a new edge set E^*_{mod} via (8).
5. Rewire the original network to obtain G_A^M via (9).
6. Feed G_A^M into ECG to obtain final community structure $\hat{C}$
7. Iterate over every community and detect the most influential node using Shapely Centrality Value.

Step 3: Coalition games based on Shapely algorithm — SV is the average marginal cost contribution across all potential coalitions of a function value.

Step 3: Coalition games based on Shapely algorithm	The Shapley algorithm is applied carefully and tries to find the top-k nodes that might be the most prominent ones. Specifically, we concentrated on five underlying network-defined coalition games that vary in degree and centrality of the network. Each game has a certain characteristic function v(C) which represents how prominent a particular node is to a given coalition C. For more insight into the working of these games and their underlying mathematics, refer to [1, 22]. The game descriptions are as follows:

(i) Game 1: In this game, we consider all the permutations of all the nodes that are immediately reachable, by one hop to the node $n_i \in$ N(G). Let each random permutation be denoted by P_i, the neighbours of node n_i, in the graph G(N, E) be denoted by n_i.*neighbours* and the degree of node n_i, be denoted by n_i.*degree*. Algorithm 2 describes the procedure involved in SV calculation.

(ii) Game 2: In many real-life social scenarios, often taking into account nodes that are directly attached to each other is not enough. A rumour source, will more often than not, affects farther nodes.

For the purpose of taking relationships with farther nodes into account, and generalising the game, we introduce a value, p, depicting the number of agents that the node is adjacent to in a coalition. In this game, a node is considered 'influenced' if at least p of its neighbours are influenced. We divide the analysis using this game into 2 parts, first, where the degree of the node is less than p and second where the degree is more than p. Algorithm 3 describes the procedure involved in SV calculation.

(iii) Game 3. In this game, we introduce the concept of weighted graph networks. This game is an extension of game 1; it uses the Dijkstra Algorithm to compute the distance between 2 nodes. The cutoff value, d, is the maximum permissible distance of a node from any member in a given coalition. The extended degree is defined as the size of the set of all nodes that are at-most distance 'd' away from the node n_i. Algorithm 4 describes the procedure involved for SV calculation.

(iv) Game 4. This is a generalization of game 3. Here we work with the assumption that a node closer to a coalition will have a greater effect on it than some other node farther away, even if both nodes satisfy the cut-off criteria as in game 3. For this purpose, we introduce a positive-valued decreasing function f(x). f(d) refers to the function which has a directly proportional effect on SV of the coalition which is 'd' units away from a node. The marginal contribution of each node n_i through node $n_i \neq n_j$, for each coalition C_i gives SV, as shown in Algorithm 5.

(v) Game 5. This is a generalization of game 2 in case of weighted networks. Here we have defined a cut-off value (n_i) for each $n_i \in$ N(G). For every coalition C, d(n_j,

C) = $\sum_{nj \in ni.\text{neighbours}} W(n_i, n_j)$, where $W(n_i, n_j)$ represents the edge's weight between nodes n_i *and* n_j (0, if no edge exists).

A node n_i marginally contributes node $n_j \in n_i$. neighbours to the value of coalition C_i if and only if $n_j \notin C_i$ and $d(n_j)$ - $W(n_i, n_j) \leq W(C_i, n_j) < d(n_j, C)$. Algorithm 6 describes the procedure for calculating the SVs.

Step 4: Estimating centrality measures — After working on the five coalition games, we introduced multiple centrality measures to determine the network's most powerful node with the highest scope or effect. To generate an elaborate comparison, various network centrality measures such as EC, BC, CC, have been used.

6 Illustration with example

A synthetic network is constructed to illustrate the working of our proposed EDGly algorithm, where nodes in communities are denoted with different colours and the edges represent the connection between communities.

A hypergraph, is formed in the first step as shown in Fig. 1. After partitioning the largest connected component into modules in Step 2, two modules can be identified and a new edge set is constructed to derive a clique from each module, A clique is a graph that has an edge connecting every pair of vertices—the new edges are shown as the dashed lines. A rewired network is obtained by combining the current edge set into the original network by reorienting the network in Step 3. Finally, ECG algorithm is used to partition the rewired network into communities.

After the community structure of the graph is obtained via ECG algorithm in Step 4, the game-theoretic Shapely centrality value is used to detect the most influential node from each community. Accumulation of these nodes from all communities in the graph gives the set of most influential nodes of the network.

Table 1 depicts the results of using EDGly on the synthetic dataset. Three communities are detected and further influential nodes are obtained by applying the Shapely centrality using Games 1–5. The three influential nodes obtained from game 1 and 4 are nodes 1, 9 and 12. Similarly, for game 2, the nodes are 3, 5 and 11. For game 3, the nodes are 1, 6 and 12 while for game 5, the nodes are 3, 9 and 11. We get the set of influential nodes by accumulating the most influential node from each community.

7 Evaluations and results

Section 7.2 explains SIR Model [37] used to compare our model with the prominent existing centrality algorithms to identify the network's most powerful node. To evaluate and compare the performance of EDGly, we have used various benchmark network centrality measures such as EC, BC, CC; these centrality measures describe the influence of nodes on the basis of the structure of the network, and are helpful to

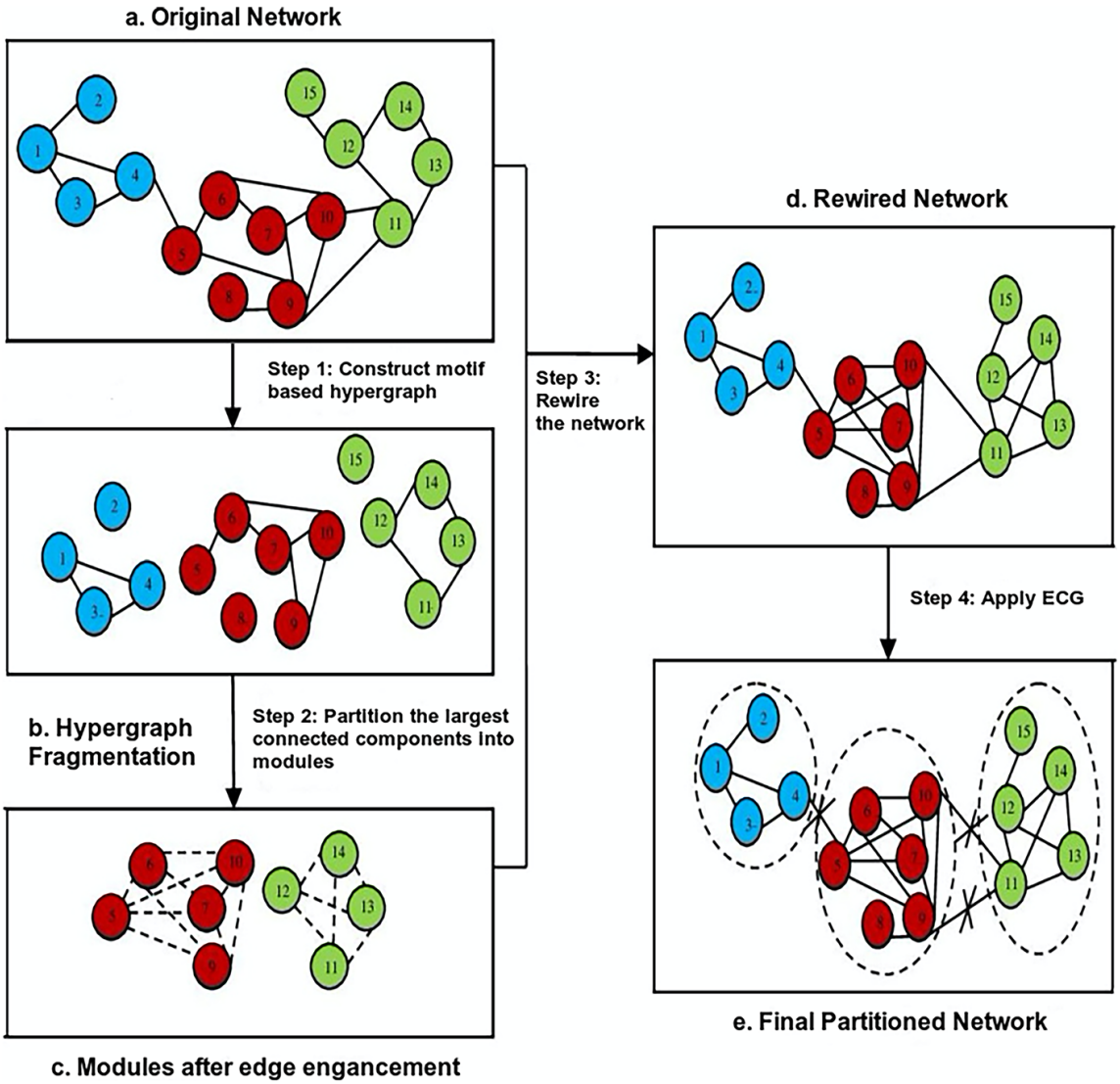


Fig. 1 Algorithmic flow for EDGly

depict an accurate effect of the impact of community structure on the influence of the nodes in a graph. We have further used the SIR model. We have compared the algorithms on the basis of their spreading abilities that can be measured by final infected scale $F(t_c)$ value, that is obtained from SIR simulation and widely used to measure the spreading ability of nodes.

To generate an elaborate comparison, Sect. 7.3 explains the evaluations performed. Sect. 7.1 shows the detailed comparison of results of our proposed EDGly algorithm with those of the Shapely centrality measure and of Shapely centrality with EdMot-Louvain for community detection.

Table 1 Result of EDGly on synthetic dataset

Community	Games				
	G1	G2	G3	G4	G5
1	1	3	1	1	3
2	9	5	6	9	9
3	12	11	12	12	11

7.1 SIR model

According to this model, the dynamics of rumour spreading can be studied. There are three types of nodes in a typical rumour propagation model: (i) susceptible nodes capable of infection, (ii) infected nodes capable of further spreading the rumour, and iii) recovered nodes that are healed and no longer capable of infection. The most influential nodes identified by any of the algorithms are considered to be the initial sources of rumour propagation. These nodes are marked as infected and all the remaining nodes are marked as susceptible.

We used the USAIR97 dataset [6] to find the most significant airports in an air-travel network. We assumed the initially infected nodes to be most influential airport per graph community. As each edge holds the attribute for the number of flights from one airport to another, we were able to determine the coverage of all locations via the initial influential airport at the end of the simulation.

In the Facebook dataset [25], we aimed to depict a simulation of rumour spreading. The most influential nodes identified by any of the algorithms were considered to be the initial sources of rumour propagation. These nodes were marked as infected and all the remaining nodes were marked as susceptible. At the end of the simulation, we were able to understand the influence of each person predicted by our algorithm. A similar understanding can be used for targeted advertising. The versatility of the model can be gauged by analyzing results on datasets from various domains.

Each iteration is considered to be one time-step, where each of the infected nodes can spread the rumour to its susceptible neighbours with a probability β. Also, each currently infected node becomes recovered with a probability γ, i.e., this node will not spread rumours in the further iterations.

7.2 Evaluations

For the purpose of our experiments, we set $\beta = 0.8$ and $\gamma = 0.2$. The iterations go on till there are no more new infected nodes. The term $R_0 = \beta/\gamma$ is known as the reproduction rate and is extremely crucial to the spreading process. If $R_0 > 1$, then rumour will spread otherwise it will die out soon. Spreading potential is assessed at each time-step 't' by the infected $F(t)$ scale [11]. The value of $F(t)$ is obtained by using the formula given by Eq. (6).

$$F(t) = \frac{n_{i(t)} + n_{r(t)}}{n} \tag{6}$$

Here, $n_{i(t)}$ is the number of infected nodes at time t, $n_{r(t)}$ is the number of recovered nodes at time t, and n is the total number of nodes. Larger the value of $F(t)$, more the influencing capacity of our initial infected set.

After a few iterations, when the rumour spread starts to slow down, we reach a point of saturation where no more new nodes get infected. This is considered to be a stable state and is represented by $F(t_c)$, as shown in Eq. (7).

$$F\left(t_c\right) = \frac{n_{r(t_c)}}{n} \tag{7}$$

Table 2 Comparison of infected scale at stable state in various benchmark algorithms for USAIR97 dataset

Method	Initial spreading nodes							$F(t_c)$
	1	2	3	4	5	6	7	
GAME 1								
EC	26	187	182	169	322	147	258	0.421
BC	144	116	46	67	183	112	219	0.370
CC	110	47	197	67	104	322	248	0.337
SHAPELY	182	255	230	201	8	166	67	0.319
SHAPELY +EdMot	8	65	182	118	201	258	261	0.352
EDGly	195	71	100	152	261	118	35	0.472
GAME 2								
EC	26	187	182	169	322	147	258	0.334
BC	144	116	46	67	183	112	219	0.391
CC	110	47	197	67	104	322	248	0.397
SHAPELY	166	152	182	230	7	147	112	0.346
SHAPELY +EdMot	4	74	140	93	312	302	272	0.415
EDGly	195	130	100	215	119	145	35	0.439
GAME 3								
EC	26	187	182	169	322	147	258	0.373
BC	144	116	46	67	183	112	219	0.352
CC	110	47	197	67	104	322	248	0.312
SHAPELY	201	248	313	67	144	182	261	0.412
SHAPELY +EdMot	8	59	182	118	313	256	230	0.418
EDGly	195	130	100	256	67	118	35	0.454
GAME 4								
EC	26	187	182	169	322	147	258	0.388
BC	144	116	46	67	183	112	219	0.388
CC	110	47	197	67	104	322	248	0.406
SHAPELY	152	201	255	230	8	67	166	0.382
SHAPELY +EdMot	8	65	182	118	201	258	261	0.313
EDGly	195	71	100	152	261	118	139	0.412
GAME 5								
EC	26	187	182	169	322	147	258	0.445
BC	144	116	46	67	183	112	219	0.430
CC	110	47	197	67	104	322	248	0.382
SHAPELY	248	255	47	293	67	166	233	0.313
SHAPELY +EdMot	8	65	67	112	313	109	221	0.346
EDGly	195	71	100	248	261	118	35	0.490

Table 3 Comparison of infected scale at stable state in various benchmark algorithms for Facebook dataset

Method	Initial spreading nodes							$F(t_c)$
	1	2	3	4	5	6	7	
GAME 1								
EC	13	68	59	870	101	186	443	0.634
BC	68	8	54	150	650	173	175	0.636
CC	75	570	150	60	650	29	176	0.620
SHAPELY	75	194	82	12	67	208	8	0.638
SHAPELY	8	59	298	121	164	10	212	0.635
+EdMot								
EDGly	164	7	19	75	10	13	59	0.648
GAME 2								
EC	13	68	59	870	101	186	443	0.599
BC	68	8	54	150	650	173	175	0.615
CC	75	570	150	60	650	29	176	0.641
SHAPELY	164	75	82	12	91	18	19	0.625
SHAPELY	67	13	384	331	405	500	212	0.604
+EdMot								
EDGly	437	600	488	316	873	435	718	0.655
GAME 3								
EC	13	68	59	870	101	186	443	0.622
BC	68	8	54	150	650	173	175	0.599
CC	75	570	150	60	650	29	176	0.610
SHAPELY	13	82	164	18	59	208	75	0.590
SHAPELY	67	2	482	441	164	10	212	0.636
+EdMot								
EDGly	151	117	225	149	453	22	91	0.650
GAME 4								
EC	13	68	59	870	101	186	443	0.586
BC	68	8	54	150	650	173	175	0.627
CC	75	570	150	60	650	29	176	0.645
SHAPELY	75	67	82	12	194	164	208	0.619
SHAPELY + EdMot	67	59	298	121	164	10	212	0.635
EDGly	164	7	19	75	10	13	91	0.656
GAME 5								
EC	13	68	59	870	101	186	443	0.647
BC	68	8	54	150	650	173	175	0.641
CC	75	570	150	60	650	29	176	0.599
SHAPELY	13	18	82	290	59	75	164	0.647
SHAPELY	67	75	107	650	164	247	212	0.604
+EdMot								
EDGly	707	174	19	17	841	13	356	0.666

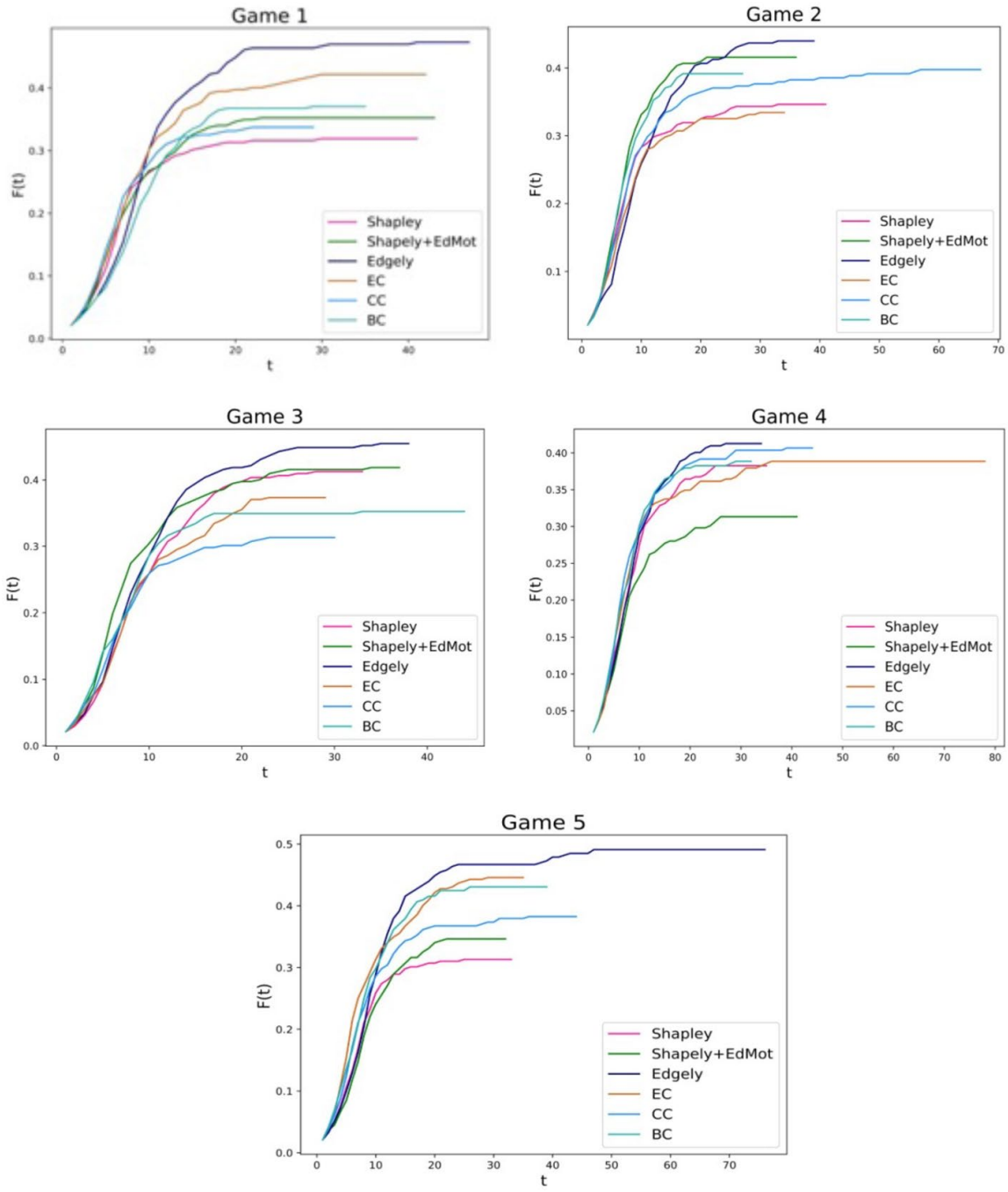


Fig. 2 Spreading speeds and influence of initially infected nodes of the various methods employed for USAIR97 dataset

7.3 Results

Tables 2 and 3 display the $F(t_c)$ results obtained in the SIR simulations for each of the five coalition games. Out of the seven communities identified, the most influential node from each community is identified via the EDGly algorithm. The final infected scale, produced by these spreading nodes are then compared for each algorithm. It can be seen that the

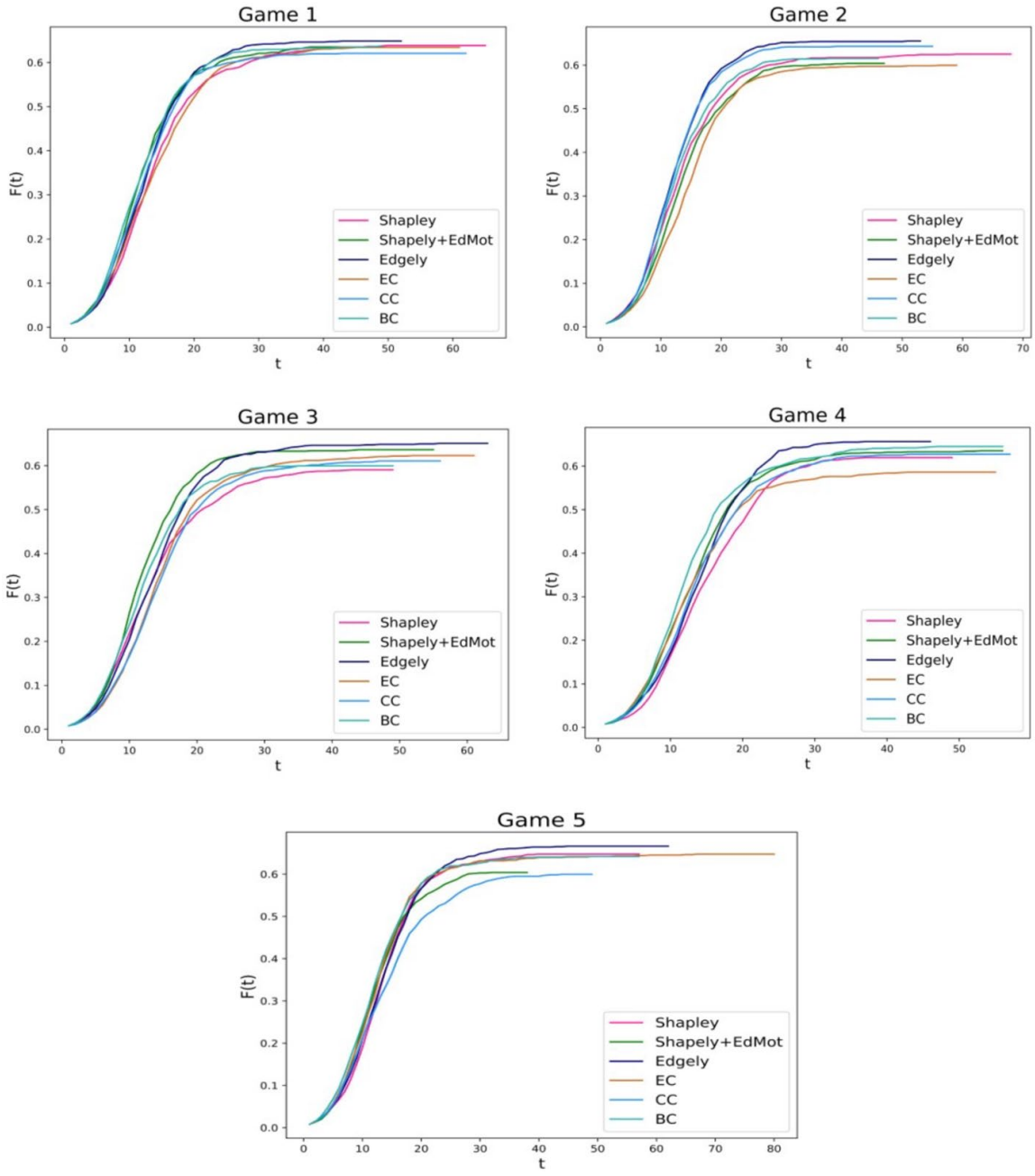


Fig. 3 Spreading speeds and influence of initially infected nodes of the various methods employed for Facebook dataset

nodes identified by the EDGly algorithm constantly result in a higher final infected scale $F(t_c)$ and hence a better coverage of the network.

It was observed that the change in the number of communities identified in the graph, does not show significant variation in the infected scale, hence we have picked the number of communities as 7.

In Figs. 2 and 3, each graph shows the experimental results of the infected scales by time, $F(t)/t$ received by the SIR Model, on the five different coalition games, using USAIR97 and Facebook datasets respectively. The results are obtained by taking the average of the results of 100 independent runs of SIR simulations. Larger $F(t)$ implies a

larger scale of control in a network with the same spreading time t, which in turn shows a faster spreading rate. It can be seen that the nodes identified by the EDGly algorithm result in a significantly higher final infected scale $F(t_c)$—depicted by navy blue—and thus, EDGly performs significantly better than the centrality measures EC, BC and CC, marked by colours—brown, light blue and sea green.

Further, it is seen that EDGly also consistently performs better than other algorithms. The addition of community detection using the EdMot—Louvain algorithm results in a significant jump in the infected scale, further the replacement of Louvain, by the ECG algorithm drastically improves the model. Thus, it can be seen from the figures that EDGly spreads faster than other benchmark methods in all games in almost all scenarios and over varied data.

8 Conclusion

We aimed at organizing the nodes into community structures and employing SV-based centrality algorithm for finding the most influential nodes from each community. EdMot approach, in tandem with ECG algorithm, was used to detect communities in a graph because of its unique quality of considering not only lower-order structures but also higher-order community structures. Finally, the SIR model was used to evaluate the capability and scope of spreading ability of each influential node identified during the process.

To evaluate our algorithm on various real-world scenarios, we examined five different game situations, thereby taking into consideration various approaches to determine the most influential nodes in a given dataset. This helped us to gain a deeper and holistic understanding of the game-theoretical algorithm.

It is observed that our proposed algorithm, EDGly algorithm, not only compares well to other commonly-accepted centrality measures, but also drastically improves results by the addition of the task of community detection, to the idea of using coalitional game theory. This is in line with our intuitive prediction, that the application of solely using coalitional game theory on an interconnected network, where each node acts as a player, and maximizes his cumulative yield over all permutations, creates a certain level of restriction in all the possible permutations.

With the application of community detection, we were able to analyze the network on a more individualistic level along with combining the positive effects of coalitions in game theory. This approach thus creates a fine balance between the two approaches and performs better than state of the art approaches.

The model has shown significant success as the most prominent nodes are successfully identified for both the datasets used. We plan to extend the idea of finding the most powerful node in social networks to numerous other similar applications for future work, such as the Internet, or urban networks, and involving a given node in disease dynamics. This will help us understand our algorithm's efficiency and accuracy in multiple applications in the real world.

Further, various optimisation techniques on the SV algorithm, for example, Fuzzy Logic will be implemented for mining much larger social networks and to improve accuracy and other relevant metrics of the project. Fuzzy-based implementation will solve various complexities and limitations that we're currently encountering.

Appendix

Algorithm 2: SVs for Game 1

Input: An unweighted graph G(N, E)

Output: SVs of all nodes in G

Initialise: $\forall\ n_i \in \mathrm{N(G)}$ set $\mathrm{SV}[n_i] \leftarrow 0$

for each $n_i \in \mathrm{N(G)}$ **do**

  $\mathrm{SV}[n_i] \leftarrow 1/(1 + n_i.degree)$

  for each $u_i \in n_i.neighbours$ **do**

    $\mathrm{SV}[n_i] \leftarrow \mathrm{SV}[n_i] + \frac{1}{(1 + u_i.degree)}$

  end for

end for

return SV

Algorithm 3: SVs for Game 2

Input: An unweighted graph G(N, E) and a positive integer p

Output: SVs of all nodes in G

Initialise: $\forall\ n_i \in \mathrm{N(G)}$ set $\mathrm{SV}[n_i] \leftarrow 0$

for each $n_i \in \mathrm{N(G)}$ **do**

  $\mathrm{SV}[n_i] \leftarrow \min\left(1, \frac{\mathrm{p}}{(1 + n_i.degree)}\right)$

  for each $u_i \in n_i.neighbours$ **do**

    $\mathrm{SV}[n_i] \leftarrow \mathrm{SV}[n_i] + \max\left(0, \frac{u_i.degree - \mathrm{p} + 1}{u_i.degree * (1 + u_i.degree)}\right)$

  end for

end for

return SV

Algorithm 4: SVs for Game 3

Input: A weighted graph G(N, E, W) and a positive cut-off value d

Output: SVs of all nodes in G

Initialise: $\forall\ n_i \in$ N(G) set SV[n_i] $\leftarrow$ 0

for each $n_i \in$ N(G) **do**
  Distance_Vector D $\leftarrow$ Dijkstra(n_i)
  extended_neighbours $\leftarrow$ empty 2D array
  extended_degree[n_i] $\leftarrow$ 0
  for each $u_i \in$ N(G) such that $u_i \neq n_i$ **do**
    if D[u_i] $\leq$ d **then**
      extended_neighbours[n_i].add(u_i)
      extended_degree[n_i]++
    end if
  end for
end for

for each $n_i \in$ N(G) **do**
  SV[n_i] $\leftarrow$ $\frac{1}{1+\text{extended_degree}[n_i]}$
  for each $u_i \in$ extended_neighbours[n_i] **do**
    SV[n_i] $\leftarrow$ SV[n_i] + $\frac{1}{(1+\text{extended_degree}[u_i])}$
  end for
end for
return SV

Algorithm 5: SVs for Game 4

Input: A weighted graph G(N,E,W) and function $f: \mathbb{R}^+ \rightarrow \mathbb{R}^+$
Output: SVs of all nodes in G
Initialise: $\forall\ n_i \in$ N(G) set SV[n_i] ← 0

for each $n_i \in$ N(G) **do**
  [Distance D , Nodes w] ← Dijkstra(n_i)
  sum ← 0, index ← |N|-1 , prev_dist ← -1, prevSV ← -1
  while index > 0 **do**
    if D(index) == prev_dist **then**
      currSV = prevSV
    else
      currSV = $\frac{f(D(index))}{1+index}$ - sum
    end if
    SV[w(index)] ← currSV + SV[w(index)]
    sum ← sum + $\frac{f(D(index))}{1+index}$
    prev_dist = D(index) , prevSV = currSV
    index ← index - 1
  end while
  SV[n_i] ← SV[n_i] + *f(0)* – sum
end for
return SV

Algorithm 6: SVs for Game 5

Input: A weighted graph G(N, E, W) and cut-offs $W_{cutoff}(n_i)$ for each $n_i \in$ N(G)

Output: SVs of all nodes in G

Initialise: $\forall\, n_i \in$ N(G) set SV$[n_i] \leftarrow 0$

for each $n_i \in$ N(G) **do**
    compute and store α_i and β_i
end for

for each $n_i \in$ N(G) **do**
    for each m in 0 to $n_i.\,degree$ **do**
        compute $\mu \leftarrow \underset{m}{\mu}(X^{ii})$, $\sigma \leftarrow \underset{m}{\sigma}(X^{ii})$
        compute $p \leftarrow \Pr\{\mathcal{N}(\mu, \sigma^2) < W_{cutoff}(n_i)\}$
        SV$[n_i] \leftarrow$ SV$[n_i] + \dfrac{p}{1+ n_i.degree}$
    end for

    for each $v_j \in n_i.\,neigh\,bours\; n_i.\,neigh\,bours$ **do**
        p $\leftarrow$ 0
        for each m in 0 to $n_i.\,degree$ **do**
            compute $\mu \leftarrow \underset{m}{\mu}(X^{ij})$, $\sigma \leftarrow \underset{m}{\sigma}(X^{ij})$
            compute $z \leftarrow \underset{m}{\mathcal{Z}^{ij}}$
            p $\leftarrow$ p + $\dfrac{\text{z * (v}_i\text{.degree-m)}}{v_j.degree\ *\ (v_j.degree+1)}$
        end for
        SV$[n_i] \leftarrow$ SV$[n_i]$ + p
    end for
end for return SV

References

1. Aadithya KV, Ravindran B, Michalak TP, Jennings NR (2010) Efficient computation of the shapely value for centrality in networks. In: Saberi A (ed) Internet and Network Economics, WINE, Lecture Notes in Computer Science, vol 6484
2. Al-Garadi MA, Varathan KD, Ravana SD (2017) Identification of influential spreaders in online social networks using interaction weighted k-core decomposition method. Phys A 468:278–288
3. Amati G, Angelini S, Gambosi G, Rossi G, Vocca P (2019) Influential users in Twitter: detection and evolution analysis. Multimed Tools Appl 78:3395–3407
4. Bhuiyan BA (2018) An overview of game theory and some applications. Philos Progress 59(1–2):111–128
5. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech. https://doi.org/10.1088/1742-5468/2008/10/P10008
6. Bureau of Transportation Statistics. https://transtats.bts.gov/
7. Chen W, Teng SH (2017) Interplay between social influence and network centrality: a comparative study on shapley centrality and single-node-influence centrality. In: WWW '17: Proceedings of the 26th international conference on World Wide Web, pp 967–976
8. Domingos P, Richardson M (2001) Mining the network value of customers. In: KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and Data mining, pp 57–66

9. Freeman LC (1978–1979) Centraity in social networks conceptual clarification. Soc Netw 1(3):215–239
10. Gao S, Ma J, Chen Z, Wang G, Xing C (2014) Ranking the spreading ability of nodes in complex networks based on local structure. Phys A 403:130–147
11. Guo C, Chen X, Chen D, Gao H, Yang L, Ma J (2020) Influential nodes identification in complex networks via information entropy. Entropy 22:242
12. Huang L, Wang C, Chao H (2018) A harmonic motif modularity approach for multi-layer network community detection. In: IEEE International Conference on Data Mining (ICDM), pp 1043–1048
13. Irfan MT, Ortiz LE (2011) A game theoretic approach to influence in networks. In: Proceedings of the twenty-fifth AAAI conference on artificial intelligence (AAAI'11), AAAI Press, pp 688–694
14. Jiang L, Zhao X, Ge B, Xiao W, Ruan Y (2019) An efficient algorithm for mining a set of influential spreaders in complex networks. Phys A 516:58–65
15. Li PZ, Cai YX, Wang CD, Liang MJ, Zheng YQ (2019) Higher-order brain network analysis for auditory disease. Neural Process Lett 49:1–19
16. Li PZ, Huang L, Wang CD, Huang D, Lai JH (2018) Community detection using attribute homogenous motif. IEEE Access 6:47707–47716
17. Li PZ, Huang L, Wang CD, Lai JH (2019) EdMot: An edge enhancement approach for motif-aware community detection. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '19), Association for Computer Machinery, New York, pp 479–487
18. Liu F, Li M (2019) A game theory-based network spreading model: based on game experiments. Int J Mach Learn Cybern 10:1449–1457
19. Liu HL, Ma C, Xiang BB, Tang M, Zhang HF (2018) Identifying multiple influential spreaders based on generalized closeness centrality. Phys A 492:2237–2248
20. Ll Ma, Ma C, Zhang HF, Wang BH (2016) Identifying influential spreaders in complex networks based on gravity formula. Phys A 451:205–212
21. Mao C, Xiao W (2018) A comprehensive algorithm for evaluating node influences in social networks based on preference analysis and random walk. Complexity 40:1–16
22. Michalak TP, Aadithya KV, Ravindran B, Jennings NR, Szczepanski PL (2013) Efficient computation of the shapley value for game-theoretic network centrality. J Artif Int Res 46(1):607–650
23. Ni C, Sugimoto CR, Jiang J (2013) Degree, closeness, and betweenness: application of group centrality measurements to explore macro-disciplinary evolution diachronically
24. Nie T, Guo Z, Zhao K, Lu ZM (2016) Using mapping entropy to identify node centrality in complex networks. Phys A 453(2016):290–297
25. Opsahl T (2013) Triadic closure in two-mode networks: redefining the global and local clustering coefficients. Soc Netw 35(2):159–167. https://doi.org/10.1016/j.socnet.2011.07.001
26. Poulin V, Théberge F (2018) Ensemble clustering for graphs. arXiv:1809.05578
27. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D (2004) Defining and identifiying communities in networks. PNAS 101(9):2658–2663
28. Shetty J, Adibi J (2005) Discovering important nodes through graph entropy: the case of Enron email database. In: LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery, pp 74–81
29. Stephenson K, Zelen M (1989) Rethinking centrality: methods and examples. Soc Netw 11(1):1–37
30. Tan CW, Yu PD, Zheng L, Lai CK, Zhang W, Fu HL (2016) Optimal detection of influential spreaders in online social networks. In: Proceeding of annual conference on information sciences and systems, pp 145–150
31. Tarkowski MK, Michalak TP, Rahwan T, Wooldridge M (2017) Game-theoretic network centrality: a review. arXiv preprint arXiv:1801.00218
32. Wei H, Pan Z, Hu G, Zhang L, Yang H, Li X, Zhou X (2018) Identifying influential nodes based on network representation learning in complex networks. PLoS ONE 13(7):e0200091
33. Yadati N, Narayanam R (2011) Game theoretic models for social network analysis. In: Proceedings of the 20th international conference companion on World Wide Web–WWW' 11, pp 291–292
34. Yang Z, Algesheimer R, Tessone CJ (2016) A comparative analysis of community detection algorithms on artificial networks. Sci Rep 6:30750
35. Zhan J, Gurung S, Parsa SPK (2017) Identification of top-k nodes in large networks using Katz centrality. J Big Data 4:16
36. Zhang X, Zhu J, Wang Q, Zhao H (2013) Identifying influential nodes in complex network with community structures. Knowl-Based Syst 42:74–84
37. Zhao L, Cui H, Qiu X, Wang X, Wang J (2013) SIR Rumour spreading model in the new media age. Phys A 392(4):995–1003

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

Minni Jain obtained her M. Tech. degree in Information Security and B. Tech. degree in Information Technology. She is currently working as an Assistant Professor in Delhi Technological University, Delhi, India. Her major research interests include Natural Language Processing, Sentiment Analysis, Information Security, Neural Networks and Fuzzy Logic

Aman Jaswani obtained his B. Tech. degree in Software Engineering from Delhi Technological University, Delhi, India. He has worked on numerous projects in the field of Machine Learning and Natural Language Processing. His areas of research include Game Theory, Machine Learning, Natural Language Processing and Data Science

Ankita Mehra obtained her B. Tech. degree in Software Engineering from Delhi Technological University, Delhi, India. She is a final year student. Her areas of research include Game Theory, Machine Learning, Deep Learning, Computer Vision and Natural Language Processing

Laqshay Mudgal obtained his B. Tech. degree in Software Engineering from Delhi Technological University, Delhi, India. His major research interests include Artificial Intelligence, Machine Learning, Deep Learning, Natural Language Processing, and Data Science

Effect of swirl flow on the performance of parallel hub axial annular diffuser

DOI : 10.36909/jer.9631

Hardial Singh[*], B.B.Arora[*,*]

[*] Department of Mechanical Engineering, Delhi Technological University, Delhi- India

* Corresponding Author: bbarora@dce.ac.in

ABSTRACT

In the present work, the parallel hub axial flow annular diffuser's performance characteristics with divergent casing varying between equivalent cone angle (10°, 15°, and 20°) with area ratio 3 have been evaluated computationally as well as experimentally. The performance of three diffusers were tested at different inlet swirl angles (from 0° to 25°) for swirling and non-swirling flow. Simulations have been carried out on a fully developed flow at Reynolds number 2.5×10^5. The results were analyzed based on the velocity profiles, static pressure recovery coefficient, and the total pressure loss coefficient. The result analysis shows that the inlet swirl flow improves the recovery of pressure and also delays the flow separation on the casing. Moreover, the findings also show that the best performance was achieved in equivalent cone angle 10° at the inlet swirl angle of 7.5° compared to other diffusers.

Key words: Annular diffuser; Static pressure recovery coefficient; Total pressure loss coefficient; Inlet swirl angle.

INTRODUCTION

The annular diffuser is a mechanical flow diffusing device that is used for many engineering applications. The diffuser's primary function is to recuperate the static pressure due to reducing the flow's mean velocity. The Annular diffuser is mainly used in the pump, compressor, gas turbine engine, turbomachinery, etc. (Kline, 1959). The fluid flow around the bearing, vanes,

and strut creates a swirl effect at the diffuser's entry. The swirl effect at the inlet is one way to increase the overall performance of the diffuser. Reneau et al. (1967) experimentally examined the effect of flow characteristics on two-dimensional straight wall diffusers. It was found that diffuser performance decreases as the inlet blockage increases in the two-dimensional conical diffusers. Coladipietro (1974) studied the inlet swirl and boundary layer thickness effect on the performance of diffusers. The efficiency and pressure recovery coefficient were more pronounced in a thin boundary layer for a small non-dimensional length. Adenubi (1976) experimentally studied the flow regimes of straight core annular diffuser with downstream turbo machine inlet conditions. It was observed that the annular diffuser's effectiveness enhances by increasing the turbulence level at the inlet. Singh et al. (1994) found that swirl improved the recovery of pressure and suppressed the separation of flow on the casing in the wide-angle annular diffuser. Mohan et al. [1998] analyzed the optimum performance of the diffusers with the computational approach with less cost than the experimentation. It shows that maximum pressure recovery improvement is achieved at a 17.5° diffuser. Shuja and Habib (1996) studied the axisymmetric annular diffuser and concluded that the increase in the flow region's heat transfer coefficient depends on the cant angle. They have also shown that the Nusselt number is a function of the Reynolds number. Ubertini and Desideri (2000) described the gas turbine annular exhaust diffuser with and without strut as having effectiveness 15 percent and less than 15 percent, respectively. It is shown that overall diffuser losses enhanced in the axial region of a strut. Arora and Pathak (2009); Arora et al. (2010); Arora (2014); Singh and Arora (2019) carried out the experimental and numerical investigation of geometrical parameters on the annular diffuser with swirl flow. They concluded that the recovery of static pressure is faster with the introduction of swirl at the diffuser's entry.

Very little research work has been carried out experimentally on the swirl flow in annular diffusers. Based on the exhaustive literature review, it is very difficult to conclude

which choice of geometric parameters and dynamic parameters will give diffusers the desirable performance because the fluid flow is quite complicated. Since there is no research work carried out on the different cone angles and has the same area ratio, it needs to be explored. The main work is focused on the parallel hub axial annular diffuser with varying geometries but having the same area ratio 3. The effect of swirl intensity is analyzed with different swirl regimes on the flow characteristics of annular diffusers. Further, the prediction of the flow separation and reattachment of the flow are carried out inside the annular diffusers. Figure 1 and Table 1 show the geometrical configuration details of the parallel hub axial annular diffuser parameters.

Table 1 Geometrical parameters of parallel hub axial annular diffusers.

R_{hi}=3.8 cm, R_{ci}=7.75 cm, θ_h=0 (deg), AR=3				
Sr. No.	**θ_c (deg)**	**L(cm)**	**L/ΔR**	**θ_e (deg)**
A	4.6	56.51	14.30	10
B	6.91	37.55	9.50	15
C	9.22	28.04	7.09	20

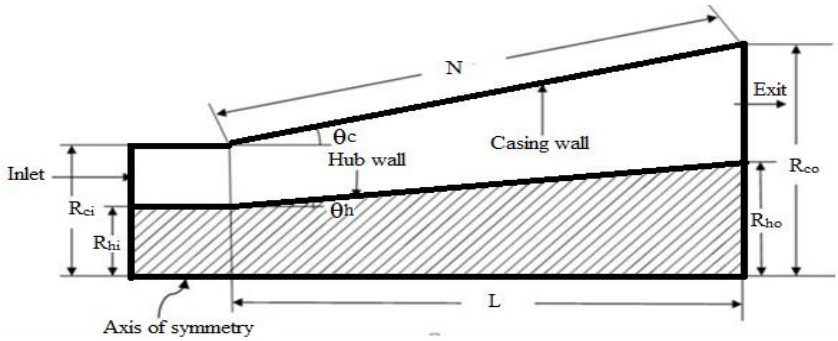


Figure 1 Diffuser geometrical parameter.

EXPERIMENTAL SET UP

Figure 2 (i) shows the annular diffuser experimental set up used for experimentation. Figure 2 (ii) represents the setup's main components i.e. blower, settling chamber, swirl plate, annular passage, and diffuser. The centrifugal blower sucks air from the ambiance then delivers it to the settling chamber with a conical divergence section, which is precisely aligned with the settling chamber. The airflow rate is controlled by changing the rpm of the impeller and throttling valve placed at the inlet. The settling chamber consists of a honeycomb and different mesh screen

sizes to prevent the flow fluctuation, reduce turbulence, and make the flow uniform. A well designed conical contraction zone of settling chamber connected to the long annulus passage in which flow commences. The swirl plate is located upstream of the test diffuser to avoid the wake effect at the test diffuser's entry to recover the pressure. Test diffuser made of transparent Perspex and hub casted of aluminum material, which is perfectly machined to get a smooth surface fabricated with a precision of ±1 percent tolerance. The numbers of static pressure taps on the casing and hub wall are made along the test diffuser's length to measure static pressure with manometers. The manometers are inclined at an angle of 10° with the horizontal to increase the sensitivity of reading. The longitudinal velocity and swirl velocity are measured along with the test diffuser's axial locations using a three-hole cobra probe using the null technique (Bryer and Pankhurst, 1971). The cobra probe yaw angle is set to zero by aligning with the incoming flow in this technique. The calibration range is associated with the cobra probe to measure the velocity within ±55 deg. The manometer's uncertainty for measurement of the static pressure is ± 1.5mm of water, and the total pressure is ±1mm of water.

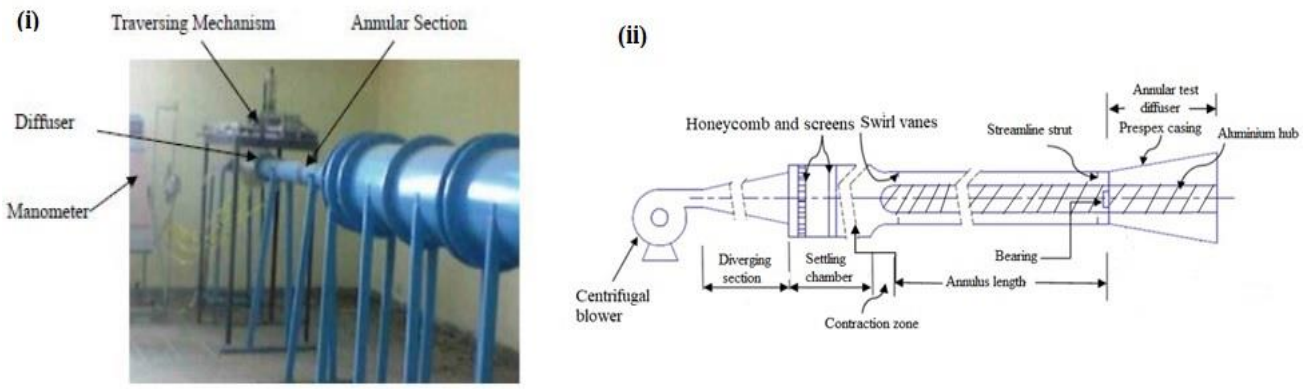


Figure 2 (i) Experimental set up; (ii) Schematic of the experimental set up.

MATHEMATICAL FORMULATIONS

The equation of continuity in 2D axisymmetric geometry is given as

$$\frac{\partial}{\partial x}(\rho v_x) + \frac{\partial}{\partial r}(\rho v_r) + \frac{\rho v_r}{r} = s_m \qquad (1)$$

Where υ_x is the axial direction velocity, and υ_r is the radial direction velocity. S_m is the mass added to the continuous phase from the dispersed phase.

The governing equations of momentum conservation for steady axisymmetric non-swirling flow in the axial and radial directions are given as

$$\frac{1}{r}\frac{\partial}{\partial x}(r\rho v_x v_x)+\frac{1}{r}\frac{\partial}{\partial r}(r\rho v_r v_x)= -\frac{\partial p}{\partial x}+\frac{1}{r}\frac{\partial}{\partial x}\left[r\mu\left(2\frac{\partial v_x}{\partial x}-\frac{2}{3}(\nabla.\vec{v})\right)\right]+\frac{1}{r}\frac{\partial}{\partial r}\left[r\mu\left(\frac{\partial v_x}{\partial r}+\frac{\partial v_r}{\partial x}\right)\right]+F_x \quad (2)$$

And

$$\frac{1}{r}\frac{\partial}{\partial x}(r\rho v_x v_r)+\frac{1}{r}\frac{\partial}{\partial r}(r\rho v_r v_r)= -\frac{\partial p}{\partial r}+\frac{1}{r}\frac{\partial}{\partial x}\left[r\mu+\left(\frac{\partial v_r}{\partial x}+\frac{\partial v_x}{\partial r}\right)\right]+\frac{1}{r}\frac{\partial}{\partial r}\left[r\mu\left(2\frac{\partial v_r}{\partial r}-\frac{2}{3}(\nabla.\vec{v})\right)\right] -2\mu\frac{v_r}{r^2}+\frac{2}{3}\frac{\mu}{r}(\nabla.\vec{v})+\rho\frac{v_z^2}{r}+F_r \quad (3)$$

Where

$$\nabla.\vec{v}=\frac{\partial v_x}{\partial x}+\frac{\partial v_r}{\partial r}+\frac{v_r}{r} \quad (4)$$

Here F_x and F_r represent axial and radial forces respectively.

The tangential momentum equation for swirling flows need to be solved, which can be written as

$$\frac{1}{r}\frac{\partial}{\partial x}(r\rho v_x v_z)+\frac{1}{r}\frac{\partial}{\partial r}(r\rho v_r v_z)= \frac{1}{r}\frac{\partial}{\partial x}\left[r\mu\frac{\partial v_z}{\partial x}\right]+\frac{1}{r^2}\frac{\partial}{\partial r}\left[r^3\mu\frac{\partial}{\partial r}\left(\frac{v_z}{r}\right)\right]-\rho\frac{v_r v_z}{r} \quad (5)$$

Where υ_z represents the swirl velocity

The turbulent flow parameters are computed, turbulence models are analyzed to the close solution of the RANS set of equations (1-5). The selection of turbulence models such as k-ε standard, k-ω standard, k-ε Realizable, k-ε RNG, and k-ω SST on the annular diffuser with swirl flow is a complicated process. The RNG k- ε turbulence model is a more advanced model as compared to the k–ε standard model.

Mohan et al. (1998); Singh et al. (2006); Arora and Pathak (2009); Arora et al.(2010) have adopted two-equation turbulence model k-ε RNG on annular diffusers for their work while work using k-ε Standard model was carried out by Sadasivan et al. (2019). The result from the simulation of a non-dimensional longitudinal velocity profile distribution at x/L=0.7 is shown in Figure 3. It is seen that the k-ε RNG turbulence model gives the lowest average deviation as compared to the experimental results. Hence k-ε RNG turbulence model is employed for simulation in the present study.

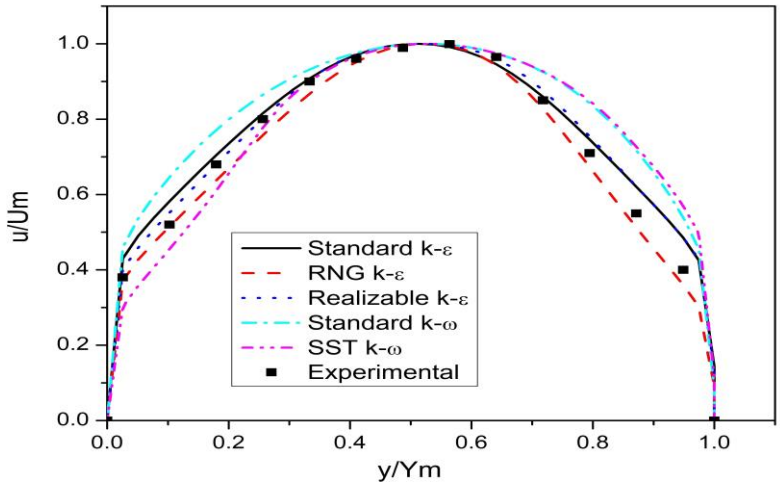


Figure 3 Comparison of the turbulence models with the experimental results of the longitudinal velocity profile at x/L=0.7 of equivalent cone angle 10°.

COMPUTATIONAL DOMAIN AND BOUNDARY CONDITIONS

An axis-symmetric two-dimensional computational domain of the annular diffuser has been sketched in the ANSYS Fluent v16.2 software. A map meshing scheme has been used with a quadrilateral element to carry out the simulations. The 2D mesh geometry of the computational domain has been shown in Figure 4. Very fine mesh is placed near the wall by maintaining the

y+<1. The first node of the mesh is placed at a distance of 0.02mm from the wall to predict the velocity gradient near the wall. The quality of the mesh is maintained in terms of skewness, aspect ratio, and orthogonality. The inlet velocity profile is fed in the axisymmetric diffuser similar to the velocity profile obtained from the experimental setup. The turbulent intensity 3% is applied at the inlet, and the value of the hydraulic diameter is calculated as per the diffuser's geometry. The boundary conditions at the outlet are applied 3% turbulent intensity normal to the boundary, zero gauge pressure, and thus calculated value of the hydraulic diameter (h_d) from the geometry of diffuser. The 2^{nd} order upwind scheme is used to control the solution for swirl velocity, momentum, turbulent kinetic energy (k), and turbulent dissipation rate (ω) for better accuracy of results. The employed residuals for convergence criteria is 10^{-6} to achieve convergence solution by an iteration method. The SIMPLE algorithm method is used for the stability of the solution.

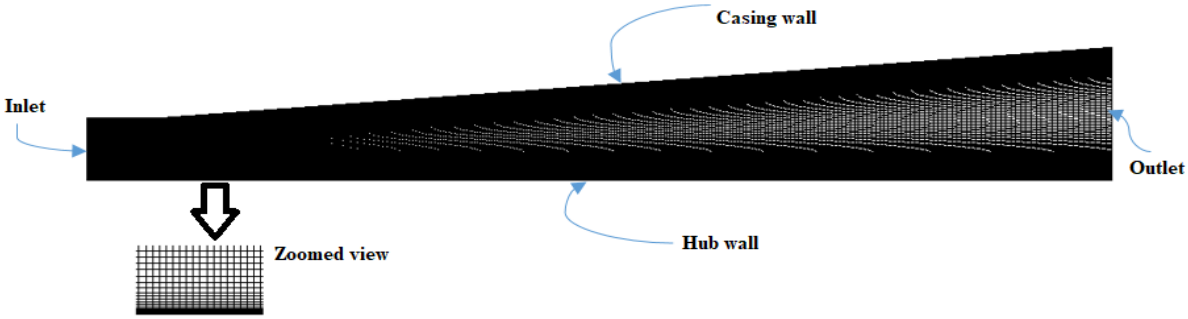


Figure 4 2D axisymmetric computational domain with meshing.

GRID INDEPENDENCE TEST

In the current work, a grid test is carried out to obtain mesh independence results. The different diffusers having various geometrical designs will have different grid sizes depending upon the geometrical dimension.

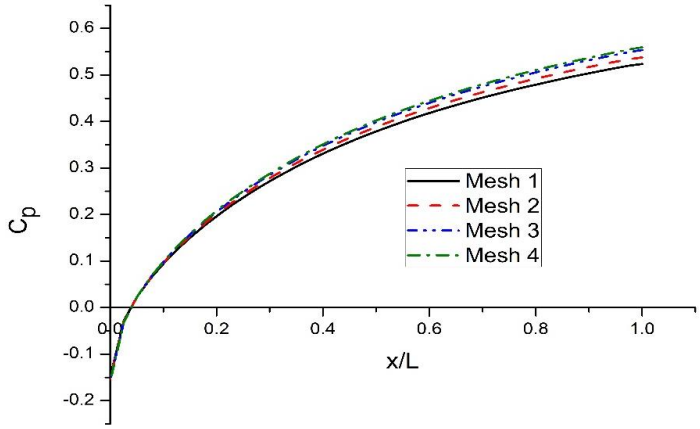


Figure 5 Pressure recovery coefficient at casing wall comparison for four mesh sizes.

In the parallel hub axial annular diffuser having an equivalent cone angle 10° with area ratio 3 of four grid sizes (mesh 1=90000, mesh 2=120000, mesh 3=160000, and mesh 4=190000) were tested at Reynolds number $2.5x10^5$ using k-ε RNG turbulence model for swirl flow. Pressure coefficient profiles for Mesh 3 and Mesh 4 are very close to each other, and hence cell count of Mesh 3 is selected for the current study, as shown in Figure 5. The simulation is performed on the i5 processor with 8GB RAM, and the time required to solve the simulation of 160000 elements is 18hrs.

RESULTS AND DISCUSSION

The computationally determined velocity vectors, longitudinal velocity profiles, swirl velocity profiles, static pressure recovery coefficient (C_p), and total pressure loss coefficient (C_{TL}) across the diffuser's length are presented and discussed.

Velocity vector

The axisymmetric 2D annular diffusers (A, B, and C) with and without swirl flow of velocity vectors at the inlet swirl angles (0°, 12°, and 25°) are shown in Fig. 6. The velocity vectors show the result across the length i.e. inlet, x/L=0.1,0.3,0.5,0.7,0.9, and outlet using CFD.

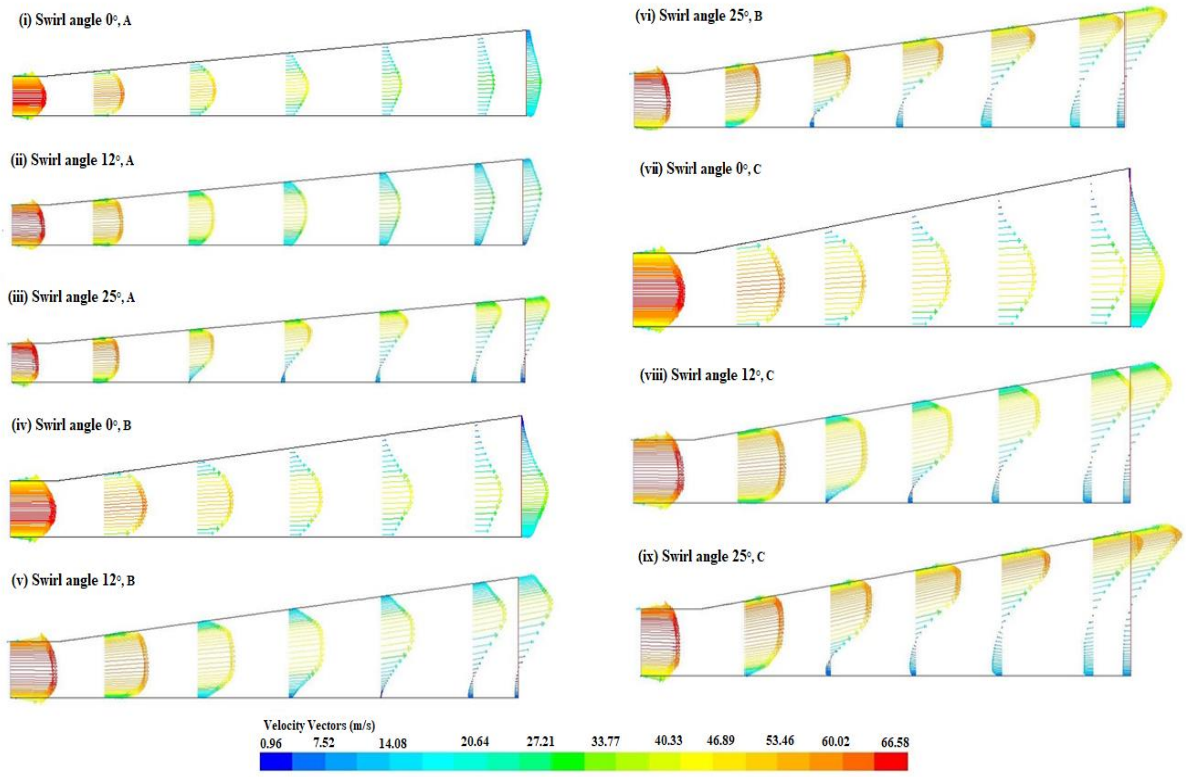


Figure 6 (i-x) Velocity vectors at inlet swirl angles 0°, 12°, and 25°of equivalent cone angle 10°, 15°, and 20°.

The uniform flow is distributed between the casing and hub wall at swirl angles 0°, and 12°as flow moves toward the downward direction in the diffuser A. The flow separation is observed on the inner wall at the inlet swirl angle 25° along the two-thirds length of diffuser A. In diffuser B, the reverse flow is observed at the swirl angles 12° and 25° on the hub walls due to the adverse pressure gradient. The high casing divergent angle in diffuser C due to that flow separation is observed on the casing wall with the non-swirling flow. The separation zone's size increases with the inlet swirl angles (12° and 25°) and results in the shifting of flow separation position and negative velocity towards the hub wall. For the 25° inlet swirl angle, the flow visualization in the velocity vectors shows that flow is accumulated on the casing wall, and a large separations zone exists on the hub wall.

Velocity profile

The flow behavior and performance of diffuser (A) have been shown in Figures (7-8) in the form of longitudinal velocity profiles and swirl velocity profiles concerning average mass velocity with different inlet swirl conditions i.e. 0°, 12°, and 25°. The hub position represents as y/Ym =0, and the casing position represents as y/Ym = 1 (Klomp, 1997). The graphs shown in Figures (7-8) are drawn at different location i.e. x/L = 0.1, 0.3, 0.5, 0.7, and 0.9 for diffuser A of area ratio 3. For no swirl condition, flow is accumulated near the hub wall at an equivalent cone angle of 10°, as shown in Figure 7 (i). The velocities recede at a faster rate on the walls as the continuous diffusion occurs throughout the diffusers. The tangential movement generated by the swirl 12° is not enough to move the bulk flow from the hub wall to the casing wall, as shown in Figure 7 (ii). Figure 7 (iii) shows a reversal trend for a 25° swirl in the longitudinal velocity distribution. In this case, bulk flow moves toward the casing wall. The reversal of flow at swirl angle 25° for x/L=0.5,0.7 and 0.9 is reversed up to the y/Ym=0.18,0.25, and 0.27, respectively. Figure 8 (i-ii) shows swirl velocity distribution at the 12°and 25° inlet swirl angles. The trends are similar for 12°and 25° swirl angles, but the value of magnitude is different. The bulk flow forced towards the casing in the case of a swirl angle 25°. These trends are based on the conservation of the angular momentum principle. The flow is moving towards the casing wall by introducing a swirl at the inlet.

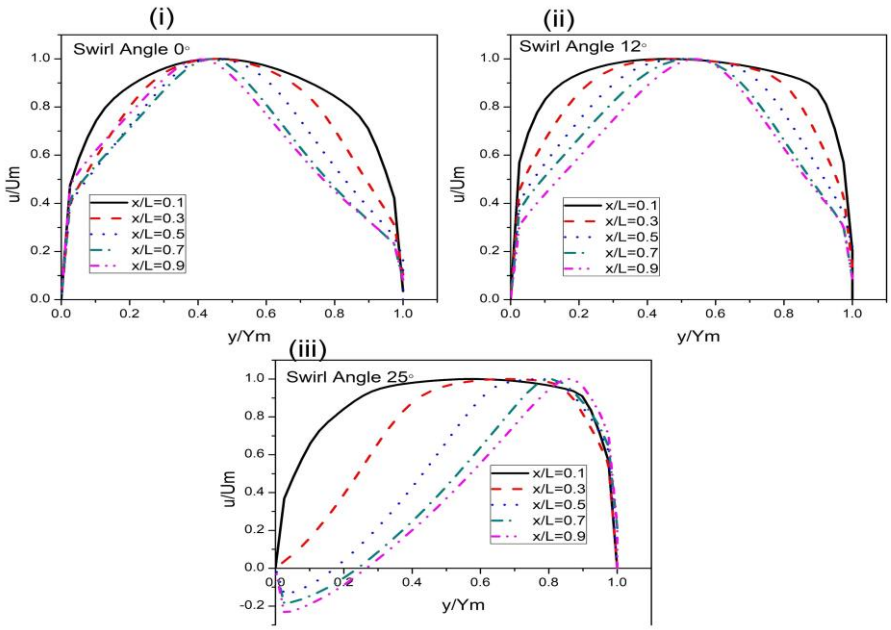


Figure 7 (i-iii) Longitudinal Velocity distribution at inlet swirl angles 0°, 12°, and 25°of equivalent cone angle 10°.

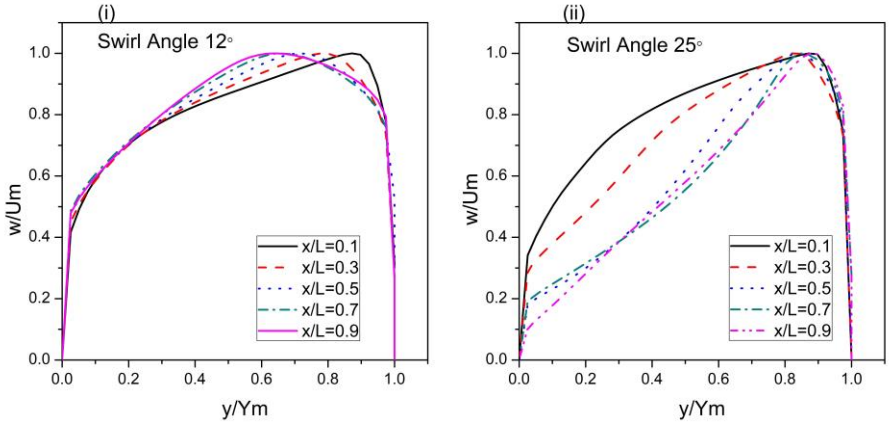


Figure 8 (i-ii) Swirl Velocity distribution at inlet swirl angles 12° and 25°of equivalent cone angle 10°.

Pressure recovery coefficient

The coefficient of pressure recovery has been determined in the annular diffuser having equivalent cone angle 10° with inlet swirl angles 0° to 25° at the diffuser's casing wall. The graph is shown in Figure 9 (i) for area ratio 3 with normalized axial length x/L (Ibrahim et al. 2007). It can be seen from plots that Cp increases with the flow downstream of the diffuser passage. The curves drawn represent the value of Cp for swirl flow almost identical to the 0° swirl. The figure shows that marginal Cp improves with the downstream diffuser passage in

the beginning and later on decreases due to separation of flow on the hub as observed from the longitudinal velocity profile distribution. Swirl increases the Cp at the entrance of the downstream diffuser passage as compared to the non-swirl flow. The highest Cp is up to x/L=0.18 at equivalent cone angle 10° for 25° inlet swirl angles of the diffuser passage. The Cp plot is shown in Figure 9 (i) shows the maximum improvement of the pressure recovery in an equivalent cone angle of 10° with a 7.5°inlet swirl angle. The variation of static pressure recovery coefficient (Cp) with the inlet swirl angle of three types of diffusers (A, B, and C) is shown in Figure 10 (i). These plots indicate that there is a maximum coefficient of static pressure recovery with an inlet swirl angles 7.5° and 12° of equivalent cone angles 10° and 15°, respectively. The pressure recovery increases up to x/L=0.60 and 0.54 for swirl angles 7.5° and 12° respectively of equivalent cone 20° but later on decreases due to separation of flow and decay swirl intensity. The coefficient of static pressure recovery (Cp) with a variation of the area ratio of diffusers A, B, and C along the length is shown in Figure 10 (ii). It is observed from the figure that maximum pressure recovery is obtained from the equivalent cone angle 10° for an A-type diffuser.

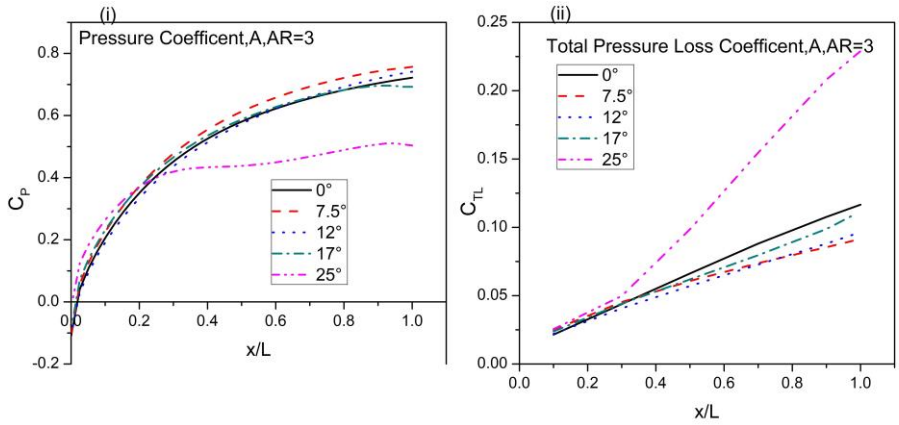


Figure 9 (i) Pressure coefficient at casing of equivalent cone angle 10°; (ii) Total pressure loss coefficient of equivalent cone angle 10°.

Total pressure loss coefficient

The total pressure loss coefficient increases almost linearly across the longitudinal length. As shown in Figure 9 (ii), the C_{TL} plot shows the minimum loss coefficient for equivalent cone angle 10° at 7.5° inlet swirl angle. This result implies that the inlet swirl angle's optimum selection suppresses the flow separation. The total pressure loss coefficient (C_{TL}) with the swirl angle is shown in Figure 11 (ii). The minimum loss of total pressure coefficient for equivalent cone angle 15° is up to 12° swirl angle and gives optimum performance at that angle. It also shows the minimum loss of total pressure coefficient at 7.5° swirl angle for equivalent cone angle 20°. The maximum loss coefficient is due to the casing's high cant angle, the swirl intensity's decay, and the adverse pressure gradient on the walls.

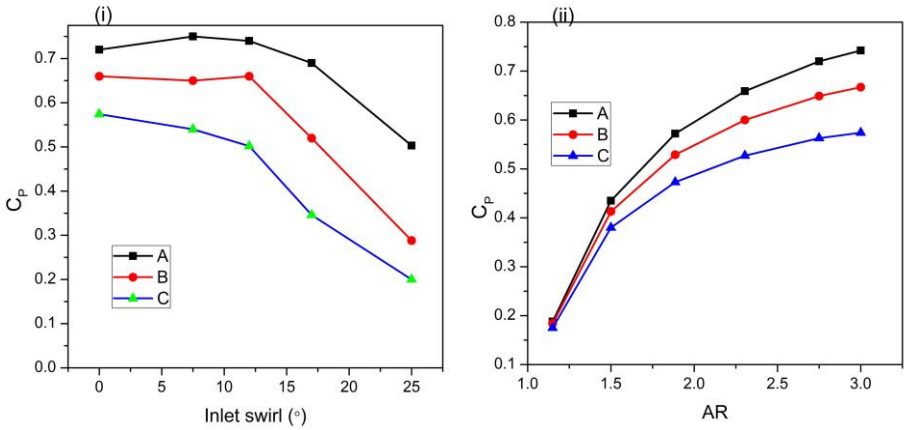


Figure 10 (i) Static pressure recovery coefficient with inlet swirl of diffusers A, B, and C; (ii) Static pressure recovery coefficient (C_p) with variation of area ratio of diffusers A, B, and C along the length.

Effectiveness

The diffuser's effectiveness (η) is plotted about the inlet swirl angle of A, B, and C types diffusers are shown in Figure 20 (i) (Hoadley, 1970). The plots indicate that η decreases with increasing equivalent cone angle and the highest value obtained for the A-type diffuser with a swirl angle of 7.5°. The diffuser's effectiveness decreases beyond the swirl angle of 12° for A and B types diffusers due to the hub's separation and reversal flow. In C type diffuser, the η decreases by the swirling flow due to wide space availability, which promotes separation of

flow either from the hub or casing wall, as seen in the velocity vector contours of equivalent cone angle 20°.

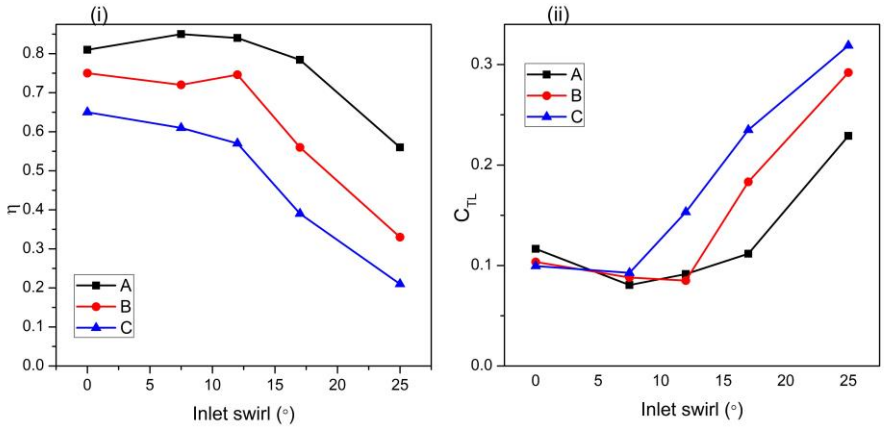


Figure 11 (i) Effectiveness with inlet swirl of diffusers A, B, and C ; (ii) Total pressure loss coefficient with inlet swirl of diffusers A, B, and C.

CONCLUSION

The following inferences have been highlighted from the numerical investigation of the incompressible flow through the parallel hub axial annular diffuser having equivalent cone angles (10°, 15°, and 20°) and the same area ratio (AR=3).

1. The longitudinal velocity profile is not symmetrical about the center at any cross-section of diffusers. The diffuser's non-swirling flow is accumulated at the hub, which is shifted towards the casing side with swirl flow.
2. The maximum value of Cp achieved at the exit of the diffuser are 0.75,0.67 and 0.56, corresponding to 7.5°,12°, and 0° swirl angle and 10°,15°, and 20° equivalent cone angles, respectively. Hence it is found that the maximum coefficient of pressure recovery increases up to 75% at swirl angle 7.5° of equivalent cone angle 10° for area ratio 3.
3. The total pressure loss coefficient (C_{TL}) is minimum at the optimum swirl angle along the diffuser passage. The simulated results show that the optimum swirl angles are 7.5°and 12° at equivalent cone angles 10° and 15°, respectively.

4. The stall is observed on the casing for equivalent cone angle 20° in non-swirl flow, but it vanishes and moves away from the casing wall with swirl flow.

NOMENCLATURE

θ	Divergence angle of wall, deg;	Subscripts	
ΔR	Annulus height, cm;	c	Casing;
L	Length of diffuser, cm;	h	hub;
x/L	Non-dimensional axial length;	i	Inlet;
y/Ym	Non-dimensional radial length.	o	Outlet;
		e	Equivalent.

REFERENCES

Adenubi, S.O., 1976. Performance and flow regime of annular diffusers with axial turbomachine discharge inlet conditions. J. Fluids Eng., 98(2): 236-242.

Arora, B. B., Pathak, B. D., 2009. Effect of Geometry on the Performance of Annular Diffuser. International Journal of Applied Engineering Research, 5(20): 2639-2652.

Arora, B.B., 2014. Performance analysis of parallel hub diverging casing axial annular diffuser with 20 equivalent cone angle. Australian Journal of Mechanical Engineering, 12(2): 179-194.

Arora, B.B., Kumar, M. and Maji, S., 2010. Analysis of flow separation in wide angle annular diffusers. International Journal of Applied Engineering Research, 5(20): 3419-3428.

Coladipietro, R.M., 1974. Effect of inlet swirl and wall layer thickness on the performance of equiangular annular diffusers. Electronic Theses and Dissertations.

D.W. Bryer, R.C. Pankhurst, 1971. Pressure-Probe Methods for Determining Wind Speed and Flow Direction, HMSO, London.

Hoadley, D., 1970. Three-dimensional turbulent boundary layers in an annular diffuser (Doctoral dissertation, University of Cambridge).

Ibrahim, K.A., Abdalla, H.A., El-Askary, W.A. and Abdel-Salam, A.M., 2007. Numerical study on turbulent swirling flow in annular diffusers. Eng Res Bull, 30(4): 551-69.

Kline, S.J. 1959. On the nature of stall. Journal of basic Engineering, 81(3): 305-319.

Klomp, E.D., 1997. Performance of straight-walled annular diffusers with swirling flow. The Aeronautical Journal, 101(1010): 467-480.

Mohan, R., Singh, S.N. and Agrawal, D.P., 1998. Optimum inlet swirl for annular diffuser performance using CFD. IJEMS, 05 (1):15-21.

Reneau, L.R., Johnston, J.P. and Kline, S.J., 1967. Performance and design of straight, two-dimensional diffusers. J. Basic Eng., 89(1): 141-150.

Sadasivan, S., Arumugam, S.K. and Aggarwal, M.C., 2019. Numerical Simulation of Diffuser of a Gas Turbine using the Actuator Disc Model. Journal of Applied Fluid Mechanics, 12(1): 77-84.

Shuja, S.Z. and Habib, M.A., 1996. Fluid flow and heat transfer characteristics in axisymmetric annular diffusers. Computers & fluids, 25(2): 133-150.

Singh, H. and Arora, B.B., 2019. Effect of Area Ratio on Flow Separation in Annular Diffuser. In Advances in Fluid and Thermal Engineering, Springer: 297-305.

Singh, S.N., Agrawal, D.P., Sapre, R.N. and Malhotra, R.C., 1994. Effect of inlet swirl on the performance of wide-angled annular diffusers. IJEMS, 1 (2):63-69.

Singh, S.N., Seshadri, V., Saha, K., Vempati, K.K. and Bharani, S., 2006. Effect of inlet swirl on the performance of annular diffusers having the same equivalent cone angle. Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering, 220(2): 129-143.

Ubertini, S. and Desideri, U., 2000. Experimental performance analysis of an annular diffuser with and without struts. Experimental Thermal and Fluid Science, 22(3-4): 183-195.

Estimation and Prediction of Deterministic Human Intent Signal to augment Haptic Glove aided Control of Robotic Hand

Rajesh Kumar*, Pimmy Gandotra†, Brejesh Lall†, Arzad A. Kherani‡, Sudipto Mukherjee*

arXiv:2110.07953v1 [cs.RO] 15 Oct 2021

***Abstract*—The paper focuses on Haptic Glove (HG) based control of a Robotic Hand (RH) executing in-hand manipulation. A control algorithm is presented to allow the RH relocate the object held to a goal pose. The motion signals for both the HG and the RH are high dimensional. The RH kinematics is usually different from the HG kinematics. The variability of kinematics of the two devices, added with the incomplete information about the human hand kinematics result in difficulty in direct mapping of the high dimensional motion signal of the HG to the RH. Hence, a method is proposed to estimate the human intent from the high dimensional HG motion signal and reconstruct the signal at the RH to ensure object relocation. It is also shown that the lag in synthesis of the motion signal of the human hand added with the control latency of the RH leads to a requirement of the prediction of the human intent signal. Then, a recurrent neural network (RNN) is proposed to predict the human intent signal ahead of time.**

***Index Terms*—Haptic Glove, Robotic Hand, Tele-operation, Recurrent Neural Network, Estimation Techniques, High-Dimensional Signal Mapping**

I. Introduction

Tele-operation systems have been focus of research for decades. The initial research (since the 1940s) in robotics was inspired by the probable development of robust tele-operation architectures and have continued to develop since then [1], [2]. Tele-operation in robots have been developed in various domains, including remote surgery [3]–[6], remote diagnosis [7]–[9], space robotics [10], [11]; and handling of hazardous materials [12], [13]. The ultimate goal is to control the robot based on the intention of the human. The tele-operation strategies draw on developments in remote sensing [14] and tele-presence [15].

The robot control can be supervisory [16], [17], or semi-autonomous [18]–[23]. During supervisory and semi-autonomous control of the robot, the human commands the motion of the robot while the robot executes the motion. The tele-operation algorithms have been primarily demonstrated using serial robotic arms [24], mobile robots [25], and Unmanned Aerial Vehicles (UAVs) [26]. The desired motion from the human is mapped to the robotic end. Researchers have demonstrated tracking of the end effector motion (of the robotic arm) [27]–[29], and the combination of human motion and the applied force (impedance-based methods [30]–[34]). The robotic human and robotic motion is usually coupled with the necessity to perform a desired task. Execution of this task invariably requires the robot to interact with the environment [35]. The early modes of supervisory control were limited to control based on audio, visual and force feedback signals of the robotic motion [36], [37]. However, with the advent of haptic based joysticks [38], [39], HGs [40]–[42], different mechanisms of tactile feedback [43]–[46] were also introduced. This allows the humans to get an approximate feel of the robot interacting with the environment. The above-said architectures are usually deployed with serial robotic arms performing the manipulation. In order to manipulate the object or to perform certain tasks, the serial robotic arms are usually appended by grippers that hold and grip the object [9], [47]–[49]. Apart from generation of signals representing the human motion [50] and subsequent control, different Electromyography (EMG) signals [51]–[54], and brain-computer interfacing (BCI) signals [55]–[59] have also been used to control the robot. The research focused on rendering the muscle and the brain signals of the human user by eliminating noise in order to estimate the human intention.

Apart from the development of the haptic interfaces and sensor modules to detect human motion [46], [60]–[62]; there has been a widespread interest in deploying estimation [63]–[65] and prediction [66] algorithms to generate both motion commands and control signals for active robotic control. Predictor based algorithms are useful to mitigate control and transmission delays during the control of the robot. For instance, Munir et al. [67] deployed a Modified Smith Detector algorithm to deploy a tele-operation system. Hauser [68] proposed a Gaussian Mixture Autoregression model (GMA) to predict the user intent signal generated from the 2D motion of a cursor in order to operate a 6 Degrees of Freedom (DoF) robotic arm. Xi et al. [69] proposed a task parameterized Hidden Markov model to encode the signal containing information about the human intent. Model-based predictive methods of the signals containing information about the human (master) intent are popular [70], [71]. Neural Network-based approaches have been utilised to predict the master side (human side) intent [72]–[74]. Neural network-based methods have also been used in reinforcement learning [75], [76] and inverse reinforcement learning methods (including deep imitation learning) [77]–[80]. Methods have also been proposed to predict the force feedback signal from the end of the robot [81]. Methods have also been proposed to ensure stability of the master - slave system amidst delays during tele-operation [82].

Although numerous modules have been deployed for analysis, estimation and prediction of the signals transmitted during tele-operation for pick and place modules with serial robotic arms, in-hand manipulation maneuvers constituting precision

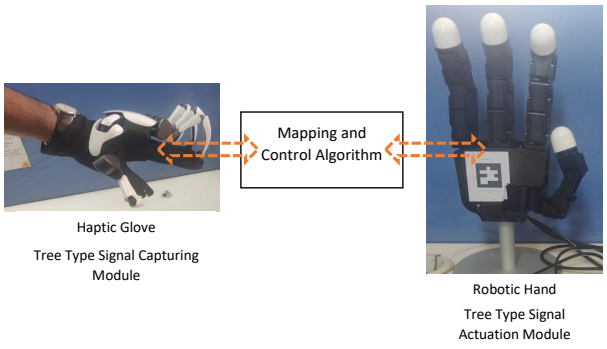


Fig. 1. The mapping and control block needs to be synthesized in the paper.

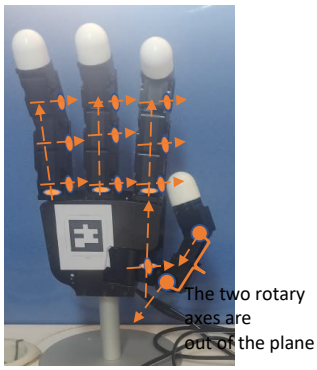


Fig. 2. Allegro RH with 16 actuated joints. The arrows representative of the axes of rotation, orthogonal to the circle.

grasp modes are limited [83]. Precision grasp based in-hand manipulation recruits motion of the fingertips of a RH such that the object moves relative to the palm. Numerous hands focus on standalone in-hand manipulation precision grasp based manipulation behaviour [84], [85]. Although in-hand manipulation has visual similarity to the traditional pick and place operations using serial robots, the former demands precise control of the motion of each of the fingertips and the force applied by the fingertips in order to ensure that the object does not slip from the grip. Difficulty is encountered in mapping the combination of the two measures, force and motion, from the human hand to the RH. This is compounded by the RHs developed having kinematics different from that of the human hand.

For the case of the RHs, multiple fingertips combine to manipulate the object. The motion commands to the joints of the RH lead to a finite motion of the fingertips. Correspondingly, the motion of the fingers of the human hand is captured at intervals via the HG. Although the HG motion signal represents the motion of the joints of the human fingers, the signal usually does not represent the complete object kinematics. As the kinematic architecture of the RH is seldom similar to that of the human hand, there is no one to one mapping between the two signals (the joint motion input to the RH and the joint motion output from the HG). Apart from the kinematic variance of the two devices, the interaction of the two with real world objects is also dissimilar. The two manipulator - object interactions differ in contact stiffness and local friction characteristics, necessitating different interpretations of the two block of signals in discussion. Although, the signal from the HG contains important information of the joint motion, the same cannot be transmitted directly to the RH. Rather, an appropriate encoding - decoding and subsequent control strategy is needed to transform the signal from the human into an interpretable human intent. Then, a subsequent strategy should be deployed to estimate the input signal to the RH in order to complete the desired map. Dex-Pilot [83] presented a strategy to perform in-hand manipulation of the object focusing on estimating a template of motion of the human fingertips and subsequent training for a particular task. It shall also be detailed later in the paper that due to a significant initial acceleration phase in the human motion, the motion of the robot tends to lag the motion of the human user. Hence, a learning algorithm is proposed to generate a human intent signal based on the measurements to mitigate the latency.

The proposed system architecture consists of two subsystems, the RH [86] and the other is the HG (Fig. 1). The human user wears HG, which senses the human motion at 'k' joints such that $k \leq n$, where 'n' is the number of finger joints on the human hand which exactly characterises the motion of all the fingers. It means that the HG does not exactly characterise the motion of the human fingertips, like in the case of Dexmo Haptic Glove (DHG) [87] that senses the motion signal at 11 points on the human hand as opposed to 19 odd joint actuation in the human finger. Also, the HG restricts motion of the human fingers at k_2 joints, such that $k_2 \leq k$ (using a restrictive torque). The other subsystem is the RH consisting of k_3 active joints with p serial chains out of the tree type robotic system [88]. The RH and the HG have different DH parameters (Denavit-Hartenberg) and might be differently oriented (for instance, during the experiment, the Allegro RH [86] (Fig. 2) is right oriented whereas the DHG is left-oriented). The only similarity between the two robotic systems seems to be that the two are tree type robotic systems with serial chains. We aim however to map the intent of the human using the HG and deploy control signals to the RH to achieve the intended manipulation. The contributions of the work can be stated as:

- Synthesis of a control algorithm such that RH achieves the desired goal pose for the object
- Determining the desired goal pose of the object being manipulated by the RH based on the HG joint motion
- Synthesis of a predictor of the goal pose, based on an intent template to leverage control latency

II. Feedback Control Architecture for Sticking Manipulation of the Object

Consider an RH with k_3 actuated joints and p serial chains within the tree (Fig. 3). Each of the serial chain holds needs to interact with the object and perform manipulation such that the goal (target) object pose is reached. The manipulation considered in this paper is a rotation of the object. Additionally, the k_3 actuated joints are torque/ current-controlled (like the Allegro RH). Let $\mathbf{r_i}$ be the vector connecting a suitable point

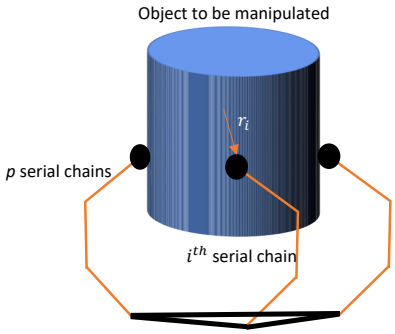


Fig. 3. A tree type robot contacting the object.

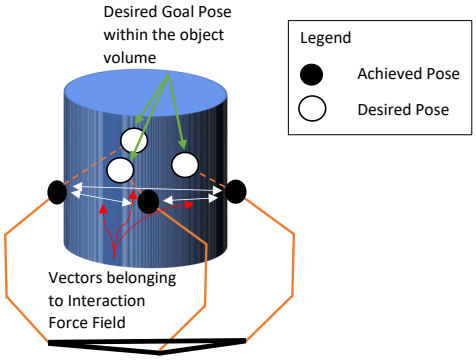


Fig. 4. Due to end effector impedance, the fingertips remain on the object surface with a residual force.

on the object to the i^{th} $(1 \leq i \leq p)$ contact point on the surface of the object. It is known that the equations governing forces that can be applied by the fingertips of the RH form an under-specified set of equations, with null space solutions. [89].

$$\underbrace{\begin{bmatrix} \mathbf{1}_{3\times3} & \mathbf{1}_{3\times3} & \dots & \mathbf{1}_{3\times3} \\ \tilde{\mathbf{r}}_\mathbf{1} & \tilde{\mathbf{r}}_\mathbf{2} & \dots & \tilde{\mathbf{r}}_\mathbf{p} \end{bmatrix}}_{\mathbf{G}} \underbrace{\begin{bmatrix} \mathbf{f_1} \\ \mathbf{f_2} \\ \vdots \\ \mathbf{f_p} \end{bmatrix}}_{\mathbf{F}} = \mathcal{W} \tag{1}$$

where $\tilde{\mathbf{r}}_\mathbf{i}$ refers to the cross product matrix corresponding to the vector $\mathbf{r_i}$, and $\mathbf{1}_{3\times3}$ represents the 3×3 identity matrix. The vector $\mathbf{f_i}$ refers to the force applied by the fingertips on the surface of the object by the i^{th} fingertip. The vector $\mathcal{W} \in \mathbb{R}^6$ refers to the net forces and moments applied to the object by the fingertips [90]. The net object dynamics is represented as:

$$\begin{bmatrix} \mathbf{M_o a_o} \\ \mathbf{I_o}\dot{\omega}_\mathbf{o} + \omega_\mathbf{o} \times \mathbf{I_o}\omega_\mathbf{o} \end{bmatrix} = \mathcal{W} \tag{2}$$

The term $\mathbf{M_o}$ is the mass of the object, $\mathbf{a_o}$ the linear acceleration of the object, $\mathbf{I}_o$ the moment of inertia tensor of the object, and $\omega_\mathbf{o}$ is the angular velocity of the object. It is known that the solution to the grasp map (Eq. (1)) is a combination of two vector fields. The fingertip forces belonging to the equilibrating force field imparts motion to the object. However, the interaction force field generates the null solutions. Though motion is not imparted, assigning suitable null forces leads to improvement of the contact condition.

It is also known that the force vectors belonging to the interaction force field can be spanned using the vectors joining the fingertips. Consider the case shown in Fig. 4, which is a three-point grasp. The general solution to the force field [91], [92] is usually represented as:

$$\mathbf{F} = \mathbf{G}^{+}\mathcal{W} + \sum_{i=1}^{3p-6} \alpha_i \mathbf{NS}(\mathbf{G}) \tag{3}$$

where $p > 2$, $\mathbf{NS}(\mathbf{G})$ represents the null space of the matrix $\mathbf{G}$ (the grasp matrix as per Eq. (1)), $\mathbf{G}^{+}$ represents the Moore-Penrose Pseudo Inverse [93] of the matrix $\mathbf{G}$, and α_i are assignable scalars to the null space matrix. The dynamics of the RH, as with all kinematic chains, with the vector $\mathbf{q}$ as the joint state of the robot is generally represented as:

$$\mathbf{M}\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}}) + \mathbf{g}(\mathbf{q}) = \tau - \mathbf{J}^\mathbf{T}\mathbf{F_{ext}} \tag{4}$$

where $\mathbf{M}$ refers to the mass matrix of the robot, $\mathbf{C}$ refers to the Coriolis term of the dynamics, $\mathbf{g}(\mathbf{q})$ refers to the gravity compensation term of the robot, τ vector refers to the torque applied at the actuated joints. The vector $\mathbf{F_{ext}}$ is the force applied to the end effectors. Derivation for dynamics of tree type robots in the same form is available as well [88]. Consider the robot grasping the object with individual fingertips controlled under impedance control. The torque vector τ is determined as:

$$\begin{aligned} \tau =& \mathbf{K}(\mathbf{q_{des}} - \mathbf{q}) + \mathbf{D}(\dot{\mathbf{q}}_\mathbf{des} - \dot{\mathbf{q}}) + \mathbf{M}\ddot{\mathbf{q}}_\mathbf{des} + \mathbf{C}(\mathbf{q_{des}}, \dot{\mathbf{q}}_\mathbf{des}) \\ &+ \mathbf{g}(\mathbf{q_{des}}) \end{aligned} \tag{5}$$

where $\mathbf{q_{des}}$ is the desired motion of the joints required to achieve the desired position. The matrices $\mathbf{K}$, $\mathbf{D}$ are the gain matrices for robotic control [94]. Another well known result is that when motors with high gear ratio are deployed, the system dynamics (coefficients for time derivatives of the joint states) can be neglected as disturbance. The impedance control hence allows an unconstrained end effector (fingertip) to achieve the set goal pose. However, when the robotic fingertip interacts with the object, the force applied by the fingertip on the object is proportional to the difference in the achieved pose and the desired pose.

As stated earlier, the interaction force components can be assigned without affecting the motion of the object. The goal pose of the fingertips are set to be within the object in order to enhance the contact condition. This is shown schematically in Fig. 4, where the goal pose of the fingertips is within the object volume. However, due to the constraint of the object in between, the fingertips cannot achieve the desired pose. The interaction force field can be viably transmitted using the projection along the line joining the fingertips on the surface of the object.

Once viable contact conditions are achieved, a kinematic resolved rate motion controller is used. It is also known that any rigid body rotation is represented as an element of the

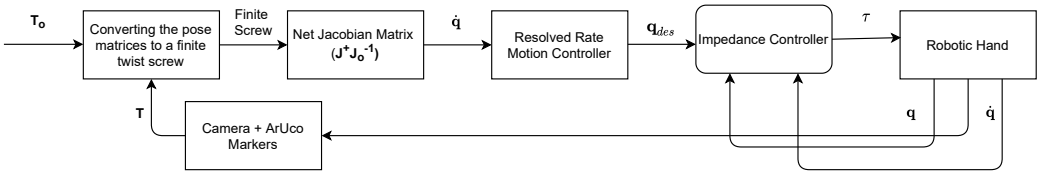


Fig. 5. Representation of the control algorithm to achieve the desired object pose using the RH

$\mathbb{SO}(3)$ manifold [95], and any general rigid body motion is represented as an element of the $\mathbb{SE}(3)$ manifold [96]. The goal pose and the current pose of the object be defined as an element of the $\mathbb{SE}(3)$ manifold. A general pose matrix $\mathbf{T}$ is represented as: $\mathbf{T} \equiv \{\begin{bmatrix} \mathbf{R} & \mathbf{o} \\ \mathbf{0} & 1 \end{bmatrix} || \mathbf{R} \in \mathbb{R}^{3\times3}, \mathbf{o} \in \mathbb{R}^3, \mathbf{R}^T\mathbf{R} = \mathbf{1}_{3\times3},\ det(\mathbf{R}) = 1\}$. The twist screw $\$_o \in \mathbb{R}^6$ represents the twist representation (combination of the linear and the angular velocity) relating the current object pose ($\mathbf{T}$) to the desired object pose ($\mathbf{T_o}$). The details for the determination of the screw $\$_o$ to translate between two object pose matrices can be referred to in the literature [97]. It is known that the least energy geodesic is determined by the screw rotation between the initial and the desired poses [97]. The object level screw rotation trajectory is utilised to transit the object from the initial pose to the desired goal pose. The end effector twist of the i^{th} finger of the RH is quantified as:

$$\$_i = \mathbf{J_i}\dot{\mathbf{q}}_\mathbf{i} \tag{6}$$

where the matrix $\mathbf{J_i}$ is the Jacobian matrix relating the joint rates to the twist of the end effector twist of the i^{th} fingertip ($\$_\mathbf{i}$). The vector $\mathbf{q_i}$ refers to the joint angles of the active joints of the i^{th} tree, whereas as the vector $\mathbf{q} = [\mathbf{q_1^T}, ..., \mathbf{q_p^T}]^\mathbf{T}$. So, the vector containing the twists of all the fingertips for the robot reduce to:

$$\underbrace{\begin{bmatrix} \$_1 \\ \vdots \\ \$_p \end{bmatrix}}_{\$_{RH}} = \underbrace{\begin{bmatrix} \mathbf{J_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{J_2} & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{J_p} \end{bmatrix}}_{\mathbf{J}} \dot{\mathbf{q}} \tag{7}$$

Given the desired twist of the object $\$_o$, the twist of the i^{th} fingertip of the RH is quantified as:

$$\$_\mathbf{i} = \underbrace{\begin{bmatrix} \mathbf{1}_{3\times3} & \mathbf{0}_{3\times3} \\ -\tilde{\mathbf{r}}_\mathbf{i} & \mathbf{1}_{3\times3} \end{bmatrix}}_{\mathbf{J_{oi}}} \$_\mathbf{o} \tag{8}$$

The Eq. (8) can then be extended to p serial chains as:

$$\underbrace{\begin{bmatrix} \$_\mathbf{1} \\ \$_\mathbf{2} \\ \vdots \\ \$_\mathbf{p} \end{bmatrix}}_{\$_\mathbf{RH}} = \underbrace{\begin{bmatrix} \mathbf{J_{o1}} \\ \mathbf{J_{o2}} \\ \vdots \\ \mathbf{J_{op}} \end{bmatrix}}_{\mathbf{J_o}} \$_\mathbf{o} \tag{9}$$

The Eq. (7) and the Eq. (9) can be collated as:

$$\dot{\mathbf{q}} = \mathbf{J}^+\mathbf{J_o^{-1}}\$_o \tag{10}$$

Utilising the standard resolved rate motion approximation, the subsequent incremental goal pose is determined using the control equation:

$$\mathbf{q_{des}} = \mathbf{q} + \eta\dot{\mathbf{q}}\Delta t \tag{11}$$

Following which, the desired joint angle vector presented in Eq. (11) is utilised in Eq. (5) to compute the actuation torques. The net control algorithm is depicted pictorially in Fig. 5. The object twist is derived from the desired object pose matrix and the current object pose matrix (determined from an external camera system using vision techniques like ArUCo Markers [98]). Having established a control algorithm to ensure that the robotic fingertips result in the RH achieving the desired object pose, we focus on the synthesis of the goal object pose from the signals from the human wearing the HG.

III. Synthesis of the Human Intent by Superposition of Serial Chain Kinematics

As noted earlier, the only kinematic similarity between the RH and the HG is both are tree-type robotic system with serial chains. Further, the joint encoders on the HG do not exactly characterise the exact motion of the human fingertips. A heuristic is proposed here for superposition of the kinematics of the serial chains of the HG to the kinematics of the serial chains of the RH. The data flow in relating the input signal on the HG to the motion of the object on the RH is presented schematically in Fig. 6. Consider the mapping of the fingertip of the i^{th} serial chain of RH and the i^{th} serial chain of the HG. The first-order differential kinematics of the i^{th} fingertip of the RH is represented as:

$$\$_\mathbf{i} = \sum_{j=1}^{k_3} \$_{ij}\dot{\theta}_j \tag{12}$$

where $\$_{ij}$ represents the screw rotation of the j^{th} joint of the i^{th} chain. The kinematics of the i^{th} serial chain, including the active joints of the HG on the i^{th} serial chain yields the hypothetical first order kinematics of the hand as:

$$\$'_\mathbf{i} = \sum_{j=1}^{k} \$'_{ij}\dot{\theta}'_j \tag{13}$$

The vector $\$'_{ij}$ represents the equivalent screw motion of the active joint on the HG to the RH. For all the k hypothetical kinematics of the HG on the RH, the estimated object screw motion is characterised as the minimum norm rigid body motion, quantified as:

$$\$'_\mathbf{o} = \mathbf{J_o}^+\$_\mathbf{HG} \tag{14}$$

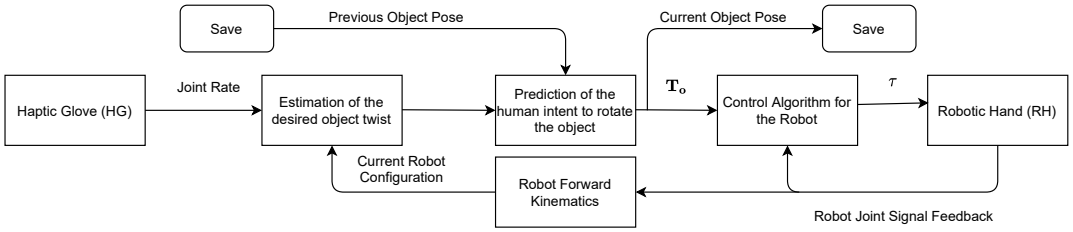


Fig. 6. Representation of the algorithm to achieve the desired object pose using the input from HG

The vector $\$_{\mathbf{HG}}$ is quantified as:

$$\$_{\mathbf{HG}} = \begin{bmatrix} \$'_1 \\ \vdots \\ \$'_p \end{bmatrix} \tag{15}$$

The instantaneous linear velocity and the angular velocity of a point on the object can be derived based on the screw representation as $[\mathbf{v},\ \omega] = \$'_o$. Upto a time t, knowing the hypothetical twist motion, the motion of the object can be synthesised as: $d = \int_0^t \mathbf{v}dt$, and $\Delta\theta = \int_0^t \omega dt$. Subsequently, the goal transformation matrix is repeatedly synthesised as: $\mathbf{T_o} = \begin{bmatrix} \mathbf{R} & \mathbf{d} \\ \mathbf{0} & 1 \end{bmatrix}$, with the matrix $\mathbf{R} = \mathbf{1}_{3\times3} + \mathbf{E}\sin\theta + \mathbf{E^2}(1-\cos\theta)$. The matrix $\mathbf{E}$ is defined as: $\mathbf{E} = \begin{bmatrix} 0 & -\Delta\theta_z & \Delta\theta_y \\ \Delta\theta_z & 0 & -\Delta\theta_x \\ -\Delta\theta_y & \Delta\theta_x & 0 \end{bmatrix}$, where $\mathbf{\Delta}\theta = \begin{bmatrix} \Delta\theta_x \\ \Delta\theta_y \\ \Delta\theta_z \end{bmatrix}$, $\theta = ||\Delta\theta||_2$. In this formulation, the integral effect of the hypothetical twist screw is used to repeatedly determine the human intention to manipulate the object. Often the intent of prehensile manipulation is primarily a rotation, in which case the velocity terms $\mathbf{v}$ are equated to $\mathbf{0}$.

It shall be seen that the human intent coded as per this methodology accumulates the effect of the instantaneous motion of the human finger joints. We introduce the observed knowledge that the human tendency is to follow a 'sigmoid' like curve while manipulating objects. This includes an acceleration phase associated with only a small cumulative rotation (or displacement of the hypothetical/real object), followed by a large displacement period and then a small motion and large deceleration period. At the same time, the RH joints require a finite time to accelerate to follow the goal trajectory. The cumulative effect of the human intent, the goal pose, is evident only after the deceleration phase is initiated. This leads to a lag in the motion of the RH vis-á-vis, the signal captured by HG. The time delay between the feedback at the HG and visual feedback of the RH is counter intuitive to a teleoperator and is usually overcome by significant on the job training.

IV. Prediction of the Goal Pose

To minimise the effect of control latency, a method which generates a prediction in finite but small time is needed. We propose to use the heuristic that the humans deploy prehensile manipulation in repeated tasks. Thereby, a Recurrent Neural Network (RNN), which have traditionally been used for the processing of sequential data for a wide variety of signals [99]–[103] based prediction methodology is proposed.

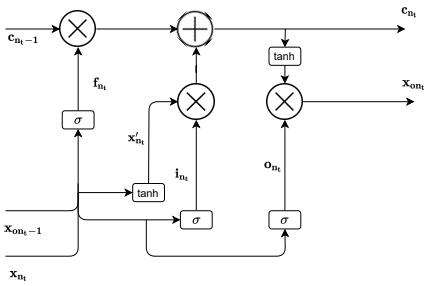


Fig. 7. Representation of an LSTM network node

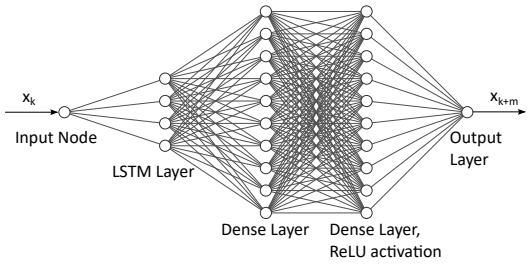


Fig. 8. Representation of the used neural network

The encoded motion of the joints of the human fingers has been reformulated as a sequence of integral solutions in the previous section. The integral solution at any time instance 't' is based on the current estimate and accumulation of the prior trajectory, giving rise to the necessity of sequence learning. Hence, sequence based RNN models are used to predict a futuristic estimate. The memory cell feature of the Long Short Term Memory (LSTM) based networks are useful for this case, as the human intended motion of the object follow three characteristic features within the signal. The LSTM network also retains characteristic components of the signal while massaging the issue of vanishing gradient in traditional RNNs [104]. As shown in the example later, the signal includes a slow start, a fast growth and then a saturation. This feature is repeated in the signal. The network is hence used to train the three features and output is the goal pose, a finite time in the future.

For a rotational motion of the object with respect to a coordinate frame, the vector $\Delta\theta$ represents the net motion of the object. Let us assume that the term $\Delta\theta_x$ at an instance t is labelled as $\Delta\theta_{x_{n_t}}$. Also, consider a sequence consisting of 'r'

elements: $\mathbf{x_{n_t}} = \{\Delta\theta_{x_{n_t-r}}, \Delta\theta_{x_{n_t-r+1}}, \ldots, \Delta\theta_{x_{n_t}}\}$, where n_t corresponds to the element index at time instance t. So, the sequence $\mathbf{x_{n_t}}$ consists of r elements sampled sequentially from the signal upto time t. Similarly, for the other two elements, the sequences $\mathbf{y_{n_t}}$ and $\mathbf{z_{n_t}}$ can be synthesized.

Corresponding to the vector $\mathbf{x_{n_t}}$, a scalar $x_{n_{ot}}$ is to be synthesized such that it is the predicted $(n_t + m_t)^{th}$ element of the sequence, where 'm_t' is suitably selected based on the time of prediction. A Long Short Term Memory (LSTM) [105] based neural network is synthesized such that the network predicts a state given the prior sequence of data. The input to the LSTM network is $\mathbf{x_{n_t}}$. An LSTM network consists of different gates to train a network on the sequence, comprising of the forget gate, input gate, and the output gate is presented in Fig. 7. For each iteration of the network (corresponding to the time index t), a sequence containing r elements is used as an input and the prediction estimate corresponding to the $(n_t + m_t)^{th}$ element of the signal is the output. The LSTM network utilises the following equations to predict the output, given an input sequence:

$$\mathbf{x'_{n_t}} = \tanh(\mathbf{W_1^T}[\mathbf{x_{n_t}^T}, \mathbf{x_{on_t-1}^T}]^\mathbf{T}) \tag{16}$$

$$\mathbf{i_{n_t}} = \sigma(\mathbf{W_2^T}[\mathbf{x_{n_t}^T}, \mathbf{x_{on_t-1}^T}]^\mathbf{T}) \tag{17}$$

$$\mathbf{f_{n_t}} = \sigma(\mathbf{W_3^T}[\mathbf{x_{n_t}^T}, \mathbf{x_{on_t-1}^T}]^\mathbf{T}) \tag{18}$$

$$\mathbf{c_{n_t}} = \mathbf{i_{n_t}} \circledast \mathbf{x'_{n_t}} + \mathbf{f_{n_t}} \circledast \mathbf{c_{n_t-1}} \tag{19}$$

$$\mathbf{o_{n_t}} = \sigma(\mathbf{W_4^T}[\mathbf{x_{n_t}^T}, \mathbf{x_{on_t-1}^T}]^\mathbf{T}) \tag{20}$$

$$\mathbf{x_{on_t}} = \mathbf{i_{n_t}} \circledast \mathbf{x'_{n_t}} + \mathbf{f_{n_t}} \circledast \mathbf{c_{n_t-1}} \tag{21}$$

where $\circledast$ represents element wise multiplication. The matrices $\mathbf{W_i}$ constitute the training parameters and the symbol σ represents the sigmoid function [104]. The vector $\mathbf{f_{n_t}}$ represents the output vector from the forget gate, the vector $\mathbf{i_{n_t}}$ represents the output vector from the input gate, whereas the vector $\mathbf{o_{n_t}}$ represents the output vector from the output gate. The vector $\mathbf{c_{n_t}}$ represents the cell state at time t, indexed as n_t. Subsequently, the vector $\mathbf{x_{on_t}}$ is the output vector from the LSTM block and input to the other subsequent layers. As shown in Fig. 8, two dense layers with 10 nodes are used, one with a 'ReLU' activation [104]. Then a single node output is used to estimate the intent in the sequence after 'm_t' elements. The backpropogation algorithm is used to successively determine the parameters, to minimize the mean squared error loss function [104].

In the example presented later, a vector consisting of 20 elements (sampled over approximately 1.2 seconds) is used as an input to the network in order to estimate a value after 10 elements (0.6 seconds). A system of overlapping windows is used such that the input signal is sequentially sampled for the past 20 elements. This ensures an overlap of 19 elements for subsequent training samples. Although, the overlapping results in redundancy of the input samples, it allows quick capture during the rapid growth of the signal, which is the one of the three main features of the task. As a general principle of prediction, the number of samples to the input vis-á-vis the futuristic estimate should be selected based on the signal variation.

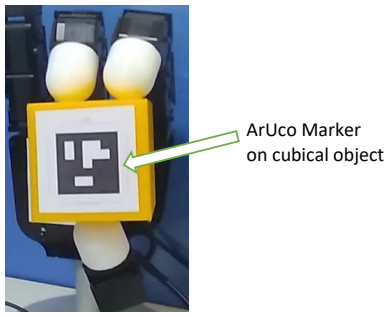


Fig. 9. Allegro Hand grasping a cubical object

TABLE I
MECHANICAL MOTION RANGE (MMR) FOR THE FOUR JOINTS OF THE THUMB AND THE FOUR FINGERS OF THE ARH IN DEGS

Joint No	Thumb MMR (deg)	Fingers MMR (deg)
1	64.9	66.84
2	84.11	116.52
3	116.52	119.40
4	119.302	116.863

V. EXPERIMENTAL ANALYSIS USING ALLEGRO RH AND THE DHG

The Allegro RH (ARH) consists of four kinematic chains (three fingers and a thumb). Each of the kinematic chain has 4 active joints powered by motors with gear ratio of 1:369. At this ratio, the end effector dynamics can be neglected. The impedance controller produces a desired torque at 333 Hz, whereas the resolved rate motion controller updates the desired joint angle value at 20 Hz. Two fingers and a thumb are used for manipulation. The the native resolution of Joint Encoders resolution maps to joint angle of 0.002 degs. However, the signal is transmitted as a 16 bit signal, reducing the resolution to 0.005 degs. The encoders used have a mechanical range of 333.33 degs but individual joints have varying motion ranges arising from hardware intersections as listed in Table I. The torque limits for different joints of the ARH is listed in Table II. Although the desired torque is approached by the control algorithm, the torque capacity of the motor limits the motion performance of joints and hence the robot.

For this paper, the end effector stiffness is constant throughout the manipulation though it could vary with the task. As an embodiment of the control algorithm enumerated in Section II, consider the robot holding a cubical object (as shown in Fig. 9). The true motion of the object is tracked using the ArUco markers pasted on the object. A motion sequence is shown in Fig. 10 using three video snips of change in the object pose. The change of object pose measured about one

TABLE II
LIMITING TORQUE (LT) FOR THE FOUR JOINTS OF THE THUMB AND THE FOUR FINGERS OF THE ARH

Joint No	Thumb LT (Nm)	Fingers LT (Nm)
1	± 0.437	± 0.312
2	± 0.337	± 0.562
3	± 0.225	± 0.375
4	± 0.225	± 0.238

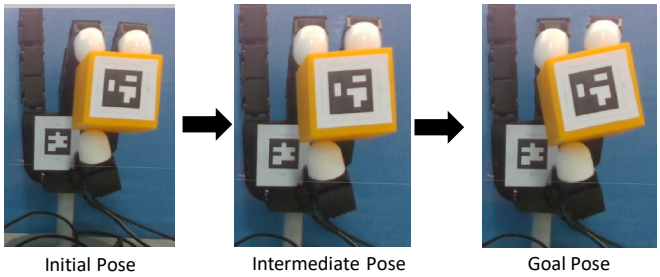


Fig. 10. Motion of the object by the ARH

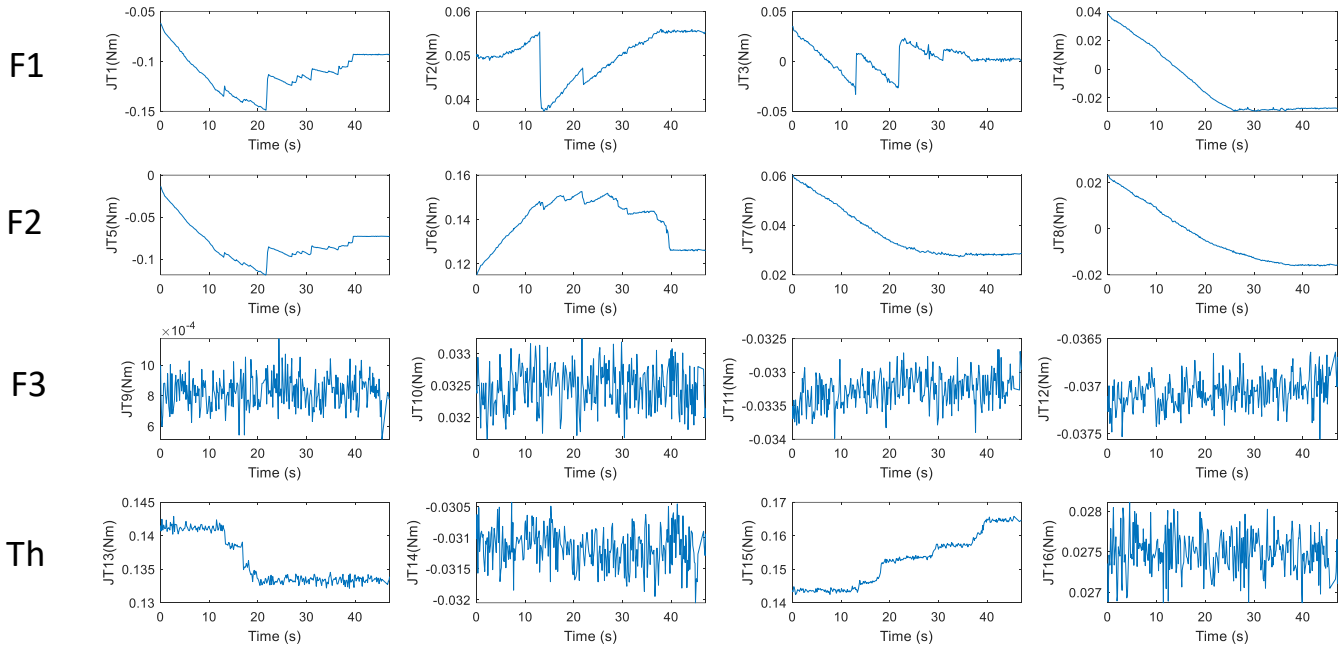


Fig. 11. Variation of the Joint Torque (JT) signal of the 16 actuated joints of the ARH corresponding to the motion in Fig. 10. 'F1' refers to the torque signals of the four joints of the first robotic finger (from right) actuating the object, 'F2' refers to the second finger. 'F3' (third finger) does not hit/ actuate the object, hence the torque signal remains constant. 'Th' refers to the four joints corresponding to the thumb signal

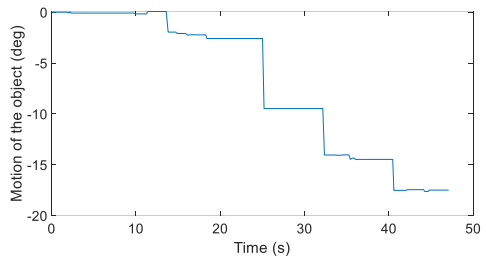


Fig. 12. Motion of the object by the ARH corresponding to Fig. 10

axis is shown in the Fig. 12. The response shows repeated step inputs leading to the desired change in the object pose. The corresponding torque applied by the 16 joints is presented in Fig. 11. The torque applied by the joints for robotic finger 3 (leftmost finger) is constant (with visible signal noise) as the finger does not participate in manipulating the object and remains free. The variation of the required torque for the rest of the 12 signals can be observed.

Also, consider the DHG which is a tree type chain of five serial chains (four fingers) and a thumb. Each of the fingers has two encoding points, whereas the thumb has three encoding points. The fingers have one encoder recording the bending action while the other encoders record the splitting motion. The DHG encodes the data at the rate of 200 Hz, with a spatial resolution of 0.08 deg for different joints. The signals corresponding to the five joints with bending motion is within the mechanical range of 147 deg, whereas the splitting motion is within 30 deg. The stiffness signal varies from (0 – 0.33 kg-cm/deg). The machine level resolution of the motor torque is 0.005 Nm. The peak torque value is limited to 0.5 Nm. In order to analyse the joint signals from the DHG, a Principal Component Analysis (PCA) [106] was performed by manipulating the object for four test cases. The variation of the joint values for a test motion of the DHG manipulating an object is shown in Fig 13. The corresponding magnitudes of the PCA components for different axes of the 11-D data

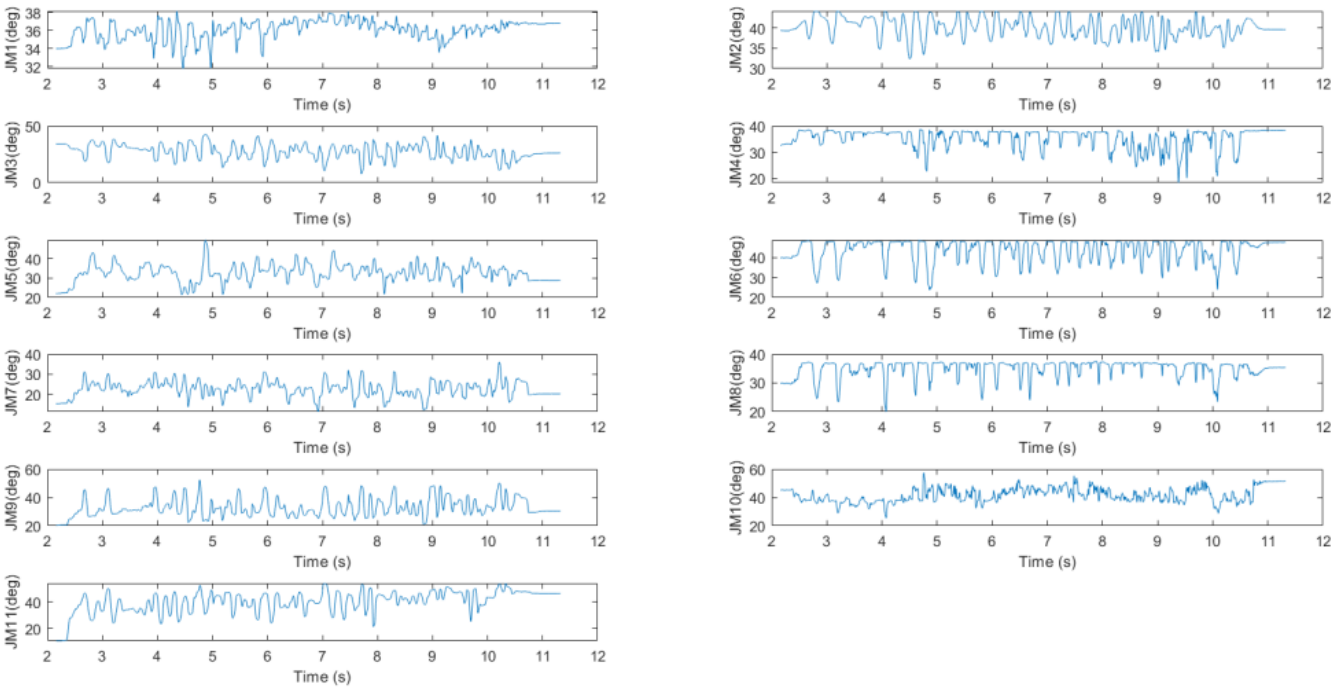


Fig. 13. Variation of the Joint Motion (JM) signal of the 11 joints of the DHG for arbitrary motion with a spherical ball; the joint motion signal for the i^{th} encoder is represented as 'JMi'

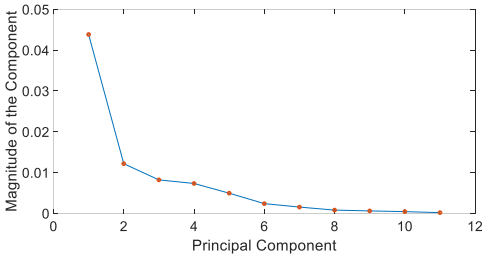


Fig. 14. PCA analysis of the data corresponding to Fig. 13

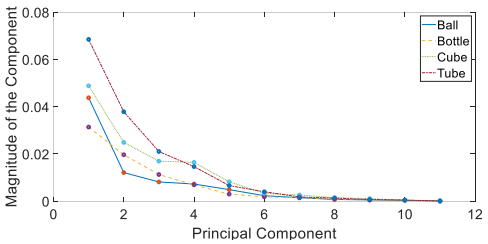


Fig. 15. PCA analysis of the data corresponding for exemplar manipulation of different objects

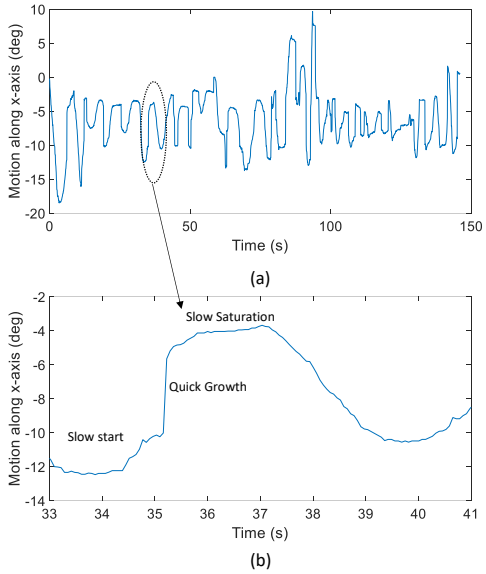


Fig. 16. (a) Exemplar human motion intent of the object (b) Detailed analyses of a single wavelet

is presented in Fig 14. For four different test cases, the PCA magnitudes are presented in Fig. 15. It can be seen that for all the cases, the first three components are dominant, with the next three principal values being of lesser significance. This can be attributed to the fact that the primary motion targetted of the object during in-hand manipulation was a general rotational motion with parasitic translation motion. Beyond six dimensions, the magnitude of the principal components is negligible, which is consistent with the motion being of a rigid body.

A static stiffness value is used to restrict the DHG at a grasp pose so that the fingers can be restricted. Then the motion of the human fingers is estimated and mapped to a

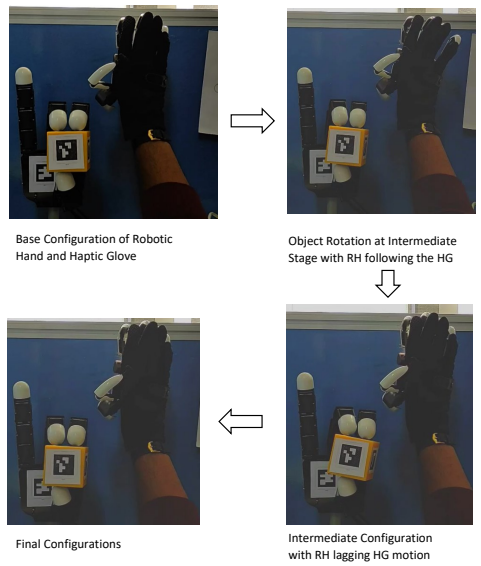


Fig. 17. Intermediate motion of the HG and RH mapping

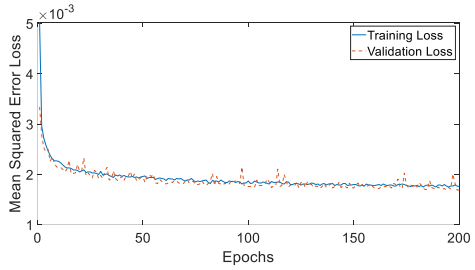


Fig. 18. Mean Squared Error corresponding to the train and the validation data set (rad)

hypothetical motion of the object. The ARH is then controlled to replicate the object motion. The motion of the robotic fingertips followed by capturing the rotational intent in one of the three cartesian directions is presented in Fig. 16. It can be seen that the motion intent of the human to manipulate the object follows an 'sigmoid' shaped curve, consisting of a slow growth followed by a quick surge and then a slow saturation. The human intent is small at the initial stages, thereby leading to lag in the motion of the robot (combined with the slow initial response of the robot). The motion of the RH and the object can be seen in Fig. 17.

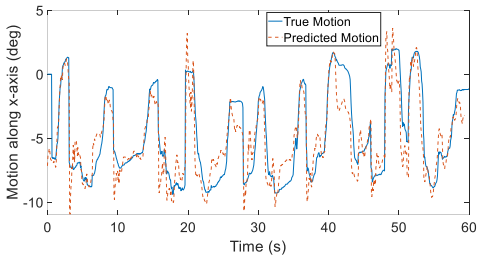


Fig. 19. Comparison of the true data and the predicted data from the algorithm

It can be seen that the motion of the object by the RH lags the motion of the HG. The RNN network trained on the DHG is then introduced, to predict the goal pose recursively. The corresponding variation of mean squared error loss during training and validation is presented in Fig. 18. Both the losses converge to 0.14 deg. The training data was used to project the motion intent 0.6 seconds into the future. For an example manipulation, the comparison of the predicted variation of the estimate and the true data is presented in Fig. 19. It can be seen that the neural network predicts the desired estimate 0.6 seconds earlier, thereby reducing the control and estimation latency.

VI. Conclusion

An introductory development towards HG based control of a kinematically different RH is shown. A stable control algorithm for the RH to relocate the object to a goal pose is presented. Then, a methodology is proposed to deterministically synthesize finger actuation to achieve the desired motion of the object even though the measurement of the hand motion is of a lower dimension. The issue of latency during intent identification and to actuate the robot is enumerated and then managed using a Long Short Term Memory Recurrent Neural Network to predict in the future by leveraging knowledge of time trajectory of typical human actions.

References

[1] P. F. Hokayem and M. W. Spong, "Bilateral teleoperation: An historical survey," *Automatica*, vol. 42, no. 12, pp. 2035–2057, 2006.

[2] M. Shahbazi, S. F. Atashzar, and R. V. Patel, "A systematic review of multilateral teleoperation systems," *IEEE transactions on haptics*, vol. 11, no. 3, pp. 338–356, 2018.

[3] P. Berthet-Rayne, K. Leibrandt, G. Gras, P. Fraisse, A. Crosnier, and G.-Z. Yang, "Inverse kinematics control methods for redundant snakelike robot teleoperation during minimally invasive surgery," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2501–2508, 2018.

[4] Z. Wang, I. Reed, and A. M. Fey, "Toward intuitive teleoperation in surgery: Human-centric evaluation of teleoperation algorithms for robotic needle steering," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 5799–5806.

[5] G. Ateş, R. Majani, and M. I. C. Dede, "Design of a teleoperation scheme with a wearable master for minimally invasive surgery," in *New Trends in Medical and Service Robotics*. Springer, 2019, pp. 45–53.

[6] L. Cheng and M. Tavakoli, "Neural network-based physiological organ motion prediction and robot impedance control for teleoperated beating-heart surgery," *Biomedical Signal Processing and Control*, vol. 66, p. 102423, 2021.

[7] P. A. Patlan-Rosales and A. Krupa, "Robotic assistance for ultrasound elastography providing autonomous palpation with teleoperation and haptic feedback capabilities," in *2020 8th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*. IEEE, 2020, pp. 1018–1023.

[8] M. Al-Badri, S. Ipsen, S. Böttger, and F. Ernst, "Robotic 4d ultrasound solution for real-time visualization and teleoperation," *Current Directions in Biomedical Engineering*, vol. 3, no. 2, pp. 559–561, 2017.

[9] G. Yang, H. Lv, Z. Zhang, L. Yang, J. Deng, S. You, J. Du, and H. Yang, "Keep healthcare workers safe: application of teleoperated robot in isolation ward for covid-19 prevention and control," *Chinese Journal of Mechanical Engineering*, vol. 33, no. 1, pp. 1–4, 2020.

[10] Z. Liu, Z. Lu, Y. Yang, and P. Huang, "Teleoperation for space manipulator based on complex virtual fixtures," *Robotics and Autonomous Systems*, vol. 121, p. 103268, 2019.

[11] Z. Wang, H.-K. Lam, B. Xiao, Z. Chen, B. Liang, and T. Zhang, "Event-triggered prescribed-time fuzzy control for space teleoperation systems subject to multiple constraints and uncertainties," *IEEE Transactions on Fuzzy Systems*, 2020.

[12] J. Aleotti, G. Micconi, S. Caselli, G. Benassi, N. Zambelli, M. Bettelli, D. Calestani, and A. Zappettini, "Haptic teleoperation of uav equipped with gamma-ray spectrometer for detection and identification of radioactive materials in industrial plants," in *Factories of the Future*. Springer, Cham, 2019, pp. 197–214.

[13] J. Aleotti, G. Micconi, S. Caselli, G. Benassi, N. Zambelli, M. Bettelli, and A. Zappettini, "Detection of nuclear sources by uav teleoperation using a visuo-haptic augmented reality interface," *Sensors*, vol. 17, no. 10, p. 2234, 2017.

[14] C. R. Jung and J. Scharcanski, "Sharpening dermatological color images in the wavelet domain," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 1, p. 4, 2009.

[15] D. J. Roberts, A. J. Fairchild, S. P. Campion, J. O'Hare, C. M. Moore, R. Aspin, T. Duckworth, P. Gasparello, and F. Tecchia, "withyou—an experimental end-to-end telepresence system using video-based reconstruction," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 3, pp. 562–574, 2015.

[16] I. Havoutis and S. Calinon, "Supervisory teleoperation with online learning and optimal control," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1534–1540.

[17] H. Latifee, A. Pervez, J.-H. Ryu, and D. Lee, "Mini-batched online incremental learning through supervisory teleoperation with kinesthetic coupling," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 5453–5459.

[18] M. Rubagotti, T. Taunyazov, B. Omarali, and A. Shintemirov, "Semi-autonomous robot teleoperation with obstacle avoidance via model predictive control," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2746–2753, 2019.

[19] I. Havoutis and S. Calinon, "Learning from demonstration for semi-autonomous teleoperation," *Autonomous Robots*, vol. 43, no. 3, pp. 713–726, 2019.

[20] D.-H. Zhai and Y. Xia, "Adaptive control of semi-autonomous teleoperation system with asymmetric time-varying delays and input uncertainties," *IEEE transactions on cybernetics*, vol. 47, no. 11, pp. 3621–3633, 2016.

[21] C. Ha, S. Park, J. Her, I. Jang, Y. Lee, G. R. Cho, H. I. Son, and D. Lee, "Whole-body multi-modal semi-autonomous teleoperation of mobile manipulator systems," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 164–170.

[22] K. Y. Lui, H. Cho, C. Ha, and D. Lee, "First-person view semi-autonomous teleoperation of cooperative wheeled mobile robots with visuo-haptic feedback," *The International Journal of Robotics Research*, vol. 36, no. 5-7, pp. 840–860, 2017.

[23] W. Li, Z. Li, Y. Liu, L. Ding, J. Wang, H. Gao, and Z. Deng, "Semi-autonomous bilateral teleoperation of six-wheeled mobile robot on soft terrains," *Mechanical Systems and Signal Processing*, vol. 133, p. 106234, 2019.

[24] J. Singh, A. R. Srinivasan, G. Neumann, and A. Kucukyilmaz, "Haptic-guided teleoperation of a 7-dof collaborative robot arm with an identical twin master," *IEEE transactions on haptics*, vol. 13, no. 1, pp. 246–252, 2020.

[25] S. Musić, G. Salvietti, F. Chinello, D. Prattichizzo, S. Hirche *et al.*, "Human–robot team interaction through wearable haptics for cooperative manipulation," *IEEE transactions on haptics*, vol. 12, no. 3, pp. 350–362, 2019.

[26] D. Zhang, G. Yang, and R. P. Khurshid, "Haptic teleoperation of uavs through control barrier functions," *IEEE transactions on haptics*, vol. 13, no. 1, pp. 109–115, 2020.

[27] G. Du, P. Zhang, J. Mai, and Z. Li, "Markerless kinect-based hand tracking for robot teleoperation," *International Journal of Advanced Robotic Systems*, vol. 9, no. 2, p. 36, 2012.

[28] J. Kofman, X. Wu, T. J. Luu, and S. Verma, "Teleoperation of a robot manipulator using a vision-based human-robot interface," *IEEE transactions on industrial electronics*, vol. 52, no. 5, pp. 1206–1219, 2005.

[29] D. Rakita, B. Mutlu, and M. Gleicher, "A motion retargeting method for effective mimicry-based teleoperation of robot arms," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017, pp. 361–370.

[30] A. Ajoudani, N. Tsagarakis, and A. Bicchi, "Tele-impedance: Teleoperation with impedance regulation using a body–machine interface," *The International Journal of Robotics Research*, vol. 31, no. 13, pp. 1642–1656, 2012.

[31] K. Hashtrudi-Zaad and S. E. Salcudean, "Analysis of control architectures for teleoperation systems with impedance/admittance master and slave manipulators," *The International Journal of Robotics Research*, vol. 20, no. 6, pp. 419–445, 2001.

[32] Z. Lu, P. Huang, and Z. Liu, "High-gain nonlinear observer-based impedance control for deformable object cooperative teleoperation with nonlinear contact model," *International Journal of Robust and Nonlinear Control*, vol. 30, no. 4, pp. 1329–1350, 2020.

[33] M. Maaref, A. Rezazadeh, K. Shamaei, and M. Tavakoli, "A gaussian mixture framework for co-operative rehabilitation therapy in assistive impedance-based tasks," *IEEE journal of selected topics in signal processing*, vol. 10, no. 5, pp. 904–913, 2016.

[34] A. Haddadi and K. Hashtrudi-Zaad, "Bounded-impedance absolute stability of bilateral teleoperation control systems," *IEEE Transactions on Haptics*, vol. 3, no. 1, pp. 15–27, 2009.

[35] E. Triantafyllidis, C. Mcgreavy, J. Gu, and Z. Li, "Study of multimodal interfaces and the improvements on teleoperation," *IEEE Access*, vol. 8, pp. 78 213–78 227, 2020.

[36] M. Massimino and T. Sheridan, "Variable force and visual feedback effects on teleoperator man/machine performance," in *Proceedings of the NASA Conference on Space Telerobotic*, 1989, pp. 1751–1756.

[37] M. J. Massimino and T. B. Sheridan, "Teleoperator performance with varying force and visual feedback," *Human factors*, vol. 36, no. 1, pp. 145–157, 1994.

[38] D. Senkal and H. Gurocak, "Haptic joystick with hybrid actuator using air muscles and spherical mr-brake," *Mechatronics*, vol. 21, no. 6, pp. 951–960, 2011.

[39] Y.-J. Nam and M.-K. Park, "Virtual excavator simulator featuring hils and haptic joysticks," *Journal of Mechanical Science and Technology*, vol. 29, no. 1, pp. 397–407, 2015.

[40] J. Perret and E. Vander Poorten, "Touching virtual reality: a review of haptic gloves," in *ACTUATOR 2018; 16th International Conference on New Actuators*. VDE, 2018, pp. 1–5.

[41] M. Zhu, Z. Sun, Z. Zhang, Q. Shi, T. He, H. Liu, T. Chen, and C. Lee, "Haptic-feedback smart glove as a creative human-machine interface (hmi) for virtual/augmented reality applications," *Science Advances*, vol. 6, no. 19, p. eaaz8693, 2020.

[42] J. Perret and E. Vander Poorten, "Commercial haptic gloves," in *Proceedings of the 15th Annual EuroVR Conference*. VTT Technology, 2018, pp. 39–48.

[43] T. Murakami, T. Person, C. L. Fernando, and K. Minamizawa, "Altered touch: miniature haptic display with force, thermal and tactile feedback for augmented haptics," in *ACM SIGGRAPH 2017 Posters*, 2017, pp. 1–2.

[44] J. D'Abbraccio, L. Massari, S. Prasanna, L. Baldini, F. Sorgini, G. Airò Farulla, A. Bulletti, M. Mazzoni, L. Capineri, A. Menciassi *et al.*, "Haptic glove and platform with gestural control for neuromorphic tactile sensory feedback in medical telepresence," *Sensors*, vol. 19, no. 3, p. 641, 2019.

[45] T. Bakker, J. Verlinden, D. Abbink, and R. van Deventer, "Development of a haptic device with tactile and proprioceptive feedback for spatial design tasks," in *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*. IEEE, 2017, pp. 223–228.

[46] R. Chaudhari, C. Schuwerk, M. Danaei, and E. Steinbach, "Perceptual and bitrate-scalable coding of haptic surface texture signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 3, pp. 462–473, 2014.

[47] C. Mizera, T. Delrieu, V. Weistroffer, C. Andriot, A. Decatoire, and J.-P. Gazeau, "Evaluation of hand-tracking systems in teleoperation and virtual dexterous manipulation," *IEEE Sensors Journal*, vol. 20, no. 3, pp. 1642–1655, 2019.

[48] G. S. Gupta, S. C. Mukhopadhyay, C. H. Messom, and S. N. Demidenko, "Master–slave control of a teleoperated anthropomorphic robotic arm with gripping force sensing," *IEEE Transactions on Instrumentation and Measurement*, vol. 55, no. 6, pp. 2136–2145, 2006.

[49] G. Zhu, X. Xiao, C. Li, J. Ma, G. Ponraj, A. Prituja, and H. Ren, "A bimanual robotic teleoperation architecture with anthropomorphic hybrid grippers for unstructured manipulation tasks," *Applied Sciences*, vol. 10, no. 6, p. 2086, 2020.

[50] M. Ye, C. Yang, V. Stankovic, L. Stankovic, and A. Kerr, "A depth camera motion analysis framework for tele-rehabilitation: Motion capture and person-centric kinematics analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 5, pp. 877–887, 2016.

[51] P. K. Artemiadis and K. J. Kyriakopoulos, "An emg-based robot control scheme robust to time-varying emg signal features," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 3, pp. 582–588, 2010.

[52] P. Artemiadis, "Emg-based robot control interfaces: past, present and future," *Advances in Robotics & Automation*, vol. 1, no. 2, pp. 1–3, 2012.

[53] K. Gui, H. Liu, and D. Zhang, "A practical and adaptive method to achieve emg-based torque estimation for a robotic exoskeleton," *IEEE/ASME Transactions on Mechatronics*, vol. 24, no. 2, pp. 483–494, 2019.

[54] T. Teramae, T. Noda, and J. Morimoto, "Emg-based model predictive control for physical human–robot interaction: Application for assist-as-needed control," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 210–217, 2017.

[55] B. Xu, W. Li, X. He, Z. Wei, D. Zhang, C. Wu, and A. Song, "Motor imagery based continuous teleoperation robot control with tactile feedback," *Electronics*, vol. 9, no. 1, p. 174, 2020.

[56] S. Zhao, Z. Li, R. Cui, Y. Kang, F. Sun, and R. Song, "Brain–machine interfacing-based teleoperation of multiple coordinated mobile robots," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 6, pp. 5161–5170, 2016.

[57] S. Qiu, Z. Li, W. He, L. Zhang, C. Yang, and C.-Y. Su, "Brain–machine interface and visual compressive sensing-based teleoperation control of an exoskeleton robot," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 1, pp. 58–69, 2016.

[58] Y. Liu, W. Su, Z. Li, G. Shi, X. Chu, Y. Kang, and W. Shang, "Motor-imagery-based teleoperation of a dual-arm robot performing manipulation tasks," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 11, no. 3, pp. 414–424, 2018.

[59] Y. Liu, M. Habibnezhad, and H. Jebelli, "Brain-computer interface for hands-free teleoperation of construction robots," *Automation in Construction*, vol. 123, p. 103523, 2021.

[60] B. Fang, F. Sun, H. Liu, and D. Guo, "A novel data glove using inertial and magnetic sensors for motion capture and robotic arm-hand teleoperation," *Industrial Robot: An International Journal*, 2017.

[61] Q. He, Y. Wu, Z. Feng, W. Fan, Z. Lin, C. Sun, Z. Zhou, K. Meng, W. Wu, and J. Yang, "An all-textile triboelectric sensor for wearable teleoperated human–machine interaction," *Journal of Materials Chemistry A*, vol. 7, no. 47, pp. 26 804–26 811, 2019.

[62] G. Cheung, D. Florencio, P. Le Callet, C.-W. Lin, and E. Magli, "Introduction to the issue on interactive media processing for immersive communication," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 3, pp. 381–383, 2015.

[63] C. P. Vo, X. D. To, and K. K. Ahn, "A novel force sensorless reflecting control for bilateral haptic teleoperation system," *IEEE Access*, vol. 8, pp. 96 515–96 527, 2020.

[64] W. Qi, X. Liu, L. Zhang, L. Wu, W. Zang, and H. Su, "Adaptive sensor fusion labeling framework for hand pose recognition in robot teleoperation," *Assembly Automation*, 2021.

[65] A. Dong, Z. Du, and Z. Yan, "A sensorless interaction forces estimator for bilateral teleoperation system based on online sparse gaussian process regression," *Mechanism and Machine Theory*, vol. 143, p. 103620, 2020.

[66] R. Uddin and J. Ryu, "Predictive control approaches for bilateral teleoperation," *Annual Reviews in Control*, vol. 42, pp. 82–99, 2016.

[67] S. Munir and W. J. Book, "Internet based teleoperation using wave variables with prediction," in *2001 IEEE/ASME International Conference on Advanced Intelligent Mechatronics. Proceedings (Cat. No. 01TH8556)*, vol. 1. IEEE, 2001, pp. 43–50.

[68] K. Hauser, "Recognition, prediction, and planning for assisted teleoperation of freeform tasks," *Autonomous Robots*, vol. 35, no. 4, pp. 241–254, 2013.

[69] B. Xi, S. Wang, X. Ye, Y. Cai, T. Lu, and R. Wang, "A robotic shared control teleoperation method based on learning from demonstrations," *International Journal of Advanced Robotic Systems*, vol. 16, no. 4, p. 1729881419857428, 2019.

[70] C. Smith and P. Jensfelt, "A predictor for operator input for time-delayed teleoperation," *Mechatronics*, vol. 20, no. 7, pp. 778–786, 2010.

[71] K. Hosseini-Suny, H. Momeni, and F. Janabi-Sharifi, "Model reference adaptive control design for a teleoperation system with output prediction," *Journal of Intelligent & Robotic Systems*, vol. 59, no. 3, pp. 319–339, 2010.

[72] M. Nikpour, B. Yazdankhoo, B. Beigzadeh, and A. Meghdari, "Adaptive online prediction of operator position in teleoperation with unknown time-varying delay: simulation and experiments," *Neural Computing and Applications*, pp. 1–18, 2020.

[73] J.-Q. Huang and F. L. Lewis, "Neural-network predictive control for nonlinear dynamic systems with time-delay," *IEEE Transactions on Neural Networks*, vol. 14, no. 2, pp. 377–389, 2003.

[74] H. Wang, P. X. Liu, and S. Liu, "Adaptive neural synchronization control for bilateral teleoperation systems with time delay and backlash-like hysteresis," *IEEE transactions on cybernetics*, vol. 47, no. 10, pp. 3018–3026, 2017.

[75] M. Ewerton, G. Maeda, D. Koert, Z. Kolev, M. Takahashi, and J. Peters, "Reinforcement learning of trajectory distributions: Applications in assisted teleoperation and motion planning," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 4294–4300.

[76] K. Kamali, I. A. Bonev, and C. Desrosiers, "Real-time motion planning for robotic teleoperation using dynamic-goal deep reinforcement learning," in *2020 17th Conference on Computer and Robot Vision (CRV)*. IEEE, 2020, pp. 182–189.

[77] A. Agrawal, "Automating endoscopic camera motion for teleoperated minimally invasive surgery using inverse reinforcement learning," Ph.D. dissertation, Master's thesis, Worcester Polytechnic Institute, 2018.

[78] D. R. Scobee, V. R. Royo, C. J. Tomlin, and S. S. Sastry, "Haptic assistance via inverse reinforcement learning," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2018, pp. 1510–1517.

[79] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel, "Deep imitation learning for complex manipulation tasks from virtual reality teleoperation," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 5628–5635.

[80] D. J. Brown and M. J. Proulx, "Audio–vision substitution for blind individuals: Addressing human information processing capacity limitations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 5, pp. 924–931, 2016.

[81] O. Dena, T. Faghanimakrani, and A. H. Aghvami, "Gbdt-based modules for force prediction in a model-mediated teleoperation system," in *2020 27th International Conference on Telecommunications (ICT)*. IEEE, 2020, pp. 1–6.

[82] A. Haddadi and K. Hashtrudi-Zaad, "Robust stability of teleoperation systems with time delay: A new approach," *IEEE transactions on haptics*, vol. 6, no. 2, pp. 229–241, 2012.

[83] A. Handa, K. Van Wyk, W. Yang, J. Liang, Y.-W. Chao, Q. Wan, S. Birchfield, N. Ratliff, and D. Fox, "Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 9164–9170.

[84] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray *et al.*, "Learning dexterous in-hand manipulation," *The International Journal of Robotics Research*, vol. 39, no. 1, pp. 3–20, 2020.

[85] S. Cruciani, B. Sundaralingam, K. Hang, V. Kumar, T. Hermans, and D. Kragic, "Benchmarking in-hand manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 588–595, 2020.

[86] M. Li, H. Yin, K. Tahara, and A. Billard, "Learning object-level impedance control for robust grasping and dexterous manipulation," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 6784–6791.

[87] X. Gu, Y. Zhang, W. Sun, Y. Bian, D. Zhou, and P. O. Kristensson, "Dexmo: An inexpensive and lightweight mechanical exoskeleton for motion capture and force feedback in vr," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 1991–1995.

[88] S. V. Shah, S. K. Saha, and J. K. Dutt, "Dynamics of tree-type robotic systems," in *Dynamics of Tree-Type Robotic Systems*. Springer, 2013, pp. 73–88.

[89] V. R. Kumar and K. J. Waldron, "Force distribution in closed kinematic chains," *IEEE Journal on Robotics and Automation*, vol. 4, no. 6, pp. 657–664, 1988.

[90] J. K. Davidson, K. H. Hunt, and G. R. Pennock, "Robots and screw theory: applications of kinematics and statics to robotics," *J. Mech. Des.*, vol. 126, no. 4, pp. 763–764, 2004.

[91] Y. Zhang, F. Gao, and W. A. Gruver, "Determination of contact forces in grasping," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS'96*, vol. 3. IEEE, 1996, pp. 1038–1044.

[92] R. Kumar and S. Mukherjee, "Enhanced dynamic capability of cable-driven parallel manipulators by reconfiguration," *Robotica*, pp. 1–19, 2021.

[93] J. C. A. Barata and M. S. Hussein, "The moore–penrose pseudoinverse: A tutorial review of the theory," *Brazilian Journal of Physics*, vol. 42, no. 1-2, pp. 146–165, 2012.

[94] B. Siciliano and O. Khatib, *Springer handbook of robotics*. springer, 2016.

[95] J. Mäkinen, "Rotation manifold so (3) and its tangential vectors," *Computational Mechanics*, vol. 42, no. 6, pp. 907–919, 2008.

[96] J.-L. Blanco, "A tutorial on se (3) transformation parameterizations and on-manifold optimization," *University of Malaga, Tech. Rep*, vol. 3, p. 6, 2010.

[97] M. Zefran, V. Kumar, and C. B. Croke, "On the generation of smooth three-dimensional rigid body motions," *IEEE Transactions on Robotics and Automation*, vol. 14, no. 4, pp. 576–589, 1998.

[98] M. F. Sani and G. Karimian, "Automatic navigation and landing of an indoor ar. drone quadrotor using aruco marker and inertial sensors," in *2017 international conference on computer and drone applications (IConDA)*. IEEE, 2017, pp. 102–107.

[99] L. Ruiz, F. Gama, and A. Ribeiro, "Gated graph recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 68, pp. 6303–6318, 2020.

[100] Y. She, Y. He, and D. Wu, "Learning topology and dynamics of large recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 62, no. 22, pp. 5881–5891, 2014.

[101] Y. Xia, "A compact cooperative recurrent neural network for computing general constrained l_1 norm estimators," *IEEE transactions on signal processing*, vol. 57, no. 9, pp. 3693–3697, 2009.

[102] L. Zhang, G. Wang, and G. B. Giannakis, "Real-time power system state estimation and forecasting via deep unrolled neural networks," *IEEE Transactions on Signal Processing*, vol. 67, no. 15, pp. 4069–4077, 2019.

[103] N. Farsad and A. Goldsmith, "Neural network detection of data sequences in communication systems," *IEEE Transactions on Signal Processing*, vol. 66, no. 21, pp. 5663–5678, 2018.

[104] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.

[105] N. M. Vural, S. Ergut, and S. S. Kozat, "An efficient and effective second-order training algorithm for lstm-based adaptive learning," *IEEE Transactions on Signal Processing*, pp. 1–1, 2021.

[106] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.



Rajesh Kumar is a PhD from Indian Institute of Technology Delhi. His research focuses on Robot Kinematics, Control and AI based methods in order to improve robot autonomy. He is a B.Tech in Mechanical Engineering from Indian Institute of Technology Delhi, India.



Pimmy Gandotra received the B.E. degree in Electronics and Communication Engineering from Jammu University, Jammu and Kashmir, India, in 2013 and the M.Tech degree in Electronics and Communication Engineering from Shri Mata Vaishno Devi University in 2016. She pursued her Ph. D degree in Electronics and Communication Engineering at Shri Mata Vaishno Devi University, Katra, Jammu and Kashmir, India. Her research interest includes the emerging technologies of 5G wireless communication network. Currently she is pursuing her post-doctoral research at Indian Institute of Technology as a Principal Project Scientist. She has received teaching assistantship from MHRD from 2014-2016. She is a student member of Institute of Electrical and Electronics Engineers (IEEE).



Brejesh Lall received the B.E. and M.E. degrees in electronics and communications engineering from Delhi College of Engineering, Delhi University, in 1991 and 1992, respectively, and the Ph.D. degree in the area of multi-rate signal processing from the Indian Institute of Technology, Delhi, in 1999. His research interests are in signal and image analysis, computer vision, multimedia systems, multi-scale modelling of stochastic processing wide-sense, cyclostationary process representation physical layer in wireless communication. He has over 50 papers in international journals and refereed conferences. From September 1997 to June 2005, he worked at Hughes Software Systems, in the digital signal processing group. Since July 2005, he has been with the faculty of the Department of Electrical Engineering, Indian Institute of Technology, Delhi, where he is currently a Professor.



Arzad A. Kherani Dr. Arzad Alam Kherani is currently Associate Professor (Department of Electrical Engineering and Computer Science) at IIT, Bhilai. He has worked on broad areas of Computer networks, Queueing systems, Wireless communications. He has also worked for industry for a decade and gained experience with product conceptualization, product design and development, industrial R&D and standardization. He did Master's in 1999 and PhD in 2003 from IISc Bangolore.



Sudipto Mukherjee is a Professor in Department of Mechanical Engineering, Indian Institute of Technology Delhi, India. He holds a PhD and a M.Tech degree from Ohio University, USA and a B.Tech degree from Indian Institute of Technology Kanpur, India.

Computers and Electrical Engineering 96 (2021) 107552

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Exploiting feature space using overlapping windows for improving biometric recognition☆

Surinder Kaur [a,b], Gopal Chaudhary [b], Smriti Srivastava [c], Manju Khari [d], Ruben Gonzalez Crespo [e,*], Javalkar Dinesh Kumar [a]

[a] Lingaya's Vidyapeeth, Haryana, India
[b] Bharati Vidyapeeth's College of Engineering, New Delhi, India
[c] Netaji Subhas University of Technology, Delhi, India
[d] Jawaharlal Nehru University, New Delhi, India
[e] Universidad internacional de La Rioja, Spain

ARTICLE INFO

Keywords:
Biometrics
Windowing
Difference feature
ROI extraction
Overlapping
Score level fusion

ABSTRACT

Biometrics is a highly researched topic due to its importance in security, surveillance, and authentication systems. Granulation is the procedure of partitioning data into windows. Two novel feature extraction techniques using overlapped granules based on the texture features of palm-prints are exploited here. The first category of attribute involves overlapping features that increase the information content by increasing granulation, resulting in additional features. The second attribute is differential information feature (DIF) which involves the first derivative of intensity representing feature dynamics. To further improve the performance, score level fusion is applied. The improvement in the identification values of 2 to 4 percentage points in general and up to 6 percentage points in some cases is seen. In most cases, score level fusion has shown better performance than unimodal methods. The ROC curves showed the superiority of the proposed method over other existing methods.

1. Introduction

Biometrics is a highly researched topic due to its importance in security, surveillance, and authentication systems [1–7]. Besides other biometric modalities, Palm-print-based authentication systems are widely discovered by researchers, and several types of features are investigated owing to their high accuracy and other abilities. Feature extraction techniques for the palm-print are categorized as- Line-based approaches [8,9], subspace-based approaches [10,11] and statistical methods [12–15].

In these approaches, initially, images are transformed into another domain using Fourier transforms, Wavelets, Gabor transforms, etc. Then the transformed images are divided into non-overlapping windows, and local statistics like mean, variance, moments, and centers of gravity are calculated [16]. Then, these statistics are used as features. In human recognition, multimedia components such as palm-print image, speech, signature text, or gait video are the data sources. Granulation is the procedure of partitioning or dividing data into windows or frames. However, if the samples are image or text, then it is windowing, whereas if samples are speech or video, then it is framing. After granulation, they are referred to as the information sources [17].

In the general feature extraction method, the image is first pre-processed and enhanced using an image enhancement process. Rather than analyzing the entirety of the pre-processed image at once, it is divided into square matrices of the same sizes as generally

☆ This paper is for special section VSI-hcac. Reviews processed and recommended for publication by Guest Editor Dr. Vicente García Díaz.

* Corresponding author.
E-mail address: ruben.gonzalez@unir.net (R.G. Crespo).

https://doi.org/10.1016/j.compeleceng.2021.107552
Received 22 January 2021; Received in revised form 10 March 2021; Accepted 8 October 2021
Available online 25 October 2021
0045-7906/© 2021 Elsevier Ltd. All rights reserved.

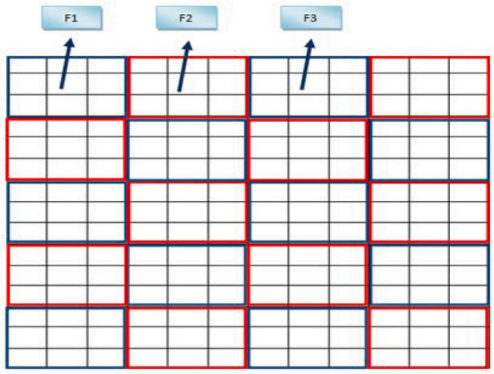


Fig. 1. Non-overlapping windowing method.

non-overlapping windows, granulation, and then functions like Mean, Standard Deviation Gaussian Mean are applied to get the feature vector. The non-overlapping windowing method is depicted in Fig. 1.

1.1. Motivation

All these image-based biometrics consist of limited information in the form of a feature space that may be corrupted by noise and scars. In these limited spaces, features are calculated based on the intensity information that is present. The degree of variability in features increases if intensity sets vary. Where intensity sets are the vectors of information present in a window, information and features are highly dependent on intensity sets; if windows are increased, intensity sets will also increase. To increase the number of features, the number of windows may be increased, while this will lead windows to be smaller in size. This will also reduce the size of the intensity vector, and a smaller window will de-correlate the data. This concludes that non-overlapping windows limit the feature size.

To increase the feature length, the following methods may be used.

1. Overlapping windows may be used, which will increase feature set length without reducing the window's size.
2. Differential information may be extracted, as there is a lot of information present in the difference of intensities between the windows that may forecast feature dynamics.
3. Feature concatenation may be done where we may apply a feature level fusion.
4. Score level fusion may be applied. These rules may be studied in the literature categorized as conventional, T-norm and T-conorm based rules [18–21].

In this study, intensities sets are to be increased using overlapping widows to increase the feature space. For comparison, the mean feature, the absolute average deviation (AAD) feature, and the Gaussian membership function feature (GMF) are considered. In the overlapping features, 20%, 40%, 50%, and 80% overlapping rates of the window are taken into account. The second novel feature is the differential information feature (DIF) which involves important difference features in the palm-print.

1.2. Contributions of the paper

1. In this work, texture analysis of the palm-print is evaluated with two novel features extracted from statistical features based on the overlapped granules.
2. The first category of features involves overlapping features, which increase the information content due to increased granulation. Hence, increasing the total number of features. In overlapping window features (OWF), 20% to 80% overlapping of the window takes place. The second novel feature is the differential information feature (DIF) which involves the first derivative of intensity representing feature dynamics.
3. The mathematical formula for calculating the number of features in overlapping windows for images is derived.
4. For comparison, the mean feature, the absolute average deviation (AAD) feature, and the Gaussian membership function feature (GMF) are considered. All these results are verified using the k-Nearest Neighbor Classifier (kNN).
5. Then, the score level fusion schemes are compared.

The rest of the paper is as follows. Section 2 presents the methodology. Section 2.6 deliberates the score level fusion. Section 3 describes Databases for palm-print used. Section 4 demonstrates the simulations and the analysis of the result. Lastly, Section 5 concludes the proposed study.

2. Methodology

Here, the methods of feature extraction, information retrieval from the windows, gray levels intensities, corresponding membership function, and increasing feature size methods are explained.

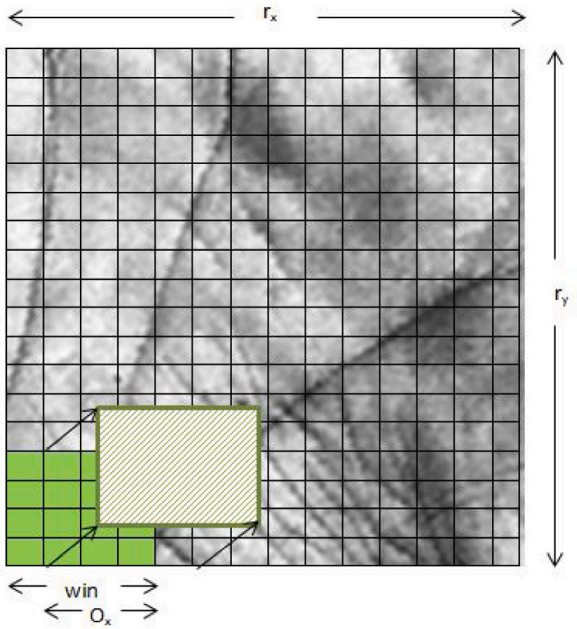


Fig. 2. Overlapping windowing method from a palm-print.

2.1. Feature extraction

To calculate the texture feature, an image is partitioned into different windows. Each window consists of a set of gray-level intensities. Using these intensities, a single feature is extracted from each window. After pre-processing, the enhanced images of size 150 × 150 are granulated into different windows of size 10 × 10 each. Considering a set of gray level intensities, I = $\{I_{ij}\}$ in a window, and $\{\mu_{ij}\}$ is the corresponding membership function which shows the association of gray levels with the set. The importance of this method is that the gray level intensities I = $\{I_{ij}\}$ and the corresponding membership function values $\{\mu_{ij}\}$ are used by a single entropy function as a single entity.

Then the Gaussian membership function-based features are calculated from the information sets of each window. The formula for the GMF-based feature extraction to extract the GMF feature a_{ij} from (i, j) the window is expressed in Eqs. (1) and (2),

$$\mu_{ij}^{g} = e^{-\left[\frac{I_{ij}-\bar{I_{ij}}}{\sqrt{2}\sigma}\right]^2} \tag{1}$$

$$a_{ij} = \frac{1}{m \times n} \sum_{i=0}^{m} \sum_{j=0}^{n} I_{ij} \mu_{ij}^{g} \tag{2}$$

where, I_{ij} is intensity at (i, j), $\bar{I}_{ij}$ is mean and σ is the standard deviation of the intensities in the window, μ_{ij}^{g} is the membership function and a_{ij} is the feature vector from the (i, j) window. It is seen that a_{ij} is dependent on $\sum_{i=0}^{m} \sum_{j=0}^{n} I_{ij} \mu_{ij}^{g}$. Hence, the information present in the texture pattern may be evaluated using a single entropy function which is a function of $\{I_{ij} \mu_{ij}\}$.

Along with GMF, the statistical mean and average absolute deviation (AAD) features are also extracted. The AAD features are calculated as:

$$aa_{ij} = \frac{1}{m \times n} \sum_{i=0}^{m} \sum_{j=0}^{n} (I_{ij} - \bar{I_{ij}}) \tag{3}$$

where $m \times n$ is the image size, $\bar{I_{ij}}$ is the mean and I_{ij} is the intensity at point (i, j).

2.2. Overlapping windows

To increase the feature space, the overlapping of the window is achieved in place of non-overlapping windows as shown in Fig. 2. It is another feature extraction technique using advanced windowing, which reconsiders a small portion of the intensity set in calculating the new feature value.

The number of features that are extracted may be derived by the formula:

$$N = \begin{cases} \frac{(r_x - win + \frac{(100-O_x)}{win}) \times win}{(100-O_x)} \times \frac{r_y}{win}, & O_x > 0 \\ \frac{r_x}{win} \times \frac{r_y}{win}, & O_x = 0 \end{cases} \tag{4}$$

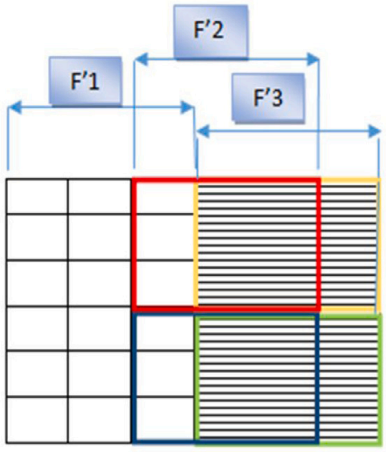


Fig. 3. This depicts the features calculated from each window.

where $r_x \times r_y$ is the size of ROI, O_x is the overlapping percentage in the x direction, $win \times win$ is the size of the window (generally, square windows are taken), and N is the total number of features that are extracted. Overlapping in the y direction is not considered here. For example, if $r_x \times r_y$ is 150 × 150, win is 10 × 10. The number of features that are extracted will be 225 if no overlapping is considered, and if the percentage of overlapping is 50%, then the feature size will be 435. Similarly, for 40% and 80% percent overlapping, the number of features extracted are 365 and 1065, respectively.

This has many advantages as the overlapping window function minimizes spectral leakage. On the other hand, the signal at window boundaries is almost neglected. Additionally, overlapping helps minimize the poorer time-domain resolution, which requires a longer window. It increases the sample space, which further leads to higher information content. No information loss takes place in case of overlapping.

2.3. Concatenation feature

In this case, the two evaluated features are concatenated which in broader sense is a form of feature level fusion. This way we can combine two different piece of intensity information together. For example, $OWF_{0\%}$ and 50% overlapping $OWF_{50\%}$ gives $OWF_{0\%} + OWF_{50\%} = ConC_{50\%}$ To combine the features directly, $f^1_{K\times m}$, and $f^2_{K\times n}$ are concatenated where k, m, and n are the feature dimensions and for $k = 1$, then F_{fused} is the feature given as:

$$F_{fused} = f_{K\times(m+n)} \tag{5}$$

The overlapped feature vector is appended to the non-overlapped feature vector thereby increasing the sample space.

2.4. Differential information feature

There is a lot of information stored in feature dynamics. Hence, to calculate the dynamic feature or differential information feature (DIF), the formula can be shown in Eq. (6).

$$F' = \frac{\sum_{n=1}^{255} n(F_{(n+1)} - F_n)}{\sum_{n=1}^{255} n^2} \tag{6}$$

where F is the dynamic feature value, the value of n varies from 1 to the number of features that the vector values present. The differential information feature vector enables us to extract the hidden differential information vector from the database without expanding the sample space, i.e., vector size, ensuring even lesser loss of pixel information. This is also helpful in reducing the noise in the database.

The outline of proposed work is shown in Fig. 5. It is shown that biometric samples are first pre-processed using proposed overlapping methods, then, features are calculated and then fusion is performed. Then finally, performance is measured.

2.5. Feature percentage increase (%)

This section shows the total increase in the number of features with different methods. The non-overlapping windowing method is depicted in Fig. 1. Fig. 3 shows the method of the features calculated from each window.

Fig. 4 shows the feature vectors calculated from different windowing schemes. Fig. 4a shows the sample feature vector $OWF_{0\%}$ using non-overlapping windowing. Initially, the image was of size 150 × 150. Each 10 × 10 window gives a single feature value. Assuming the feature values of F_1, F_2, ….., F_N for different feature extraction methods, a feature vector of size 225 is obtained.

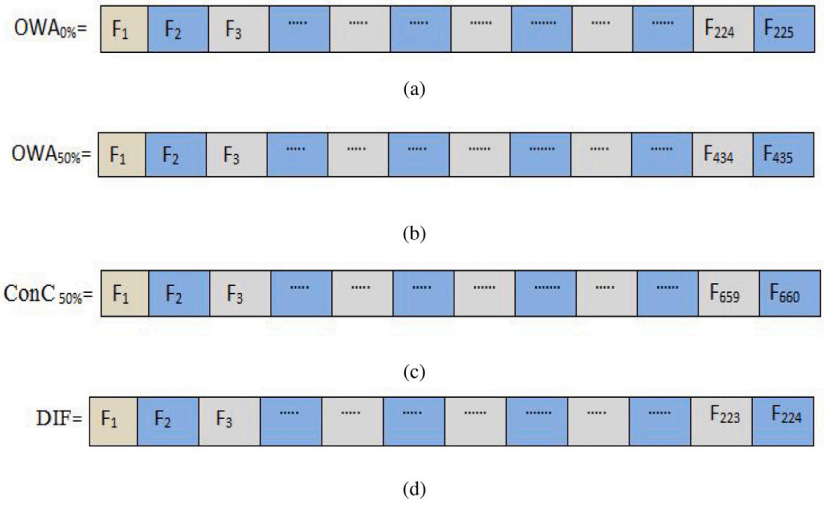


Fig. 4. Different windowing schemes (a) Sample feature vector using Non-overlapping windowing (b) Sample feature vector with 50% overlapping (c) Concatenation feature vector with 50% overlapping and 0% overlapping (d) Differential information feature.

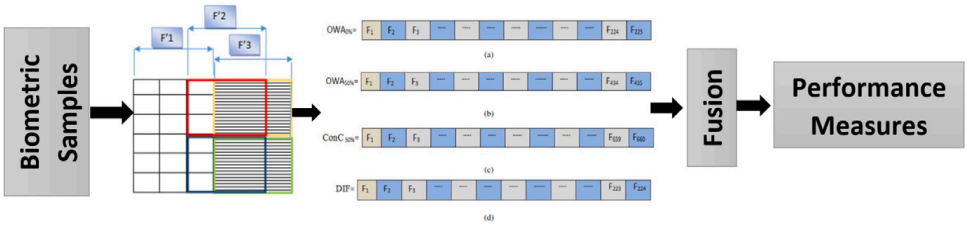


Fig. 5. Overlapping windowing method from a palm-print.

Table 1
Feature percentage increase (%).

Technique	Feature size	$F_{increase}(\%)$
$OWF_{0\%}$	225	–
$OWF_{20\%}$	277	23.11
$OWF_{40\%}$	365	62.22
$OWF_{50\%}$	435	93.33
$OWF_{80\%}$	1065	373.7
$ConC_{50\%}$	660	193.3
DIF	224	−0.44

In overlapping widowing, 20%, 40%, 50%, and 80% overlapping of the window takes place. The length of the feature vector with overlapping of 50% $OWF_{50\%}$ is 1×435. The feature vector $OWF_{50\%}$ that was obtained is shown in Fig. 4b. It increased the sample space, which further led to higher accuracy. The length of the feature vector with concatenation of $OWF_{0\%}$ and 50% overlapping was $1 \times (225 + 435) = 660$ which is shown in Fig. 4c. The DIF is shown in Fig. 4d. The size of the DIF was reduced by 1.

Table 1 shows the percentage increase in the number of features $F_{increase}(\%)$ calculated using Eq. (7).

$$F_{increase}(\%) = \frac{F_{extracted} - OWF_{0\%}}{OWF_{0\%}} \times 100 \tag{7}$$

where the reference feature size is $OWF_{0\%}$, $F_{extracted}$ is a new feature that is extracted. For example, $OWF_{O_x\%}$ where O_x can be 20, 40, 50, and 80.

2.6. Score level fusion

In our experiments, a separate score matrix was calculated on test palm-print images for each feature ($OWF_{50\%}, ConC_{50\%}$ and DIF) separately. In this work, we used *Min Max score Normalization*, which is evaluated as $\delta_{k'} = \frac{\delta_k - min}{max - min}$. This is an important step to scale all scores in the same range.

2.6.1. Score fusion rules

In this work, we used only t norm-based fusion rules. Assuming δ_i to be the property score obtained from the i_{th} feature and δ is the score obtained after score level fusion and N is the each feature $OWF_{50\%}, ConC_{50\%}$ and DIF.

Table 2
Results of identification values (%) for GMF, AAD and Mean for different features extraction method using IITD and PolyU database.

Feature	IITD database			PolyU database		
	AAD	Mean	GMF	AAD	Mean	GMF
$OWF_{0\%}$	89.8	89.1	91.2	96.8	96.1	97.2
$OWF_{20\%}$	90.2	89.6	91.8	96.8	96.3	97.4
$OWF_{40\%}$	90.8	90.6	92.6	98.1	97.6	98.8
$OWF_{80\%}$	90.6	90.5	92.6	96.9	96.3	97.2
$OWF_{50\%}$	93.2	93	94.8	98.8	98.6	99.2
DIF	92.1	92.9	94.1	98.4	97.9	98.9
$ConC_{50\%}$	93.9	93.6	95.4	99.1	99.1	99.6

1. Hamacher t-norm : $\delta = \frac{\delta_1\delta_2}{\delta_1+\delta_2-\delta_1\delta_2}$.
   If $\delta_1, \delta_2, \delta_3$ are the scores obtained from the $OWF_{50\%}, ConC_{50\%}$ and DIF, then, the final score δ will be $T-norm(\delta_3, T-norm(\delta_1, \delta_2))$.
2. Frank t-norm : $\delta = \log_s(\frac{1+(s^{\delta_1}-1)(s^{\delta_2}-1)}{s-1})$, where $s > 0$.
3. Yager t-norm : $\delta = Max(1-((1-\delta_1)^s + (1-\delta_2)^s), 0)$, where $s > 0$.
4. Schweizer-Sklar t-norm : $\delta = (Max(\delta_1^s + \delta_2^s - 1, 0))$, where $s > 0$.
5. Dubois-Prade t-norm : $\delta = (\frac{\delta_1 * \Delta_2 * \ldots * \delta_N}{(Max(\delta_1, \delta_2, \ldots, \delta_N), s)})$, where $s \in [0, 1]$.

3. Databases for palm-print

In this work, the Hong Kong Polytechnic University (PolyU) database and the Indian Institute of Technology Delhi (IITD) database were used to validate the proposed method.

The IIT Delhi palm-print database version 1.0 contains left-and right-hand anterior samples of approximately 230 persons in the age group 14–56 years. This database consists of 5 to 6 samples of each hand. So, from each group, 200 people are selected, having six samples for the experiments. Database acquisition was based on a contact-less type of scanning system with a digital CMOS camera. This type of system is effortless and highly useful in the office environment. There is no use of pegs for the placement of the hand.

The PolyU palm-print database contains anterior samples of approximately 386 persons with overall 7752 gray-scale samples. Every person had 20 samples collected in two sessions with a difference of approximately 60 days. In this system, pegs were used to position the hand sample to remove position variance. From the database, only ten samples of each individual were selected to perform the training and testing.

Both the databases are different concerning the placement of the hand. Different databases are selected to show that the system is environment-independent and works well with two databases. It is also independent of the position of the hand. As in the PolyU database, pegs are used; hence data samples are highly static while the sample's position is variable in the IIT database as they do not have pegs.

4. Results and discussion

For validation, the experiments were performed in both verification and identification modes where verification is the process of validating an individual from its own previous enrolled samples, while identification is the validation of a person from a database pool.

Receiver operating characteristic curves were adopted for verification. Euclidean distance was calculated from the training and test samples. The recognition rate was used to evaluate the performance of the proposed method in identification. The recognition rate is the fraction of the test samples which are correctly recognized by the identification system. The dimensions of ROI were kept as 150 × 150 in the experiments. After pre-processing, the size of the windows was kept 15 × 15 at each case.

The results of identification values (%) of GMF, AAD and Mean for different features extraction methods using IITD and PolyU database are tabulated in Table 2. It is clear that overlapping of windows enhances the performance of the system in the case of both databases. From the overlapping features, $OWF_{50\%}$ showed the better performance, followed by $OWF_{40\%}$. $OWF_{80\%}$ and $OWF_{20\%}$ showed similar results due to same ratio of overlapping in-spite of their different lengths of feature. $ConC_{50\%}$ is a method of the simplest feature level fusion, and hence the performance of the system was increased on fusing $OWF_{0\%}$ and $OWF_{50\%}$, which was also justified from the results. It showed the 99.6% and 95.4% of identification results for the PolyU and IITD databases, respectively. It may be concluded that the identification values obtained by using different feature value methods such as mean, absolute average deviation, and Gaussian membership function showed that both overlapping (with and without concatenation) and differentiation improved the identification values by 2 to 4 percentage points in general and up to 6 percentage points in some cases.

In Fig. 6, the ROC of (a) mean features (b) average absolute deviation (AAD), and (c) gaussian membership features (GMF) of Overlapping, Differential and Concatenation windows using the IITD database are shown. It is seen from the ROC in Fig. 6 that area under the curve (AUC) is the maximum in the case of $ConC_{50\%}$ concatenation window followed by $OWF_{50\%}$ and reaches 100%

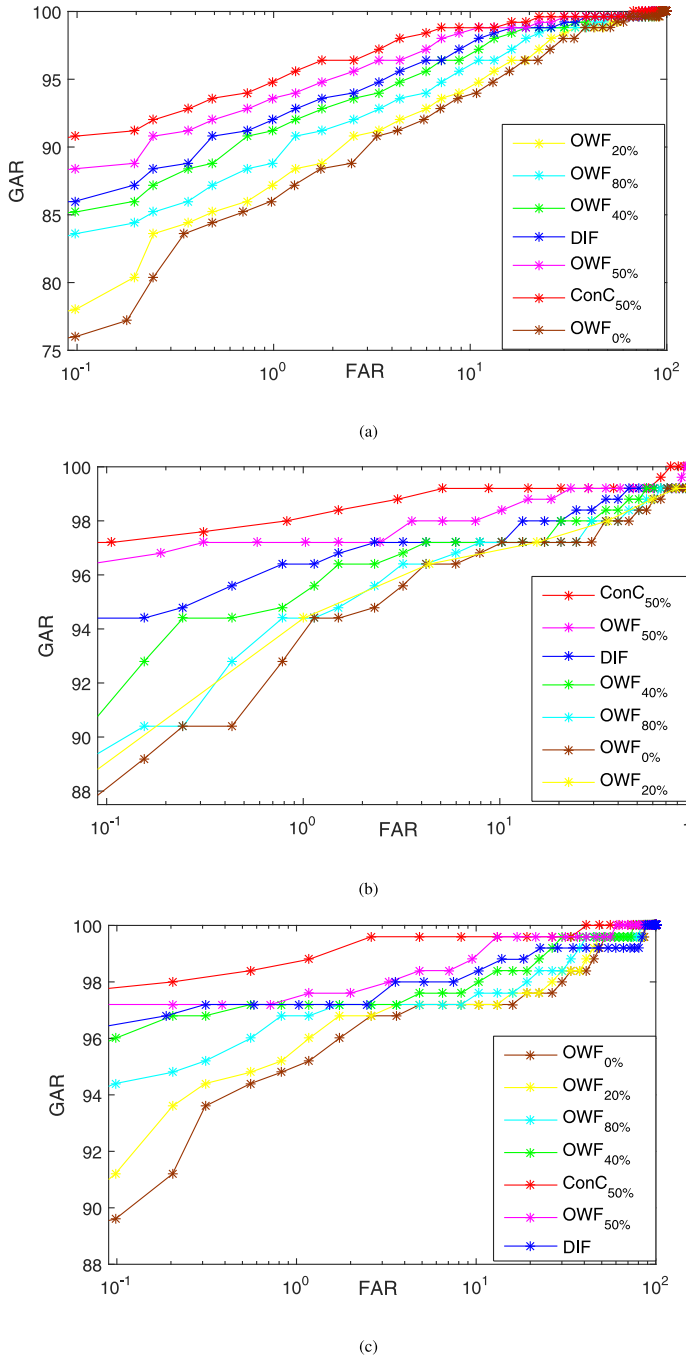


Fig. 6. ROC of (a) mean (b) average absolute deviation (AAD) and (c) gaussian membership features (GMF) of Overlapping, Differential and Concatenation windows using IITD database.

GAR at a faster rate. DIF also showed good performance over $OWF_{40\%}$ and $OWF_{80\%}$. This is a usable residual piece of information that is stored in the differential intensity information. The Receiver Operating Characteristics that were obtained using the PolyU database are shown in Fig. 7.

4.1. Score level fusion results

To further enhance the performance of the system, score level fusion techniques were applied. Because all the features were extracted from a single modality, hence, these techniques were inter-modal score level fusion schemes. We evaluated three major features, $OWF_{50\%}$, $ConC_{50\%}$ and DIF. To compare the score level fusion, four combinations were made $OWF_{50\%}$ -DIF, $OWF_{50\%}$ -$ConC_{50\%}$, DIF-$ConC_{50\%}$ and $OWF_{50\%}$ -$ConC_{50\%}$-DIF ways. The identification results of the score level fusion of $OWF_{50\%}$ -DIF, $OWF_{50\%}$ -$ConC_{50\%}$, DIF-$ConC_{50\%}$ and $OWF_{50\%}$ -$ConC_{50\%}$-DIF strategies on IITD database and PolyU database are tabulated in Tables 3 and 4, respectively.

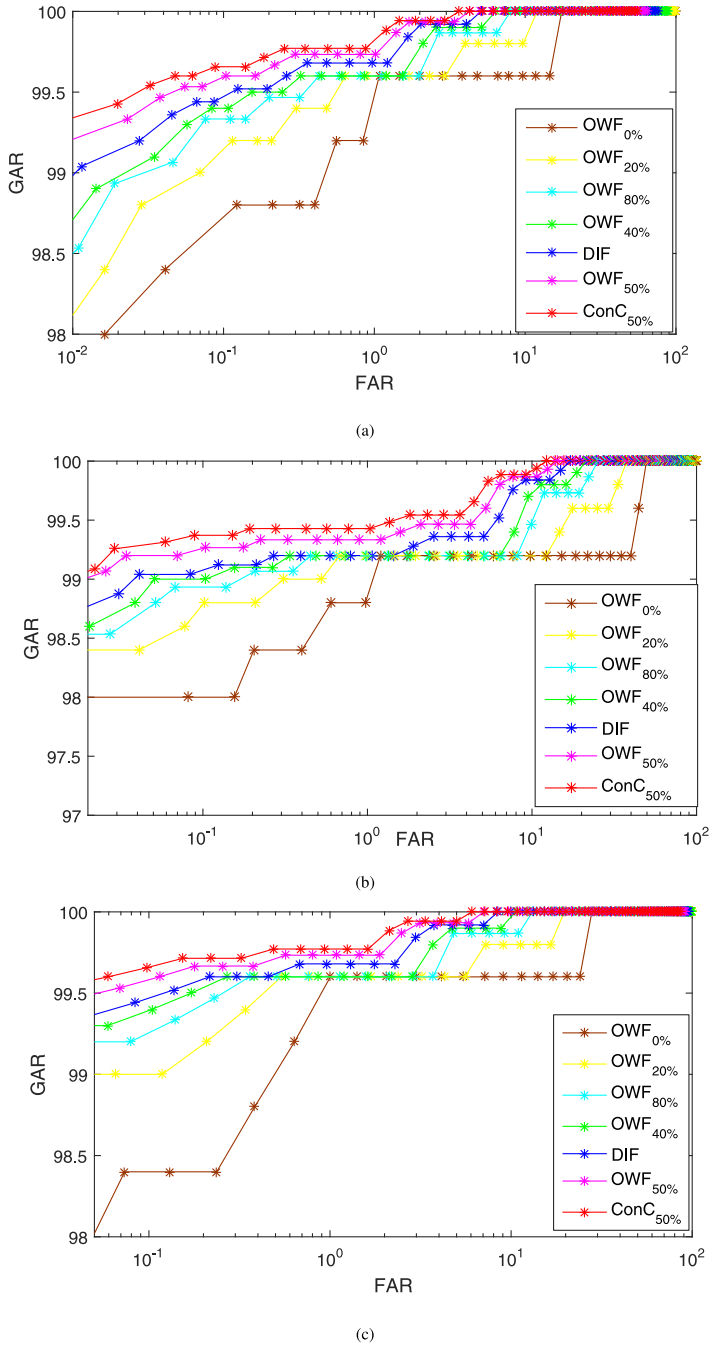


Fig. 7. ROC of (a) mean (b) average absolute deviation (AAD) and (c) gaussian membership features (GMF) of Overlapping, Differential and Concatenation windows using PolyU database.

Table 3
Score level fusion of $OWF_{50\%}$-DIF, $OWF_{50\%}$-$ConC_{50\%}$, DIF-$ConC_{50\%}$ and $OWF_{50\%}$-$ConC_{50\%}$-DIF strategies using IITD database.

Rules	$OWF_{50\%}$-DIF		$OWF_{50\%}$-$ConC_{50\%}$		DIF-$ConC_{50\%}$		$OWF_{50\%}$-$ConC_{50\%}$-DIF	
False Acceptance Rate (FAR %)	0.1	1	0.1	1	0.1	1	0.1	1
Yager OWA	97.48	98.8	94.04	96.96	98.65	99.1	99.2	99.7
Hamacher T-norm	95.82	98.8	93.46	97.2	98.8	99.2	99.2	100
Schweizer-Sklar	98.71	99.2	97.82	98.94	99.2	98.87	100	100
Frank T-norm	98.41	99.2	98.8	98.8	99.6	99.6	100	100
Dubois-Prade	99.6	99.6	99.2	99.2	99.7	100	100	100

In most of the cases, the Frank T-norm and Dubois-Prade have shown better performance than the other three. For $OWF_{50\%} - ConC_{50\%} - DIF$ using IITD database, it was clear that with Yager's OWA, at 0.1 FAR, GAR was 99.2%, for the Hamacher T-norm, it

Table 4
Score level fusion of $OWF_{50\%}$-DIF, $OWF_{50\%}$-$ConC_{50\%}$, DIF-$ConC_{50\%}$ and $OWF_{50\%}$-$ConC_{50\%}$-DIF strategies using PolyU database.

Rules	$OWF_{50\%}$-DIF		$OWF_{50\%}$-$ConC_{50\%}$		DIF-$ConC_{50\%}$		$OWF_{50\%}$-$ConC_{50\%}$-DIF	
False Acceptance Rate (FAR %)	0.1	1	0.1	1	0.1	1	0.1	1
Yager OWA	99.64	100	99.6	100	100	100	100	100
Hamacher T-norm	98.8	100	98.33	99.66	99.88	100	100	100
Schweizer-Sklar	99.9	100	98.89	99.6	100	100	100	100
Frank T-norm	100	100	99.42	99.6	100	100	100	100
Dubois-Prade	99.6	99.6	99.6	99.6	99.65	100	100	100

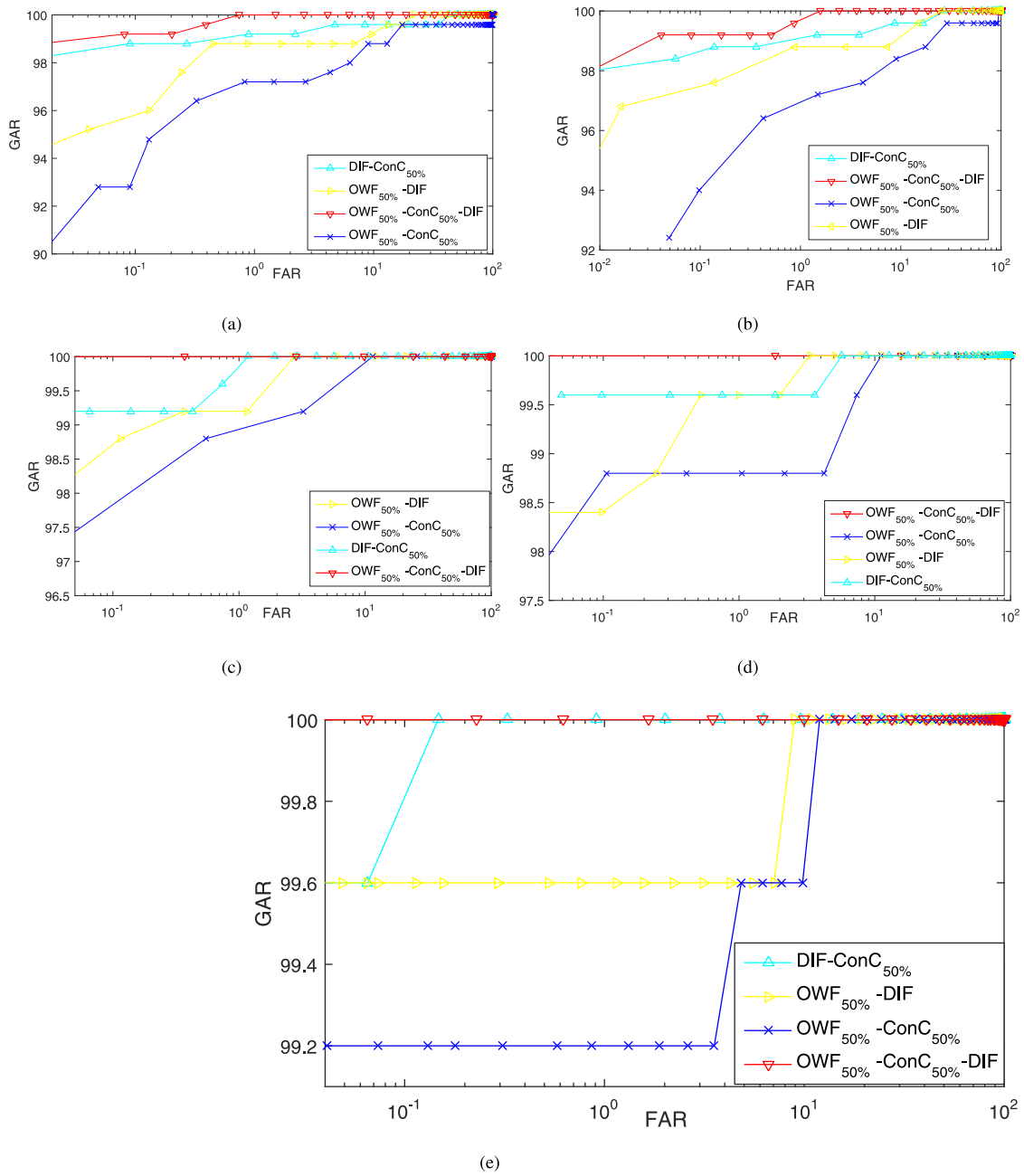


Fig. 8. ROC using different T-norm methods over IITD database (a) Hamacher T-norm (b) Yager OWA T-norm (c) Schweizer-Sklar T-norm (d) Frank T-norm (e) Dubois-Prade T-norm.

was 99.2% and for FAR = 1, GAR was 100% for both the cases. For Frank T-norm, Schweizer-Sklar and Dubois-Prade, GAR reached to 100% at FAR< 0.001. Using the PolyU database, the Frank T-norm outperformed the other methods.

The ROC curves of the score level fusions with different fusion rules using the IITD database are shown in Fig. 8. While for the PolyU database, the ROC curves are shown in Fig. 9. The performance of the system is also shown in terms of area under the

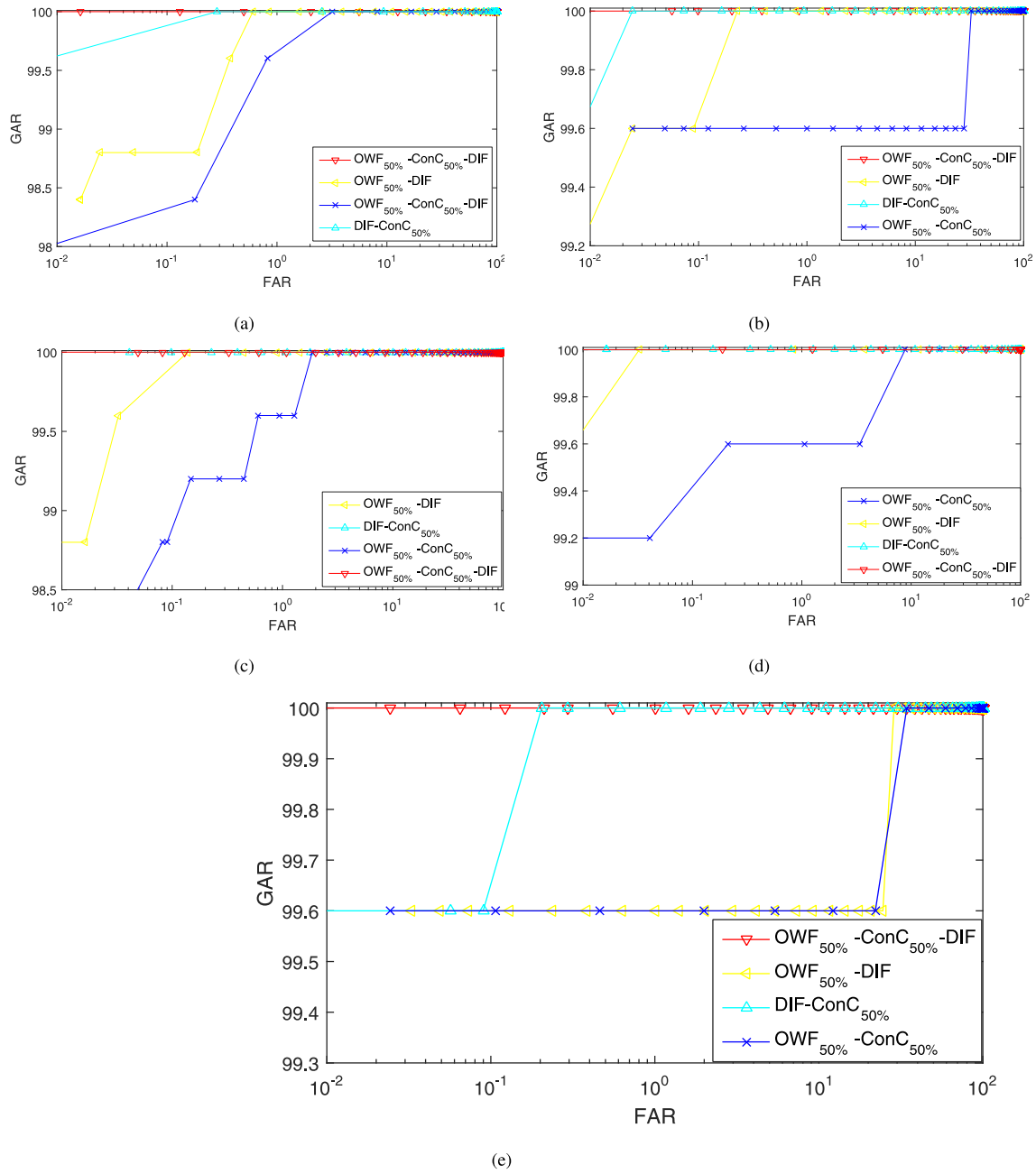


Fig. 9. ROC using different T-norm methods over PolyU database (a) Hamacher T-norm (b) Yager OWA T-norm (c) Schweizer-Sklar T-norm (d) Frank T-norm (e) Dubois-Prade T-norm.

Table 5
AUC of $OWF_{50\%}$-DIF, $OWF_{50\%}$-$ConC_{50\%}$, DIF-$ConC_{50\%}$ and $OWF_{50\%}$-$ConC_{50\%}$-DIF strategies using IITD database.

Rules	$OWF_{50\%}$-DIF	$OWF_{50\%}$-$ConC_{50\%}$	DIF-$ConC_{50\%}$	$OWF_{50\%}$-$ConC_{50\%}$-DIF
Yager OWA	0.9738	0.9720	0.9882	0.9986
Hamacher T-norm	0.9726	0.9710	0.9890	0.9996
Schweizer-Sklar	0.9786	0.9760	0.9920	1.0000
Frank T-norm	0.9818	0.9810	0.9952	1.0000
Dubois-Prade	0.9796	0.9780	0.9960	1.0000

curve (AUC) of ROC in Tables 5 and 6 for the IITD database and PolyU database, respectively. The range of AUC is 0–1. But for satisfactory performance of the system, it should be greater than 0.5 and must reach 1. It is clear that the proposed result covers the maximum AUC and reaches 100% GAR at a faster rate as compared to ROC's of other technique.

Table 6
AUC of $OWF_{50\%}$-DIF, $OWF_{50\%}$-$ConC_{50\%}$, DIF-$ConC_{50\%}$ and $OWF_{50\%}$-$ConC_{50\%}$-DIF strategies using PolyU database.

Rules	$OWF_{50\%}$-DIF	$OWF_{50\%}$-$ConC_{50\%}$	DIF-$ConC_{50\%}$	$OWF_{50\%}$-$ConC_{50\%}$-DIF
Yager OWA	0.9820	0.9790	0.9956	1.0000
Hamacher T-norm	0.9820	0.9780	0.9933	1.0000
Schweizer-Sklar	0.9880	0.9860	1.0000	1.0000
Frank T-norm	0.9910	0.9910	1.0000	1.0000
Dubois-Prade	0.9890	0.9880	0.9966	1.0000

5. Conclusion

In this work, texture analysis of palm-print was evaluated to exploit the feature space for improving the biometric recognition with two novel features extracted from their statistical features. For comparison, the mean feature, the absolute average deviation (AAD) feature, and the Gaussian membership function feature (GMF) were considered. To increase the feature dimension, firstly, the overlapping features were proposed. In the overlapping features, 20%, 40%, 50%, and 80% overlapping of the window took place. The second novel feature was the differential information feature (DIF) which involved important difference features contained in the feature dynamics. The concatenation features were also evaluated. AUC was the maximum in the case of $ConC_{50\%}$ concatenation window followed by $OWF_{50\%}$ and reached 100% GAR at a faster rate. In most of the score level fusion cases, the Frank T-norm and Dubois-Prade have shown better performance than the other three. $OWF_{50\%} - ConC_{50\%} - DIF$ was superior to the other owing to the three levels of score level fusion. The ROC curves showed the superiority of the proposed method over other existing methods. However, this has many advantages as the overlapping window function minimizes spectral leakage. On the other hand, the signal at window boundaries is almost neglected. Additionally, overlapping helps minimize the poorer time-domain resolution, which requires a longer window. It increases the sample space, which further leads to higher information content. In the future, image enhancement methods can be exploited to reduce boundary spectral leakage.

CRediT authorship contribution statement

Surinder Kaur: Conception and design of study, Acquisition of data. **Gopal Chaudhary:** Conception and design of study, Acquisition of data, Writing – original draft. **Smriti Srivastava:** Conception and design of study, Analysis and/or interpretation of data. **Manju Khari:** Conception and design of study, Analysis and/or interpretation of data. **Ruben Gonzalez Crespo:** Conception and design of study, Writing – original draft, Writing – review & editing. **Javalkar Dinesh Kumar:** Conception and design of study, Writing – original draft, Writing – review & editing.

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.

Acknowledgment

All authors approved the version of the manuscript to be published.

References

[1] Dua M, Gupta R, Khari M, Crespo RG. Biometric iris recognition using radial basis function neural network. Soft Comput 2019;23(22):11801–15.
[2] Gupta R, Khari M, Gupta D, Crespo RG. Fingerprint image enhancement and reconstruction using the orientation and phase reconstruction. Inform Sci 2020;530:201–18.
[3] Khari M, Garg AK, Crespo RG, Verdú E. Gesture recognition of RGB and RGB-d static images using convolutional neural networks. Int J Interact Multimedia Artif Intell 2019;5(7).
[4] Hernández FL, de Ory EG, Aguilar SR, Crespo RG. Residue properties for the arithmetical estimation of the image quantization table. Appl Soft Comput 2018;67:309–21.
[5] Pacheco A, Barón HB, Crespo RG, Espada JP. Reconstruction of high resolution 3D objects from incomplete images and 3D information. Int J Interactive Multimed Artif Intell 2014;2(6):7–16.
[6] Lombardo JM, Lopez MA, León M, Mirón F, Arranbarri J, Álvarez D. MOBEEZE. Natural interaction technologies, virtual reality and artificial intelligence for gait disorders analysis and rehabilitation in patients with Parkinson's disease. Int J Interact Multimed Artif Intell 2019;5(6):54–62.
[7] Hemeida AM, Mansour R, Hussein ME. Multilevel thresholding for image segmentation using an improved electromagnetism optimization algorithm. Int J Interact Multimedia Artif Intell 2019;5(4).
[8] Zhang D, Shu W. Two novel characteristics in palm-print verification: datum point invariance and line feature matching. Pattern Recognit 1999;32:691–702.
[9] Duta N, Jain AK, Mardia KV. Matching of palmprints. Pattern Recognit Lett 2002;23:477–85.
[10] Lu G, Zhang D, Wang K. Palmprint recognition using eigenpalms features. Pattern Recognit Lett 2003;24:1463–7.
[11] Wu X, Zhang D, Wang K. Fisherpalms based palmprint recognition. Pattern Recognit Lett 2003;24:2829–38.
[12] Kumar A, Shen HC. Palmprint identification using palmcodes. In: 2004 Third International Conference on Image and Graphics. Hong Kong, China. New York, NY, USA: IEEE.; 2004, p. 258–61.

[13] Lu GM, Wang KQ, Zhang D. Wavelet based independent component analysis for palmprint identification. In: 2004 International Conference on Machine Learning and Cybernetics. Shanghai, China. New York, NY, USA: IEEE; 2004, p. 3547–50.
[14] Chaudhary G, Srivastava S, Bhardwaj S. Multi-level fusion of palmprint and dorsal hand vein. In: 3rd International Conference on Information System Design and Intelligent Applications. Visakhapatnam, India. Berlin, GERMANY: Springer; 2016, p. 321–30.
[15] Srivastava S, Bhardwaj S, Bhargava S. Fusion of palm-phalanges print with palmprint and dorsal hand vein. Appl Soft Comput 2016;47:12–20.
[16] Li W, Zhang D, Xu Z. Palmprint identification by fourier transform. Int J Pattern Recognit Artif Intell 2002;16:417–32.
[17] Farrukh S, Hanmandlu M. Properties of information sets and information processing with an application to face recognition. Knowl Inf Syst 2017;52:485–507.
[18] Alexandre LA, Campilho AC, Kamel M. On combining classifiers using sum and product rules. Pattern Recognit Lett 2001;22:1283–9.
[19] Hanmandlu M, Grover J, Madasu VK, Vasirkala S. Score level fusion of hand based biometrics using t-norms. In: 2010 IEEE International Conference on Technologies for Homeland Security. Waltham, MA, USA. New York, NY, USA: IEEE; 2010, p. 70–6.
[20] Srivastava S. Accurate human recognition by score-level and feature-level fusion using palm–phalanges print. Arab J Sci Eng 2018;43:543–54.
[21] Grover J, Hanmandlu M. The fusion of multispectral palmprints using the information set based features and classifier. Eng Appl Artif Intell 2018;67:111–25.

Surinder Kaur is currently working as an Assistant Professor in Bharati Vidyapeeth's College of Engineering, Delhi, India. She received B.E. degree in Computer Science and Engineering in 2000 from Sant Longowal Institute of Engineering and Technology, Longowal, Punjab, and the M.Tech degree in Punjabi University, Patiala, Punjab, India, in 2003. Her current research interests include soft computing, Biometrics, Artificial Intelligence.

Gopal Chaudhary is currently working as an assistant professor in Bharati Vidyapeeth's College of Engineering, Delhi, India. He holds a Ph.D. in Biometrics at University of Delhi, B.E. in ECE, 2009 and the M.Tech. degree in Microwave and optical communication from Delhi Technological University, New Delhi, India, in 2012 and research interests include soft computing, information fusion and pattern recognition.

Smriti Srivastava received the B.E. degree in electrical engineering and the M.Tech. degree in heavy electrical equipment from Maulana Azad College of Technology [now Maulana Azad National Institute of Technology (MANIT)], Bhopal, India, in 1987 and 1991, respectively, and the Ph.D. degree in intelligent control from the Indian Institute of Technology, New Delhi, India, in 2005.

Manju Khari an Associate Professor in Jawaharlal Nehru University, New Delhi, India. She holds a Ph.D. in CSE from NIT Patna and ME in Information Security from AIACTR, Delhi, India. Her research interests are software testing, information security, optimization, Image processing and machine learning.

Rubén González Crespo has a Ph.D. in Computer Science Engineering with 200+ publications. Currently, he is Vice-Chancellor of Academic Affairs and Faculty from UNIR and Global Director of Engineering Schools from PROEDUCA Group. He is an advisory board member for the Ministry of Education at Colombia and evaluator from the National Agency for Quality Evaluation and Accreditation of Spain (ANECA).

Javalkar Dinesh Kumar is working as assistant professor in the Department of Electronics & communication Engineering, Lingaya's University, Haryana. He holds a Ph.D. in ECE, VLSI through IOT from Lingaya's Vidyapeeth (Deemed to be University), Faridabad, Haryana in 2019 and B.E. degree in ECE in 2012 and M.Tech. degree in VLSI System Design from Jawaharlal Nehru Technological University, Hyderabad in 2014.

Fake News Detection using Semi-Supervised Graph Convolutional Network

Priyanka Meel, Dinesh Kumar Vishwakarma
Biometric Research Laboratory, Department of Information Technology
Delhi Technological University, New Delhi, India-110042

Abstract: Social media becomes the central way for people to obtain and utilise news, due to its rapidness and inexpensive value of data distribution. Though, such features of social media platforms also present it a root cause of fake news distribution, causing adverse consequences on both people and culture. Hence, detecting fake news has become a significant research interest for bringing feasible real time solutions to the problem. Most current techniques of fake news disclosure are supervised, that need large cost in terms of time and effort to make a certainly interpreted dataset. The proposed framework concentrates on the text-based detection of fake news items while considering that only limited number of labels are available. Graphs are functioned extensively under several purposes of real-world problems on the strength of their property to structure things easily. Deep neural networks are used to generate great results within tasks that utilizes graph classification. The Graph Convolution Network works as a deep learning paradigm which works on graphs. Our proposed framework deals with limited amount of labelled data; we go for a semi-supervised learning method. We come up with a semi-supervised fake news detection technique based on GCN (Graph Convolutional Networks). The recommended architecture comprises of three basic components: collecting word embeddings from the news articles in datasets utilising GloVe, building similarity graph using Word Mover's Distance (WMD) and finally applying Graph Convolution Network (GCN) for binary classification of news articles in semi-supervised paradigm. The implemented technique is validated on three different datasets by varying the volume of labelled data achieving 95.27 % highest accuracy on Real or Fake dataset. Comparison with other contemporary techniques also reinforced the supremacy of the proposed framework.

Keywords: Semi-supervised, Graph Convolution Network, Word Mover's Distance, K-Nearest Neighbour, Global Vectors, Euclidean Space

1. Introduction

The constant extension of social media presented multiple parallel channels for users with added versatile facilities to obtain news than a few decades back. In accordance with the Pew

Research Center [1] , nearly 2/3 of U.S. grown-ups made headlines of social media. Since people proceed to profit of the service and simple convenience from social media, they additionally reveal themselves to some captivating and incorrect news published on social media platforms particularly fake knowledge. These fake news stories are purposely composed to carry incorrect information as a kind of idea so as business or governmental manipulation. The widespread of fake headlines could induce severe consequences on people and culture. Thus, identifying and alleviating fake news becomes a critical obstacle in current social media themes. The term “fake news” can be described in varied configurations that slightly differs from each other but most of the times can be used interchangeably as: misinformation, disinformation, hoax, propaganda, clickbait, misconception, fabrication, deceit etc. [2] [3] [4] [5] as pictured in word cloud representation in Figure 1.

The ubiquitous nature of web platforms encouraged their overwhelming use that fosters information sharing in an unrestricted way. This uncontrolled nature usually leads to generation and dissemination of biased contents. These untruthful events cause damage to social stability, economy and politics. One such rumour circulated on social media on 30 may 2018 regarding the spread of Nipah Virus through broiler chicken; causing huge financial loss to the dealers. On 4 January 2018 a fake news regarding collapse of metro pillar in Bangalore created unrest and panic in the public. A tidal wave of false claims and remedies has been burst out following COVID19 pademic such as curing the infection with garlic water, inhaling hot hair with hairdryer, use of certain vitamins or snake oils as remedy, use of vodka or vinegar as most effective hand sanitizer, unleashing lions on Russian streets to keep people indoor etc. These are few of the numerous instances of fake news circulating like wild fire on social media every day.

Paramount research in literature is engrossed around detecting fraudulent contents using rule-based methods or traditional machine learning classifiers. These traditional techniques fail to provide exemplary performance in the presence of large volumes of data, suffers with annotation inconsistencies, necessitate huge efforts for feature engineering and cannot provide near real time solutions. The reasons of inaccurate fact circulation spans across a broad spectrum incorporating genuine reporting faults, false interpretations, incorrect use of idioms or phrases, monetary benefits, political propaganda, marketing gimmick etc. but in all its camouflaged formats it poses a serious threat in front of the mankind. Therefore, in this work we tried to come up with a solution to this problem taking into consideration three key

technologies involving GloVe word embedding, Word Mover's Distance (WMD) and Graph Convolutional Networks (GCN).



Figure 1: Word cloud representing different connotation of fake news

Graphs normally occur in various real-world problems employing cultural and social issues. Information that is accessible to us can be transformed in a structured way by means of connections between various objects to give promising insights regarding data [5]. One crucial issue in Artificial Intelligence problems is to draw meaningful relation between one part of data to other parts of data. Currently the best way to describe/represent these relations and connections between data is by employing graphs data structure that exist universally in modern technology solutions by providing us with a plan to mathematically describe complicated data as well as associated parameters.

Artificial Intelligence problem solving techniques can be broadly categorized into three classes: supervised, unsupervised and semi-supervised. Supervised technologies are well established to provide benchmark solutions in healthcare, data authentication, web services, security and many more fields. The realm of semi-supervised and unsupervised paradigm still needs to be judiciously explored. The major pitfall of supervised training methods are annotation inconsistencies, cost and manpower required due to massive data size. In order to create a good balance between labelled and unlabelled data we tried to use both at their best level of performance. Labelled data steers the training process towards right direction and unlabelled data is used to enhance model generalization as well as performance. Therefore, a good balance of both is highly required that we categorically attempted to incorporate in our suggested framework.

Kusner et al [6] established Word Mover's Distance in 2015 as a distinct case of Earth Mover's Distance [7] and also proved that WMD gives unprecedented performance when coupled with K-nearest neighbour in classification tasks. This measure is suitable for calculating the distance between two text documents based on alignment between the words that are semantically close but syntactically different. It basically measures the dissimilarity between two text documents as the minimum distance the word vectors of one document need to travel to reach the word vectors of another document. GCN and its variants are being well tested to provide outstanding performance in variety of real-life applications incorporating both textual and visual data such as automatic seizure detection using temporal GCN [8], spatio-temporal GCN for picking up and returning demand of public bike sharing by exploring information from multi-view data [9], Autism Spectrum Disorder and Alzheimer's Disease prediction with imaging and non-imaging information [10] , semi-supervised hyperspectral image (HIS) classification framework using spectral-spatial graph convolution network (S^2GCN) [11], multi-stream attention-enhanced adaptive graph convolutional neural network (MS-AAGCN) for skeleton-based action recognition [12] etc.

The major problem we face in classification techniques is less no. of available labels for data. Apart from this cost, manpower and labelling inconsistencies are major concerns for annotating huge volumes of news articles. Thus, we make use of semi-supervised learning method. Graphs are capable of representing structural and contextual dependencies between its nodes, each representing an object and subsequently can be utilized in cases where the number of labels is limited [13]. This encourages us to take benefits of Graph Convolution Network for classification. Given one node, for all the classes can virtually start training procedure. Such a model, can very quickly lead to the creation of a 2d vector space with word embeddings involved which could be visualized easily. As a result, it can be seen that with just 3 layers in the architecture along with one label per actual class and absolutely no features in place, a linear classification is made. This produces quite exceptional results with given constraints and conditions. The key contributions of proposed framework are as follows:

- The architecture presents a semi-supervised fake news detection technique based on GCN (Graph Convolutional Networks) trained with limited amount of labelled data.
- Elaboratively elucidate three building blocks of the framework: collecting word embeddings from the news articles in datasets utilising GloVe, constructing similarity graph using Word Mover's Distance (WMD) and finally applying Graph Convolution Network (GCN) for binary classification of news articles in semi-supervised paradigm.

- Graph Convolution Network can harness the best advantage of convolutions as well as data structuring capabilities of graphs to draw meaningful insights out of complicated data and associated parameters.
- The implemented technique is validated on three different datasets by varying the volume of labelled data.
- Experimental results are analysed for two different graph formulations constructed by taking k=3 and k=5 .
- Comparison with other contemporary techniques also reinforced the supremacy of the proposed framework.

The organisation of the work is done under five different sections and sub-sections. Section 1 highlights the problem of fake news circulation and its adverse effects on mankind. Significant works done in literature to eradicate the problem of veracity analysis and content authentication on web platforms is detailed in section 2. Section 3 deals with the technical specifications of the proposed framework and its associated components in different sub-sections. Implementation, experimentation, datasets, testing and contemporary comparison is being described in section 4. Finally, section 5 concludes the work with possible future directions to improve it.

2. Related Works

The subject of fake news becomes an emerging issue within modern social media soundings. Current veracity analysis methods can be broadly classified into a couple of sections [13], [14] : utilising news data for handling social information/context, extracting linguistic, syntactic, lexical features or visual characteristics [15], [16] from news articles to obtain particular features as well as interesting patterns that usually happen in fraudulent articles. Visual features are adopted to recognise fake pictures that are deliberately produced to exert realistic effects in fake news. Context-based approaches of veracity analysis includes characteristics of user-data on social sites and social systems [17]. User social site's profile is utilised to estimate that users' features and reliability. Characteristics derived from their social posts describe their social behaviour. Pathak et al. [18] presented elaborated discussions on characteristics of publicly available datasets for veracity analysis techniques and performance comparison with contemporary machine learning, deep learning, hybrid approaches of supervised and unsupervised domain. Contemporary survey of semi-supervised and unsupervised methods of web content credibility analysis is detailed by Saini et al [19]. We

studied and categorized the available literature in the domain of veracity analysis into three different classes: supervised, semi-supervised and unsupervised methods. Apart from this also highlighted the unprecedented performance of different variants of GCN in a variety of real-life applications including healthcare, recommendation systems, traffic management, e-commerce, retail, marketing etc. that serves as a motivation to utilize it further for our designed fake news detection framework.

2.1 Supervised Methods

The supervised methods principally concentrate on selecting useful characteristics and utilise them to develop supervised training structures. A neural network-based Graph-aware Co-Attention Networks (GCAN) [20] framework considers realistic scenario of social network to predict whether a source tweet is fake or real at the same time producing the reasonable explanations as well as highlighting evidences on suspicious tweeters. The network requires the input in the form of short text tweets and the corresponding sequence of retweet users without text comments. GCAN significantly outperforms contemporary techniques by 16% in accuracy on average on Twitter 15 and Twitter 16 datasets. To model the sematic representations of fake news by jointly modelling the textual representations, knowledge concepts and visual information into a unified framework Wang et al. [21] proposed a novel Knowledge-driven Multimodal Graph convolutional Network (KMGCN). After knowledge distillation process a well-designed graph structure is used to effectively represent textual, knowledge and visual information to apply GCN succeeded by pooling layer, fully connect layer and finally softmax function for classification. The framework is experimented and validated on WEIBO and PHEME datasets. A weakly supervised reinforcement learning framework (WeFEND) proposed by Wang et al. [22] comprises of three components: annotator, reinforced selector and fake news detector. The annotator automatically assigns weak labels for news articles, the reinforced selector filters out high-quality samples using reinforcement learning and the fake news detector classifies fake articles based on content of news. Wu et al. [23] proposed a graph-based method of rumour detection in which a propagation graph is being constructed based on who replies to whom relationship on Twitter. Two models GLO-PGNN (global embedding with propagation graph neural network) and ENS-PGNN (ensemble learning with propagation graph neural network) adopting diverse classification approaches for rumour detection task are designed. Both the methods employ attention mechanism for dynamically adjusting the weights to improve performance. Meel and Vishwakarma [24] proposed multimodal fake news detection method based on ensemble of

hierarchical attention network, image captioning and forensics analysis. Raj and Meel [25] experimented with Resnet50 , VGG16, VGG19,InceptionV3, DenseNet, Xception, AlexNet , MobileNet deep arcitectures on T1-CNN, EMERGENT, MICC-F220 multimodal datasets to compare and contrast the capacities of different ConvNet architectures for textual as well as visual forgery detection.

2.2 Semi-supervised methods

A semi-supervised naïve architecture Factual News Graph (FNG) [26] based on graphical social context representation is proposed that demonstrates significant improvements in fake news detection task and robustness in the presence of limited training data. The FANG representation learning framework is optimized by three concurrent losses unsupervised proximity loss, self-supervised stance loss and supervised fake news detection loss. Guacho et al. [27] recommended semi-supervised misinformation detection in graphs via embedding the input articles in multi-dimensional tensors which are further decomposed to capture spatial and contextual information by creating article-by-article-graph. A temporal ensembling based semi-supervised fake news detection framework using self-ensembling of the combined prediction of all previous epochs to serve as a proxy label for the current epoch is proposed by Meel and Vishwakarma [2] . Online opinion reviews have an economic impact on goods and services. Opportunistic firms often manipulate online reviews to make undue profits by unethically influencing people's choices. Rout et al. [28] proposed a semi-supervised method of classifying deceptive and fake opinion reviews and validated the framework on hotel reviews. A novel semi-supervised framework for multimodal fake news detection is designed by Mansouri et al. [29] using Linear Discrimination Analysis (LDA) and Convolutional Neural Network (CNN). Li et al [30] proposed a self-learning semi-supervised deep learning network that adds a confidence network layer to made it possible to automatically return and add correct results for improving the accuracy of the neural network.

2.3 Unsupervised methods

Annotated samples cannot characterize the authenticity of the news on recently developed event as they become obsolete very quickly due to the dynamic nature of news. To obtain high-quality up-to-the-minute labelled sample is the foremost challenge in supervised and semi-supervised learning methodologies. The motive behind development of unsupervised technologies is to completely eliminate the need of annotation and real time solution to the problem of information trustworthiness. A novel unsupervised method of detecting fake news

circulating on WhatsApp is implemented by Gaglani et al. [31] leveraging Transfer Learning techniques. The framework utilizes semantic similarity between claims circulated on WhatsApp and associated articles scrapped from web to classify the claims as true or false. Hosseinimotlagh and Papalexakis [32] recommended an unsupervised ensemble method that consolidates results from different tensor decompositions into coherent, comprehensible and high precision groups of articles that belong to different categories of false news. Unsupervised microblog rumour detection framework is proposed [33] based on recurrent neural network and autoencoders by analysing the temporal dynamics and crowd wisdom. Yang et al. [34] recommended a generative approach of unsupervised fake news classification on social media by exploiting user's engagement, opinion and their credibility factors. A three phased graph-based approach utilizing biclique identification, graph-based feature vector learning and label spreading for fake news detection in the absence of annotated data is successfully designed by Gangireddy et al. [35].

2.4 Application areas of GCN

Detailed survey analysis of GCN in different real-life applications is being presented by Singh et al. [36]. Graph Convolution Network is being explored by Yao et. al [37] in conjunction with LSTM network to explore between objects in an image based on their spatial and semantic connections for generating image captions. Jeong et al. [38] proposed a deep learning based context aware citation recommendation framework encompasses a document encoder and content encoder using GCN and BERT .Other remarkable contributions of GCN for designing solutions to challenging problems can be highlighted as: semi-supervised hyperspectral image (HIS) classification framework using spectral-spatial graph convolution network (S^2GCN) [11], Infusing knowledge into the textual entailment tasks [39], multi-label image classification framework using GCN [40] , robot navigation in crowds using GCN with attention trained by human gaze data that accurately predicts human attention to different agents in the crowd [41] etc.

2.5 Motivation and Research Objectives

The inspirations for developing semi-supervised fake news detection framework and prime research targets can be regarded as follows:

- To design a semi-supervised Fake News Detection framework that can learn from the semantics of labelled data and intrinsic patterns of unlabelled data to optimize in terms of time, cost, labour and annotation inconsistencies.

- To address the highly circulated forgery format of a news story coupled with headline text along with partially labelled training samples.
- To take advantage of self-learning cognition of designed framework which facilitates the architecture to learn from up-to-the minute latest forgery formats fed into the process in the form of unlabelled articles.
- Kusner et al [6] established the fact that Word Mover's Distance gives unprecedented performance when coupled with K-nearest neighbour in classification tasks.
- Graph Convolution Network can harness the best advantage of convolutions as well as data structuring capabilities of graphs are capable of representing structural and contextual dependencies between its nodes (documents) to draw meaningful insights out of complicated data and associated parameters.
- The framework is aimed to avoid cumbersome feature engineering of supervised learning methods, crowdsourcing or human expert annotation inconsistencies, reduces cost and manpower requirements.

3. Proposed Model

The semi-supervised framework of Graph Convolutional Network combines best features of convolutions and data modelling capabilities of graphs. The proposed architecture incorporates a three-step approach to classify the articles as fake or real. First step is to transform the given textual data from a dataset in the form of vectors. These vectors represent the linguistic features of the text and can be utilized to represent an article in Euclidean Space. Global Vector (GloVe) embedding is used to transform the articles into vectors. Each article is interpreted as the mean of vectors of the words it contains. The result of this embedding is used to construct a similarity graph between articles in the dataset. Each node represents an article and most similar nodes are connected by an edge in the graph. To calculate the similarity, i.e. distance between two articles, Word Mover's Distance is used. It uses semantically meaningful relations between words to find the similarity between two articles. Before converting text into vectors basic pre-processing steps such as: dropping unused columns, removing null and missing data, removing stope words, tokenizing the articles, lemmatizing each word, converting labels name etc. are done for all the datasets. The subsequent procedure involves word embedding and similarity graph construction. This results in creation of graphs containing three types of nodes labelled as real, fake and unlabelled which can be inputs to the graph

convolutional neural networks for finally predicting the output. The process is pictorially emphasized in figure 2 and also comprehensively detailed in algorithm1.

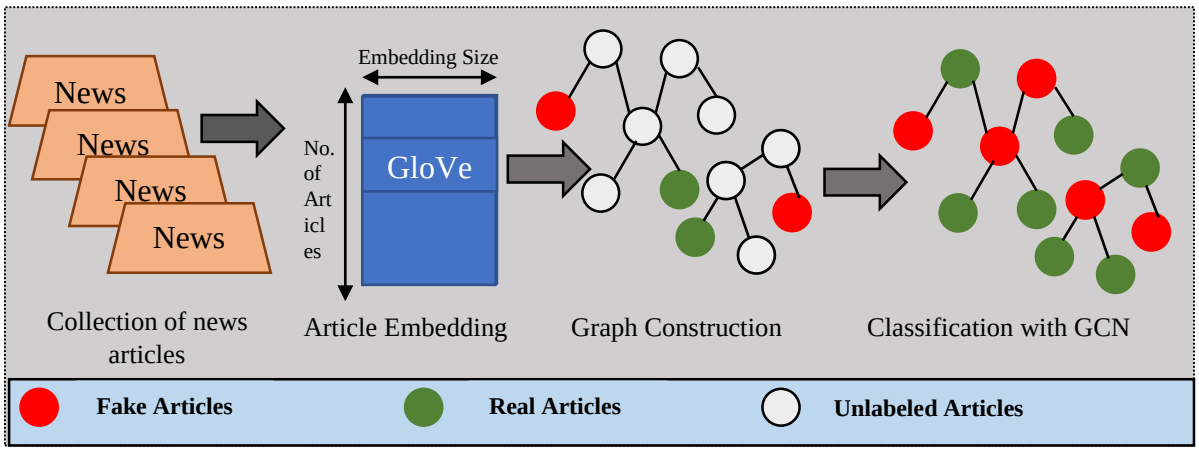


Figure 2: Proposed semi-supervised fake news classification framework

Algorithm 1: Learning in Semi-supervised GCN Framework

1: Perform initial pre-processing on input samples
2: Split the input samples into 80% training,10% validation,10% testing
3: Article embedding in Euclidean space using Glove with Embedding Dimension=300
4: Represent every article as the mean vector of the corresponding GloVe embedding
5: Similarity graph construction using WMD with K-nearest neighbour; K=3 and K=5
6: Train GCN of 4 layers up to 120 epochs with dropout=0.5 and learning rate=0.01
7: Evaluate the performance on test set

3.1 Article Embeddings in Euclidean Space

Article embedding in Euclidean Space means transforming text to multidimensional vectors by using word embedding. Word embedding is the process of constructing mathematical equivalents, i.e., representation in the mathematical form of each word from some corpus. It is implied that words with similar or same sense will have close representations. In Natural Language Processing, word embedding has proved to be one of the concepts that have huge applications, as it lays the foundation for solving many real-world tough problems. The process involves constructing vectors for each word in any number of dimensions such that it can be visualized as a vector in a vector space in those number of dimensions. This conversion of words to vectors uses neural network methodology to find the values of vectors. The vector can have any number of dimensions ranging from a single digit to a few hundred. As words with similar or same sense will have close representations, it can be implied that the meaning of words has been learnt by a model, which in turn can be useful for solving challenging problems. Our method will utilize the final word vectors created by the embedding of words for each article by representing it in an euclidean space. Further this will also be

converted into low dimensions using a dimensionality reduction technique and contextual associations in news articles will be obtained by using a similarity graph representation.

To create a vector which maps to each news article, our proposed semi-supervised method uses a pre-trained GloVe model of 300 dimensional embeddings and for all the words appearing in that article we compute a mean vector which is finally mapped to that article. GloVe (Global Vectors) an unsupervised learning method was developed at Stanford university as an open-source project by Pennington et al [42] and was launched in 2014. Word Vectors in such an n-dimensional space, occurs in the manner as similar words appear close to each other while words with different meanings appear far from each other. One major advantage that GloVe offers is its dependency on global statistics instead of local statistics as used by Word2Vec or some other embedding techniques. GloVe is based on the basic principle of co-occurrence matrix. A simple example of a co - occurrence matrix of window size 1 is depicted in Figure 3, illustrates how words befall collectively that ultimately produces the relations between individual words. Co-occurrence matrix for words is calculated just by adding the mutuality of occurrence of words in an obtained text. In general use cases this co-occurrence matrix is decomposed by employing dimensional reduction techniques like PCA and SVD.

	the	cat	sat	on	mat
the	0	1	0	1	1
cat	1	0	1	0	0
sat	0	1	0	1	0
on	1	0	1	0	0
mat	1	0	0	0	0

Figure 3: An example co-occurrence matrix for the sentence “the cat sat on the mat”

3.2 Similarity Graph Construction

The classification step requires a graph input, which is done by constructing a similarity graph representing closeness in between different nodes, i.e., each news article. Such a graph can be represented as a k-nearest neighbour graph. In particular, for every news article, we look forward towards the k–nearest neighbours in the embedding space. The k neighbours can be determined by calculating the Word Mover’s distance between one article to all the other news articles.

A k-NN (k-nearest neighbour) graph is one in which node p and node q have an edge between them if node p is one of the top k nearest of all its neighbours or vice-versa. Every k-nearest-neighbours of a location in the n-dimensional space will be determined to utilise a

"closeness" relationship wherever closeness is usually described within words of a distance measure as Word Mover's Distance. Thus, with given articles in a vector space, a k-NN graph of points can be created by calculating the remoteness between each pair of points and connecting each point with the k most proximal ones.

The Word Mover's Distance (WMD) proposed by Kusner et al. [6] in 2015 is a distance between documents that takes benefit of semantic relations among words that are apprehended by their embeddings. It is a special case of Earth Mover's Distance [7] and also provides extraordinary performance when coupled with K-nearest neighbour in classification tasks. It basically measures the dissimilarity between two text documents as the minimum distance the word vectors of one document need to travel to reach the word vectors of another document. This distance proved to be quite effective, obtaining state-of-art error rates for classification tasks.

WMD metric is a distance function between text documents that leads to unparalleled low k-nearest neighbour document classification error rate. Precisely representing the remoteness between two documents has far reaching utilities in document retrieval, multi-lingual document matching, text clustering etc. To calculate a distance between two text documents the basic unit is "travel cost" between two words. Let "A" and "B" be the representations of two text documents with |A| and |B| are the number of unique words in both the documents respectively. Each word w_i in document A is allowed to be converted into any word in document B. $T_{ij} \geq 0$ denotes the amount of distance word w_i in A has to travel to convert into word w_j in B. T is a sparse transportation flow matrix. To transform A completely into B we have to confirm that the complete outgoing flow from ith word equals A_i and incoming flow to jth word must match B_j represented mathematically as equation 1.

$$\sum_j T_{ij} = A_i \text{ and } \sum_i T_{ij} = B_j \quad (1)$$

The WMD distance between two documents is the minimum weighted cumulative cost required to move all words from document A to document B represented as

$$\text{WMD (A, B)} = \sum_{ij} T_{ij}\, c(w_i, w_j) \quad (2)$$

Formally the WMD between two documents is defined as the value of the optimal solution of the following transportation problem which is a special case of Earth mover's distance.

$$\min_{T \geq 0} \sum_{i=1}^{|A|} \sum_{j=1}^{|B|} c(w_i, w_j)\, T_{ij} \quad (3)$$

$$\sum_{j=1}^{|B|} T_{ij} = A_i \qquad \forall\ i \in \{1, 2, 3 \ldots\ldots |A|\} \tag{4}$$

$$\sum_{i=1}^{|A|} T_{ij} = B_j \qquad \forall\ j \in \{1,2, 3 \ldots\ldots |B|\} \tag{5}$$

$$T_{ij} \geq 0 \qquad \text{for all i , j} \tag{6}$$

Word Mover's Distance (WMD) is derived upon current events in embeddings of the words which determine the semantically significant description of those words of local co-occurrences within those sentences among new articles. Salient features of Word Mover's Distance can be characterized as:

- It is easy to learn, practice and free from the effect of any hyperparameters.
- It can be easily described as the distance among two text contents that can be split and described as the distances which are sparse within different words.
- It commonly includes some information represented as the word2vec or Glove and heads to huge operations accuracy.

3.3 GCN for Graph Classification

Graph Convolutional Network is a resourceful development of convolutional neural networks which functions directly on graphs. The model scales linearly in the number of graph edges by using a competent layer-wise propagation rule that is based on first-order approximation of spectral convolutions on graphs. GCN model is capable of learning hidden layer representations that encodes both node feature and graph structure to an extent useful for semi-supervised classification.

Graph Convolutional Networks (GCN) are an ideological extension to Convolutional Neural Networks (CNN) where convolution process is applied on a graph instead of pixels which constitute the image. As CNN can capture information from the images and this information then can be used to classify images, a similar approach can be built over graphs as well. A filter analogous to a CNN filter can be employed in case of GCN to capture the similarity in graphs. Graph Convolutional Networks (GCN), like Convolutional Neural Networks (CNN) have bounded no. of hyper-parameters, which leads these techniques to occupy less memory and thus, multiple levels of information can be built to give a remarkable result when compared to a traditional learning algorithm. Figure 4 illustrates the basic design pipeline for a GCN model.

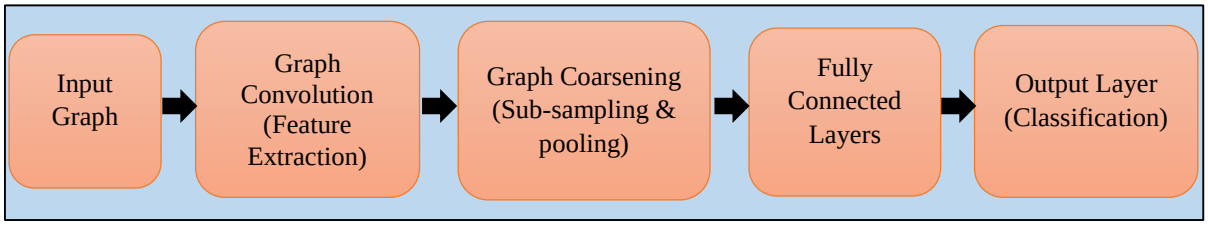


Figure 4: Design pipeline for GCN Model

The main purpose of graph classification is to foretell the labels of nodes within that particular graph. A node itself in the graph normally denotes a real-world object i.e. in our framework it represents news articles. The graph has been extensively employed to represent objects of real-world and the connection between them. Node/graph classification results prediction of labels of nodes in the graphs, for this purpose several researches take advantage of relationships among nodes to increase the classification efficiency.

GCN is one of the prominent variants of Graph Neural Network being used for a lot of real-life applications with non-Euclidean real world data having graph structure. The major difference between CNN and GNN is that GNN is a generalized variant of CNN built to operate on irregular or non-Euclidean structured data whereas the CNN can function only in case where the underlying data is regular (Euclidean). GCN is developed by Thomas Kipf and Max Welling [43] in 2016. Convolution in GCN is almost the same operation as the convolution in CNN. It implies multiplying the input neurons with a set of weights known as kernels or filters acts as sliding window and enables the network to learn features from neighbouring cells. Different layers may contain filters of different weights but within the same layer same filter will be used known as weight sharing. The particular variant of GCN that we use in our framework for semi-supervised classification is Spectral Graph Convolutional Network proposed by Thomas Kipf and Max Welling.

Neural networks apply non-linear activation functions to represent the non-linear features in latent dimensions. Forward pass equation in neural network is represented as in equation (7) where σ represents activation function, $H^{(l+1)}$ and $H^{(l)}$ represents feature representation at l^{th} and $(l+1)^{th}$ layer, $W^{(l)}$ weight and $b^{(l)}$ bias at layer l.

$$H^{(l+1)} = \sigma(W^{(l)} H^{(l)} + b^{(l)}) \quad (7)$$

Forward pass equation in Graph Convolutional Network can be represented as:

$$H^{(l+1)} = \sigma(W^{(l)} H^{(l)} A) \quad (8)$$

A is adjacency matrix representing the connections between the nodes in graph structured real world data. Adjacency matrix A enables the model to learn the feature representations based on nodes connectivity. Bias b is omitted to make the model simpler. Equation (8) is the first-order approximation of spectral graph convolution propagating the information along the neighbouring nodes within the graph.

The aim of Graph Convolutional Network framework is to learn a function of features on a graph which takes as input a feature matrix X of dimension NxD where N is the number of nodes and D is number of input features. A is the adjacency matrix for representing the overall graph structure. A and X are the input to GCN architecture producing the node level output Z (an NxF feature matrix, where F is the number of output features per node). A nonlinear function representing every neural network layer in GCN can be written as in equation (9) and equation (10) highlights layer wise propagation across the network.

$$H^{(l+1)} = f(H^{(l)}, A) \tag{9}$$

$$f(H^{(l)}, A) = \sigma(A\,H^{(l)}\,W^{(l)}) \tag{10}$$

With $H^{(0)}$=X is initial input, $H^{(L)}$=Z is node level output at last layer (or z for graph level output) , L is the number of layers , $W^{(l)}$ is a weight matrix for the l^{th} neural network layer and σ is a non-linear activation function. Instead of using matrix A for semi-supervised classification in practical scenario Kipf and Welling [43] proposed to use symmetric normalization $D^{-1/2}AD^{-1/2}$, so equation (9) can be rewritten as:

$$f(H^{(l)}, A) = \sigma(\hat{D}^{(-1/2)}\hat{A}\,\hat{D}^{-1/2}\,H^{(l)}\,W^{(l)}) \tag{11}$$

With $\hat{A}$ = A+I, I is identity matrix and $\hat{D}$ is the diagonal node degree matrix of A. In the process of semi-supervised classification with GCN the network initially starts training on the labelled nodes, subsequently propagating the information to unlabelled nodes by updating weight matrices that are shared across all nodes. This process can be summarized in the following steps:

- Accomplish forward propagation through GCN.
- Put on the sigmoid function row-wise on the last layer in the GCN.
- Calculate the cross-entropy loss on known node labels.
- Backpropagate the loss and update the weight matrices ‘W’ in each layer.

4. Experiments

In this sub-section, we extensively detail the experimentation settings, datasets, parameter selection and insights gained from the wide range of experiments performed. The implementation is done on Google Colab which offers up to 13.53 free RAM and 12 GB NVIDIA Tesla K80 GPU. The proposed framework is built and implemented in Python 3 on top of the Keras deep learning framework. To make our implementation better and effective the experiments are repeated for different volumes of labelled and unlabeled data samples. The performance scores and comparison of the proposed framework are listed in terms of F1-measure, accuracy, recall and precision evaluation metrics. Analysis of the results is being done in numerical as well as in graphical representation using the shapes of accuracy-loss curves with epochs and area under the curve plots individually for each dataset.

After initial pre-processing each dataset is split into 8:1:1 ratio for training, validation and testing respectively. Articles are embedded in vector space using 300-dimensional GloVe word embedding. Word Mover's Distance is employed for constructing similarity graph utilizing K-nearest neighbours for two different values of K i.e. K=3 and K=5. Graph convolutional network involves 4 layers convolutions for feature extraction with intermediate leaky ReLU activation and dropout layers. GCN of 4 layers is trained up to 120 epochs with dropout 0.5 , learning rate 0.01 , 16 hidden units and weight decay of 5^e-4 . After fully connected layers in output layer activation function is softmax for binary classification of news instances into real or fake.

4.1 Datasets

Three different fully labelled datasets Fake News Data, Real or Fake and Fake News Detection introduced on Kaggle platform are used for experimentation and validation of our work. The headline, body and label part of each one of the datasets are being utilized for model training and testing purpose. The details of the datasets are listed in following Table 1:

Table 1: Dataset Details

Datasets	**Details**	**Attributes Used**	**Total entries**	**Fake news count**	**Real news count**
Fake News Data [44]	Hosted on Kaggle, contains id, Title, Author, Text and Label	Headline/Title, Body/Text, Label	20700	10360	10340
Real or Fake [45]	Hosted on Kaggle platform, contains four fields Id, Headline, Body , Label	Headline/Title, Body/Text, Label	6000	3000	3000
Fake News Detection [46]	Hosted on Kaggle, contains site URL, Headline, Body and Label	Headline/Title, Body/Text, Label	3988	2121	1867

Fake News Data [44] was presented on Kaggle website for Kaggle competition two years ago, now available publicly with annotations for research and learning purpose. It comprises of 20, 800 instances with five attributes: Id, Title, Author, Text and Label. After initial pre-processing, we have 20, 700 entries in the dataset with 10360 Fake news and 10340 real news entries. We utilized Title, Text and label features for our model training, testing and comparison purpose.

Real or Fake [45] dataset was introduced on Kaggle platform three years ago, now is being extensively used for research purpose. It contains four attributes Id, Headline, Body and Label out of these four we have used headline, body and label part in our research. The dataset initially has 6335 entries reduced to 6000 after preliminary data pre-processing. Segregation of dataset is 3000 real and 3000 fake news instances.

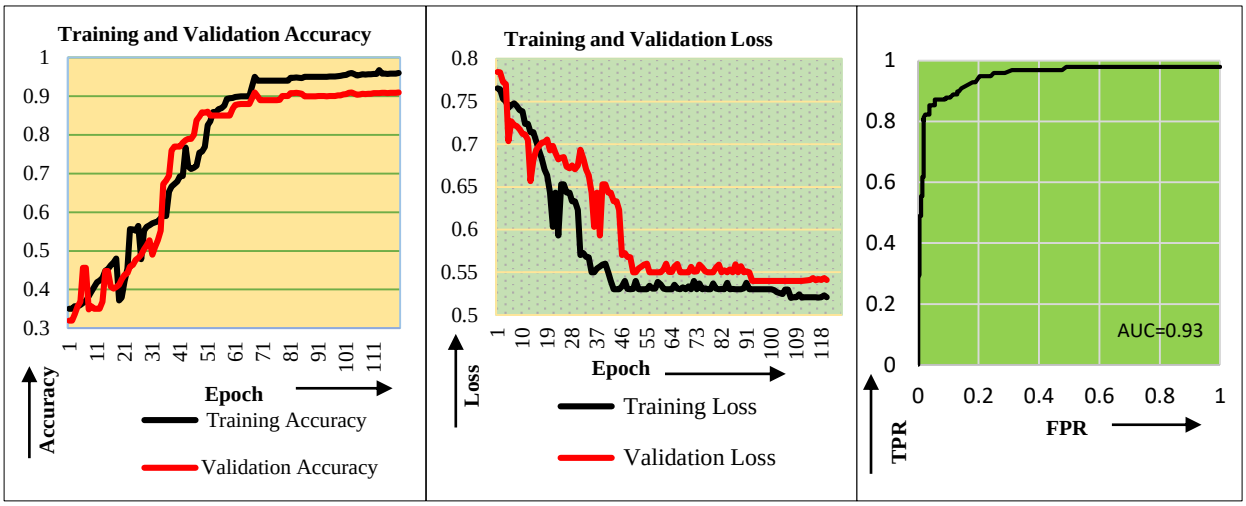
Fake News Detection [46] dataset is compiled on the Kaggle website by Jruvika. It contains four attributes site URL, Headline, Body and Label (Real/Fake). The dataset initially contains 4009 news instances. After initial data cleaning such as removing the entries with misplaced labels, missing headline and body we have 3988 rows with 2121 Fake and 1867 Real news samples.

4.2 Result Analysis

Performance of the proposed framework on three different datasets is being evaluated and compared for accuracy, precision, recall and F1 score. The experiments for each dataset have been repeated for two different values of K (K=3, K=5) and varied proportions of labelled training data ranging from 20% to 50%. Table 2 ,3 and 4 highlights the results obtained out of experiments for Fake News Data, Real or Fake and Fake News Detection datasets respectively. Figure 5 ,6 and 7 enlightens the Accuracy-Epoch curve, Loss-Epoch curve and ROC curves for each one of the datasets.

Table 2: Result analysis on Fake News Data dataset

% Labelled data	**K=3**				**K=5**			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
20%	79.29	82.52	74.65	78.39	66.03	61.00	79.27	68.94
30%	83.26	86.07	77.32	81.46	74.66	85.32	72.44	78.35
40%	87.99	75.99	88.60	81.81	82.36	69.02	87.45	77.15
50%	**91.18**	**93.05**	**94.27**	**93.65**	88.67	81.71	89.26	85.32

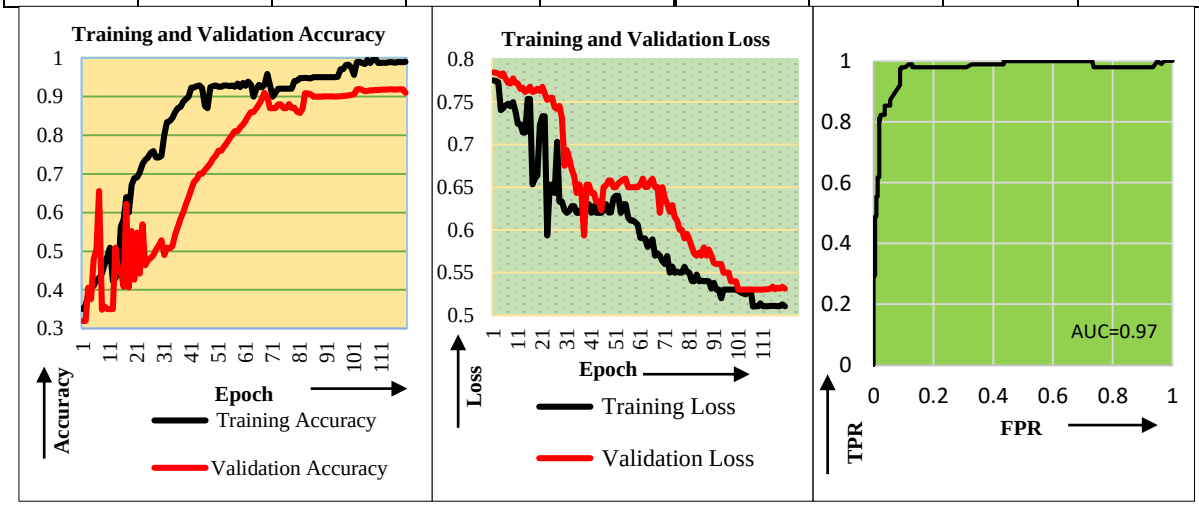


(a) (b) (c)

Figure 5: (a) Accuracy-Epoch curve (b) Loss-Epoch curve (c) ROC curve for Fake News Detection dataset

Table 3: Result analysis on Real or Fake dataset

% Labelled data	K=3				K=5			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
20%	79.55	83.75	79.99	81.83	80.00	83.69	70.08	76.28
30%	86.32	75.00	86.76	80.45	83.33	75.00	79.27	77.08
40%	91.02	89.00	90.23	89.61	87.19	76.68	93.55	84.28
50%	**95.27**	**89.47**	**95.99**	**92.61**	92.34	84.29	92.52	88.21

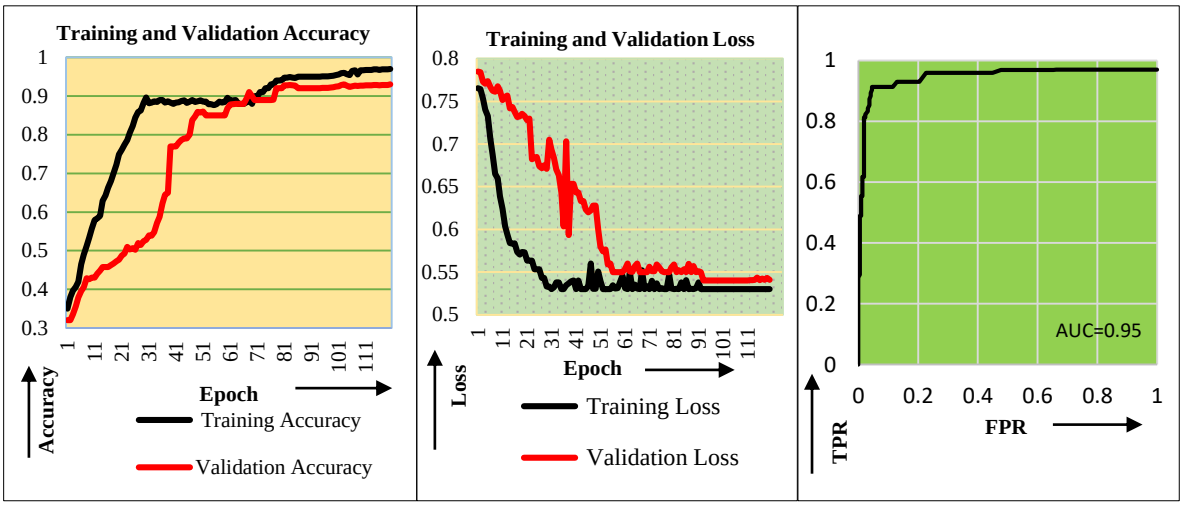


(a) (b) (c)

Figure 6: (a) Accuracy-Epoch curve (b) Loss-Epoch curve (c) ROC curve for Real or Fake dataset

Table 4: Result analysis on Fake News Detection dataset

% Labelled data	K=3				K=5			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
20%	75.03	76.06	71.27	73.58	77.32	81.33	76.65	78.92
30%	81.26	75.60	86.59	80.72	86.23	87.88	78.19	82.75
40%	85.04	88.27	87.77	88.02	88.69	86.01	86.62	86.31
50%	**92.03**	**92.07**	**94.87**	**93.45**	90.37	83.33	90.17	86.61



(a) (b) (c)

Figure 7: (a) Accuracy-Epoch curve (b) Loss-Epoch curve (c) ROC curve for Fake News Detection dataset

It is evident from the results analysed from Table 2, 3 and 4 that K=3 provides better accuracies than K=5 and as we increase the amount of labelled data in training process precision of detecting fake news also increases. The highest accuracy obtained from Fake News Data dataset is 91.18%, Real or Fake dataset is 95.27 % and Fake News Detection dataset is 92.03% respectively. Graphical representation of the experimental process in Figures 5, 6 and 7 advocates the overall effectiveness of the framework in terms of quality and quantity of performance.

4.3 State-of-the-art Comparison

To compare the efficacy of our designed architecture with contemporary methods, we calculated the performance of five different baseline methods in terms of accuracy, precision, recall and F1 score with the dataset split as 8:1:1 for training, validation and testing. State-of-the-art juxtaposition on each of the Fake News Data, Real or Fake and Fake News detection datasets are outlined in Table 5, Table 6 and Table 7, correspondingly. The approaches used as baselines for state-of-the-art comparison are as follows:

- Bali et al. [47] proposed Support Vector Classifier (SVC), Random Forest(RF), Naïve Bayes (NB), Multi-Layer Perceptron (MLP), K-Nearest Neighbour(KNN), AdaBoost

(AB) and Gradient Boosting (XGB) methods by extracting sentiment polarity, 50-dimensional GloVe word embedding , n-gram count, and cosine similarity features between title and text part of news articles.

- Agarwalla et al. [48] applied Punkt statement tokenizer from NLTK library with Support Vector Machine (SVM), Logistic Regression and Naïve Bayes with Lidstone smoothing for classification.
- Karimi and Tang [49] developed a Hierarchical Discourse Level Structure using Bi-LSTM, which extracts structure-related properties of articles by building dependency trees.
- Vishwakarma et al. [50] suggested a framework of keyword extraction with web scrapping using a rule-based classifier. The classifier uses a reality parameter (Rp) which is considered by checking the credibility of the top 15 Google search results.

Table 5: Comparative performance analysis on Fake News Data dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Bali et al. [47]	91.05	93.00	94.00	93.49
Agarwalla et al. [48]	86.32	85.89	81.26	83.51
Karimi and Tang [49]	85.90	88.80	80.70	84.56
Vishwakarma et al. [50]	85.00	95.04	80.00	86.87
Proposed Method	**91.18**	**93.05**	**94.27**	**93.65**

Table 6: Comparative performance analysis on Real or Fake dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Bali et al. [47]	93.03	92.00	92.44	92.22
Agarwalla et al. [48]	90.00	89.90	90.00	89.94
Karimi and Tang [49]	89.06	82.90	88.70	85.70
Vishwakarma et al. [50]	91.01	90.50	95.70	93.03
Proposed Method	**95.27**	**89.47**	**95.99**	**92.61**

Table 7: Comparative performance analysis on Fake News Detection dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Bali et al. [47]	86.20	92.00	92.00	92.00
Agarwalla et al. [48]	83.87	81.22	82.67	81.93
Karimi and Tang [49]	83.09	80.22	82.01	81.10
Vishwakarma et al. [50]	88.30	85.20	88.40	86.77
Proposed Method	**92.03**	**92.07**	**94.87**	**93.45**

The above discussion concludes that our proposed model for veracity analysis of web information is reasonably promising. The precision over all the three datasets Fake News Data, Real or Fake and Fake news detection is a decent development over the parallel methods. Graph Convolutional Network and Word Mover's Distance for calculating distance are the prominent technologies that have facilitated refining the preciseness of our results. Finally, experimenting with different values of K and varied proportions of labelled data helps us achieve 95.27% highest fake news detection accuracy.

5. Conclusion and Future Works

Fake news is an increasing concern for the modern society. One of the most difficult part in providing a solution to this is labelling massive volumes of data to train supervised artificial intelligence models. To counter this challenge, we tried to design a semi-supervised text fake news detection framework based on Graph Convolutional Networks. Embedding the text articles in Euclidean space, similarity graph constructing using Word Mover's Distance and Graph Classification are the three landmark components of the designed architecture. Extensive experimental analysis has been done by repeating the training and testing of the model for different proportions of labelled and unlabelled data for each dataset. The promising results compared in terms of accuracy, precision, recall, F1 score and ROC Curves very well advocates the worthiness of the method for veracity analysis. The proposed framework exhibits encouraging results by outdoing numerous state-of-the-art methods on multiple standard datasets.

In spite of all the efforts done by researchers, governments and corporates to put an end to this malice of fake news, a lot more is still to be done. Temporal component is extremely important in current real-life decision-making process. Correct information received too late is also irrelevant, so real time information authentication systems are highly in demand. Platforms such as Facebook, WhatsApp, LinkedIn, Gmail etc. are implementing solutions for user privacy as well as data security and trustworthiness. The research could be further extended to incorporate multimedia data as visual data constitute an integral part of digital online information. Advanced deep learning technologies of transformers and transfer learning can also be amalgamated with graph data structures to make the process more efficient. Stand-alone modules that can be added as browser and application plugins could also be the subsequent research tracks. The classification performance of fake news detection systems can also be improved by using attention mechanism instead of complex convolutions as well as

development of unsupervised frameworks on large scale unlabelled data further enhances the worthiness of the procedure.

References

[1] "Pew Research Center," [Online]. Available: https://www.pewresearch.org/.

[2] P. Meel and D. K. Vishwakarma, "A temporal ensembling based semi-supervised ConvNet for the detection of fake news articles," *Expert Systems with Applications,* vol. 177, pp. 1-14, 2021.

[3] D. Varshney and D. K. Vishwakarma, "A review on rumour prediction and veracity assessment in online social network," *Expert Systems With Applications,* vol. 168, 2021.

[4] D. Varshney and D. K. Vishwakarma, "A unified approach for detection of Clickbait videos on YouTube using cognitive evidences," *Applied Intelligence,* vol. 51, pp. 4214-4235, 2021.

[5] D. K. Vishwakarma and C. Jain, "Recent State-of-the-art of Fake News Detection: A Review," in *International Conference for Emerging Technology*, Belgaum, India, 2020.

[6] A. Benamira, B. Devillers, E. Lesot, A. K. Ray, M. Saadi and F. D. Malliaros, "Semi-Supervised Learning and Graph Neural Networks for Fake News Detection," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Vancouver, BC, Canada, 2019.

[7] M. J. Kusner, Y. Sun, N. I. Kolkin and K. Q. Weinberger, "From word embeddings to document distances," in *International conference on machine learning (PMLR),* Lille , France, 2015.

[8] Y. Rubner, C. Tomasi and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International journal of computer vision,* vol. 40, no. 2, pp. 99-121, 2000.

[9] I. C. Covert, B. Krishnan, I. Najm, J. Zhan, M. Shore, J. Hixson and M. J. Po, "Temporal graph convolutional networks for automatic seizure detection," in *Machine Learning for Healthcare Conference (PMLR)*, Michigan, USA, 2019.

[10] G. Xiao, RuinanWang, C. Zhang and A. Ni, "Demand prediction for a public bike sharing program based on spatio-temporal graph convolutional networks.," *Multimedia Tools and Applications,* pp. 1-19, 2020.

[11] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. Guerrero, B. Glocker and D. Rueckert, "Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease," *Medical image analysis,* vol. 48, pp. 117-130, 2018.

[12] A. Qin, Z. Shang, J. Tian, Y. Wang, T. Zhang and Y. Y. Tang, "Spectral–spatial graph convolutional networks for semisupervised hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters,* vol. 16, no. 2, pp. 241-245, 2018.

[13] L. Shi, Y. Zhang, J. Cheng and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Transactions on Image Processing ,* vol. 29, pp. 9532-9545, 2020.

[14] B. Malhotra and D. K. Vishwakarma, "Classification of Propagation Path and Tweets for Rumor Detection using Graphical Convolutional Networks and Transformer based Encodings," in *IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, New Delhi, 2020.

[15] A. Bondielli and F. Marcelloni, “A survey on fake news and rumour detection techniques,” *Information Sciences,* vol. 497, pp. 38-55, 2019.

[16] K. Sharma, F. Qian, H. Jiang and N. Ruchansky, "Combating fake news: A survey on identification and mitigation techniques," *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 10, no. 3, pp. 1-42, 2019.

[17] P. Meel and D. K. Vishwakarma, “Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities,” *Expert Systems with Applications,* vol. 153, p. 112986, 2020.

[18] Roohani, T. Rana and P. Meel, “Rumor Propagation: A State-of-the-art Survey of Current Challenges and Opportunities,” in *2nd International Conference on Intelligent Communication and Computational Techniques (ICCT),* Jaipur , India, 2019.

[19] M. D. Vicario, W. Quattrociocchi, A. Scala and F. Zollo, “Polarization and Fake News: EarlyWarning of Potential Misinformation Targets,” *ACM Transactions on the Web,* vol. 13, no. 2, pp. 1-22, 2019.

[20] A. R. Pathak, A. Mahajan, K. Singh, A. Patil and A. Nair, “Analysis of Techniques for Rumor Detection in Social Media,” in *International Conference of Computational Intelligence and Data Science (ICCIDS),* Gurugram, 2020.

[21] N. Saini, M. Singhal, M. Tanwar and P. Meel, “Multimodal, Semi-supervised and Unsupervised web content credibility analysis Frameworks,” in *IEEE 4th International Conference on Intelligent Computing and Control Systems (ICICCS),* Madurai , India, 2020.

[22] Y.-J. Lu and C.-T. Li, “GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* Online, 2020.

[23] Y. Wang, S. Qian, J. Hu, Q. Fang and C. Xu, “Fake News Detection via Knowledge-driven Multimodal Graph Convolutional Networks,” in *International Conference on Multimedia Retrieval (ICMR),* Dublin, Ireland, 2020.

[24] Y. Wang, W. Yang, F. Ma, J. Xu, B. Zhong, Q. Deng and J. Gao, “Weak supervision for fake news detection via reinforcement learning,” in *AAAI Conference on Artificial Intelligence,* New York, USA, 2020.

[25] Z. Wua, D. Pi, J. Chen, M. Xie and J. Cao, “Rumor detection based on propagation graph neural network with attention mechanism,” *Expert systems with applications,* vol. 158, 2020.

[26] P. Meel and D. K. Vishwakarma, “HAN, image captioning, and forensics ensemble multimodalfake news detection,” *Information Sciences,* vol. 567, pp. 23-41, 2021.

[27] C. Raj and P. Meel, “ConvNet frameworks for multi-modal fake news detection,” *Applied Intelligence,* pp. 1-17, 2021.

[28] V.-H. Nguyen, K. Sugiyama, P. Nakov and M.-Y. Kan, “FANG: Leveraging social context for fake news detection using graph representation,” in *29th ACM International Conference on Information & Knowledge Management,* Galway , Ireland, 2020.

[29] G. B. Guacho, S. Abdali, N. Shah and E. E. Papalexakis, “Semi-supervised Content-based Detection of Misinformation via Tensor Embeddings,” in *IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM),* Barcelona, Spain, 2018.

[30] J. K. Rout, A. Dalmia, K.-K. R. Choo, S. Bakshi and S. K. Jena, "Revisiting semi-supervised learning for online deceptive review detection," *IEEE Access,* vol. 5, pp. 1319-1327, 2017.

[31] R. Mansouri, M. Naderan-Tahan and M. J. Rashti, "A Semi-supervised Learning Method for Fake News Detection in Social Media," in *28th Iranian Conference on Electrical Engineering (ICEE),* Tabriz, Iran, 2020.

[32] X. Li, P. Lu, L. Hu, X. Wang and L. Lu, "A novel self-learning semi-supervised deep learning network to detect fake news on social media," *Multimedia Tools and Applications,* pp. 1-9, 2021.

[33] J. Gaglani, Y. Gandhi, S. Gogate and A. Halbe, "Unsupervised WhatsApp Fake News Detection using Semantic Search," in *IEEE 4th International Conference on Intelligent Computing and Control Systems (ICICCS).*, Madurai, India, 2020.

[34] S. Hosseinimotlagh and E. E. Papalexakis, "Unsupervised content-based identification of fake news articles with tensor decomposition ensembles," in *Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*, Los Angeles, California, USA, 2018.

[35] W. Chen, Y. Zhang, C. K. Yeo, C. T. Lau and B. S. Lee, "Unsupervised rumor detection based on users' behaviors using neural networks," *Pattern Recognition Letters,* vol. 105, pp. 226-233, 2018.

[36] S. Yang, K. Shu, S. Wang and F. W. ,. H. L. Renjie Gu, "Unsupervised Fake News Detection on Social Media: A generative approach," in *AAAI Conference on Artificial Intelligence*, Hawaii, USA, 2019.

[37] S. C. R. Gangireddy, D. P, C. Long and T. Chakraborty, "Unsupervised Fake News Detection: A Graph-based Approach," in *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, New Yark, USA, 2020.

[38] R. Singh, S. Bathla and P. Meel, "State-of-the-Art Applications of Graph Convolutional Neural Networks," in *6th International Conference on Recent Trends in Computing: ICRTC 2020*, Ghaziabad , India , 2020.

[39] T. Yao, Y. Pan, Y. Li and T. Mei, "Exploring visual relationship for image captioning," in *European conference on computer vision (ECCV),* Munich , Germany, 2018.

[40] C. Jeong, S. Jang, H. Shin, E. Park and S. Choi, "A context-aware citation recommendation model with BERT and graph convolutional networks," *Scientometrics* , vol. 124, no. 3, pp. 1907-1922, 2020.

[41] P. Kapanipathi, V. Thost, S. S. Patel, S. Whitehead, I. Abdelaziz, A. Balakrishnan, M. Chang, K. Fadnis, C. Gunasekara, B. Makni, N. Mattei, K. Talamadupula and A. Fokoue, "Infusing knowledge into the textual entailment task using graph convolutional networks," in *Thirty-Fourth AAAI Conference on Artificial Intelligence*, New York, 2020.

[42] Z.-M. Chen, X.-S. Wei, P. Wang and Y. Guo, "Multi-label image recognition with graph convolutional networks.," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, California , USA, 2019.

[43] Y. Chen, C. Liu, M. Liu and B. E. Shi, "Robot navigation in crowds by graph convolutional networks with attention learned from human gaze," *IEEE Robotics and Automation Letters,* vol. 5, no. 2, pp. 2754-2761, 2020.

[44] J. Pennington, R. Socher and C. D. Manning, “GloVe: Global Vectors forWord Representation,” in *conference on empirical methods in natural language processing (EMNLP)*, Doha , Qatar, 2014.

[45] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv:1609.02907*, 2016.

[46] “Fake News Data,” 2018. [Online]. Available: https://www.kaggle.com/c/fake-news/data.

[47] “real_or_fake,” 2018. [Online]. Available: https://www.kaggle.com/rchitic17/real-or-fake.

[48] “Fake News Detection,” 2017. [Online]. Available: https://www.kaggle.com/jruvika/fake-news-detection.

[49] A. P. S. Bali, M. Fernandez, S. Choubey, M. Goel and P. K. Roy, “Comparative Performance of Machine Learning Algorithms for Fake News Detection,” in *International Conference on Advances in Computing and Data Sciences , Springer ,* Singapore, 2019.

[50] K. Agarwalla, S. Nandan, V. A. Nair and D. D. Hema, “Fake News Detection using Machine Learning and Natural Language Processing,” *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 7, no. 6, pp. 844-847, 2019.

[51] H. Karimi and J. Tang, “Learning Hierarchical Discourse-level Structure for Fake News Detection,” in *The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, 2019.

[52] D. K. Vishwakarma, D. Varshney and A. Yadav, “Detection and veracity analysis of fake news via scrapping and authenticating the web search,” *Cognitive Systems Research*, vol. 58, pp. 217-229, 2019.

Noname manuscript No.
(will be inserted by the editor)

Friedmann cosmology with decaying vacuum density in Brans-Dicke theory

C. P. Singh · **Joan Solà Peracaula**

arXiv:2110.12411v1 [gr-qc] 24 Oct 2021

Received: date / Accepted: date

Abstract In this paper, we study Friedmann cosmology with time-varying vacuum energy density in the context of Brans-Dicke theory. We consider an isotropic and homogeneous flat space, filled with a matter-dominated perfect fluid and a dynamical cosmological term $\Lambda(t)$, obeying the equation of state of the vacuum. As the exact nature of a possible time-varying vacuum is yet to be found, we explore $\Lambda(t)$ given by the phenomenological law $\Lambda(t) = \lambda + \sigma H$, where λ and σ are positive constants. We solve the model and then focus on two different cases Λ_{H1} and Λ_{H2} by assuming $\Lambda = \lambda$ and $\Lambda = \sigma H$, respectively. Notice that Λ_{H1} is the analog of the standard ΛCDM, but within the Brans-Dicke cosmology. We find the analytical solution of the main cosmological functions such as the Hubble parameter, the scale factor, deceleration and equation of state parameters for these models. In order to test the viability of the cosmological scenarios, we perform two sets of joint observational analyses of the recent Type Ia supernova data (Pantheon), observational measurements of Hubble parameter data, Baryon acoustic oscillation/Cosmic microwave background data and Local Hubble constant for each model. For the sake of comparison, the same data analysis is performed for the ΛCDM model. Each model shows a transition from decelerated phase to accelerated phase and can be viewed as an effective quintessence behavior. Using the model selection criteria AIC and BIC to distinguish from existing dark energy models, we find that the Brans-Dicke analog of the Λ-cosmology (i.e. our model Λ_{H1}) performs at a level comparable to the standard ΛCDM, whereas Λ_{H2} is less favoured.

Keywords Cosmology · FLRW model · Dark energy · Observational data.

C. P. Singh
Department of Applied Mathematics, Delhi Technological University, Bawana Road, Delhi-110 042, India.
E-mail: cpsphd@rediffmail.com

Joan Solà Peracaula
Departament de Física Quàntica i Astrofísica, and Institute of Cosmos Sciences, Universitat de Barcelona, Av. Diagonal 647, E-08028 Barcelona, Catalonia, Spain
E-mail: sola@fqa.ub.edu

1 Introduction

The recent observational data from Type Ia supernova[1,2,3], cosmic microwave background radiation [4], galaxy clustering [5] and other cosmological observations [6,7,8,9, 10] suggest that our Universe is currently experiencing a phase of accelerated expansion. It has been learnt that the Universe is dominated by dark energy (DE) with negative pressure which provides the dynamical mechanism for the accelerating expansion of the Universe. However, the nature of this substance is still undetermined. The cosmological constant (CC), initially introduced by Einstein, is a natural candidate of DE. Such model is also known as standard Lambda-cold dark matter (ΛCDM) model. In it Λ =const. and sometimes it is referred to for short as the Λ-cosmology. In this scenario, DE is associated to the energy density of the quantum vacuum $\rho_\Lambda = \Lambda/8\pi G$. However, it faces a long-standing cosmological constant problem [11]. This CC problem stems from tremendous discrepancy between the theoretical value associated with quantum vacuum energy and the value required to confirm with observations. As a matter of fact, all sorts of cosmological models predict a large value of the DE and they require of an unnatural fine tuning to solve such discrepancy. In this sense the vacuum energy is not to be blamed more than many other DE models[12]. Although several possible approaches have been adopted to explain or alleviate the CC problems [13], there is no convincing fundamental theory for why vacuum energy dominance happened only recently and why its value is currently so close to the matter energy density (the so-called cosmic coincidence problem).

One possibility to mitigate certain aspects of the CC problem is to consider time evolving vacuum models, $\Lambda = \Lambda(t)$. A great deal of attention was dedicated to this possibility even before the discovering of the accelerating Universe [14,15,16,17,18]. $\Lambda(t)$ models may be an important alternative to the ΛCDM model. The original proposals were essentially phenomenological [18], but a new generation of proposals are theoretically better rooted. They are based on the idea that DE is the manifestation of vacuum quantum fluctuations in the curved space-time, after a renormalization in which the divergent vacuum contribution in the flat space-time is subtracted. The resulting effective vacuum energy density will depend on the space-time curvature, decaying from high initial values to smaller ones as the Universe expands. This idea underlies e.g. the class of running vacuum models [19,20] (see [12] for a review and references therein) and it has been supported by recent calculations in quantum field theory in curved spacetime [21].

In recent years, a large class of flat non-singular Friedmann-Robertson-Walker type cosmologies, where the vacuum energy density evolves like a truncated power-series in the Hubble parameter H, have been discussed in the literature [22,23,24,25,26,27,28,29,30, 31,32,33]. The functional form of $\Lambda(t)$ in most of them has usually been proposed on phenomenological grounds as it occurs with the vast majority of DE models. In this regard, a viable form of decaying vacuum energy density, namely $\Lambda(a) \propto H$, was proposed by Schützhold [22]. Such proposal was subsequently extended in the literature in the form of the the so-called ghost dark energy models [35,36,37]. These models together with the aforementioned class of running vacuum models [12,34] both use expansions of the vacuum energy density in powers of H, but of a different kind. These examples show the significant interest raised by the dynamical dark energy models from different perspectives.

In the present paper, we focus our attention on the analytical and observational aspects of the $\Lambda(t)$ models in the scalar tensor theory proposed by Brans and Dicke [38]. Brans-Dicke (BD) theory was the first gravity theory in which the dynamics of gravity were described by a scalar field while spacetime dynamics were represented by the metric tensor. In this theory, the gravitational constant G is replaced with a inverse of time-dependent scalar field

ϕ, which couples to gravity with a coupling parameter ω (Brans-Dicke parameter). This theory passed the experimental tests from the solar system [39]. In recent years, this theory got a new impetus as it arises naturally as the low energy limit of many theories of quantum gravity such as superstring theory or Kaluza-Klein theory. An attractive feature of BD theory is that the scalar field is a fundamental element of the theory, quite contrary to other models in which the scalar field is introduced separately in an ad hoc manner. The studies on Friedmann-Robertson-Walker model in the framework of BD theory have been carried out in Refs.[42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61].

The aim of the work is to find a cosmological scenario of the model in BD theory with varying cosmological term which would be capable to link the dynamics of the early Universe with that of our late Universe. This work extends the successful approach recently presented on BD cosmology with a rigid cosmological term [62,63], and reinforces the idea that dynamical models of the vacuum energy can be very helpful to improve the fit to cosmological and cosmographical observations [64,65,66,67,68]. In particular, they help alleviating the so-called H_0 and σ_8 tensions[69,70,71], see e.g. [72,73] and the very recent work [74]. In the present study, we compare the consequent cosmological scenario with the constraints imposed by the observational data of Type Ia supernova (Pantheon), observational Hubble parameter data, baryon acoustic oscillations data/cosmic microwave background and local H_0. The analysis of Hubble- redshift relation has shown a good fit with best fit values of the model parameters.

The structure of the paper is as follows. In Section 2, we introduce the basic cosmological equations. The solution of the field equations is presented in Section 3 with time varying cosmological constant. Section 4 describes and places constraints on the main parameters of our vacuum models by performing two sets of joint likelihood analysis consisting of Type Ia supernova (Pantheon) data, the observational Hubble parameter data (OHD), baryonic acoustic oscillations/ cosmic microwave background (BAOs/CMB) data and local H_0. Section 5 is divided in subsections. In Subsection 5.1, we discuss the evolution of the cosmological parameters using fitting values and in subsection 5.2, the model selection criterion is discussed. Finally, in Section 6, we present the summary of the work.

2 BD field equations with time-dependent vacuum

The action for BD theory extended to the cosmological constant (CC) in Jordan frame reads as follows [75,76,77].

$$S=\int d^4x\sqrt{-g}\left[\frac{1}{16\pi}\left(\phi R-\frac{\omega}{\phi}\nabla_\alpha\phi\,\nabla^\alpha\phi\right)-\rho_\Lambda+\mathscr{L}_m\right], \tag{1}$$

where ϕ is the BD scalar field representing the inverse of the Newton constant, which is allowed to vary with space and time, ω is the dimensionless constant which is known as a coupling parameter, or Brans-Dicke parameter, of the theory and $\mathscr{L}_m$ is the matter Lagrangian. It is to be noted that there is no potential for BD scalar field ϕ in the original BD theory, however, we admit the presence of CC term associated with vacuum energy density, ρ_Λ.

Variation of this action with respect to the metric $g_{\mu\nu}$ and the BD scalar field ϕ yield the following field equations, respectively.

$$G_{\mu\nu}=R_{\mu\nu}-\frac{1}{2}g_{\mu\nu}R=\frac{8\pi}{\phi}\left(T^m_{\mu\nu}-g_{\mu\nu}\rho_\Lambda\right)+\frac{8\pi}{\phi}T^{BD}_{\mu\nu} \tag{2}$$

and

$$\Box\phi = \frac{8\pi}{(2\omega+3)}\left(T_{\mu}^{m\,\mu} - 4\rho_{\Lambda}\right), \tag{3}$$

where $T^{m}_{\mu\nu}$ is the the energy-momentum tensor of matter and $T_{\mu}^{m\,\mu}$ is the trace of $T^{m}_{\mu\nu}$, and other symbols have their usual meaning. It is convenient to introduce the effective energy-momentum tensor for the two fluids, matter and vacuum energy density through $\tilde{T_{\mu\nu}} = T_{\mu\nu} - g_{\mu\nu}\rho_{\Lambda}$ and adopts the perfect fluid form:

$$\tilde{T}_{\mu\nu} = (\rho + p)u_{\mu}u_{\nu} + pg_{\mu\nu}, \tag{4}$$

where $\rho = \rho_m + \rho_{\Lambda}$ and $p = p_m + p_{\Lambda}$. We assume that the matter part contains the pressureless contribution of cold dark matter. The vacuum energy density ρ_{Λ} follows the usual equation of state (EoS) as $p_{\Lambda} = -\rho_{\Lambda}$. Also, $T^{BD}_{\mu\nu}$ is the energy-momentum for the BD scalar which is defined by

$$T^{BD}_{\mu\nu} = \frac{1}{8\pi}\Big[\frac{\omega}{\phi}\left(\nabla_{\mu}\phi\nabla_{\nu}\phi - \frac{1}{2}g_{\mu\nu}\nabla_{\alpha}\phi\nabla^{\alpha}\phi\right) + \nabla_{\mu}\nabla_{\nu}\phi - g_{\mu\nu}\nabla_{\alpha}\nabla^{\alpha}\phi\Big]. \tag{5}$$

Let us start with the homogeneous and isotropic flat Friedmann-Lemaître-Robertson-Walker (FLRW) line element

$$ds^2 = -dt^2 + a^2(t)\left[dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2)\right], \tag{6}$$

where $a(t)$ is the scale factor of the universe. Throughout we use units such that the speed of light, $c = 1$.

The field equations (2) and (3) for metric (6) and energy-momentum tensors (4) and (5) are simplified to

$$3H^2 + 3H\frac{\dot{\phi}}{\phi} - \frac{\omega}{2}\frac{\dot{\phi}^2}{\phi^2} = \frac{8\pi}{\phi}\rho, \tag{7}$$

$$2\dot{H} + 3H^2 + \frac{\ddot{\phi}}{\phi} + 2H\frac{\dot{\phi}}{\phi} + \frac{\omega}{2}\frac{\dot{\phi}^2}{\phi^2} = -\frac{8\pi}{\phi}p, \tag{8}$$

$$\ddot{\phi} + 3H\dot{\phi} = \frac{8\pi}{(2\omega+3)}(\rho - 3p). \tag{9}$$

where an overdot denotes derivative with respect to cosmic time t and $H = \dot{a}/a$ is the Hubble parameter. The first equation (7) corresponds to the Friedmann equation and the second equation (9) is the equation of motion of the BD scalar field.

If we ignore the inhomogeneities arising from the (linear) field perturbations, the BD field can be treated as a perfect fluid $T^{BD}_{\mu\nu} = (\rho_{BD} + p_{BD})u_{\mu}u_{\nu} + p_{BD}g_{\mu\nu}$ with energy and pressure are respectively given by

$$\rho_{BD} = \frac{1}{8\pi}\left[\frac{\omega}{2}\left(\frac{\dot{\phi}^2}{\phi}\right) - 3H\dot{\phi}\right], \tag{10}$$

$$p_{BD} = \frac{1}{8\pi}\left[\frac{\omega}{2}\left(\frac{\dot{\phi}^2}{\phi}\right) + 2H\dot{\phi} + \ddot{\phi}\right]. \tag{11}$$

Finally, the geometric Bianchi identity of $\nabla_{\nu}G^{\mu\nu} = 0$ in Eq.(2), which plays a role of the consistency relation, leads to

$$\nabla_{\nu}\left(R^{\mu\nu} - \frac{1}{2}g^{\mu\nu}R\right) = 0 = \nabla_{\nu}\left(\frac{8\pi}{\phi}\tilde{T}^{\mu\nu} + \frac{8\pi}{\phi}T^{\mu\nu}_{BD}\right). \tag{12}$$

One interesting thing about working in Jordan frame is that the conservation equation holds for matter and scalar field separately, i.e., equations of motion of matter do not enter into the BD scalar field. It means that $\tilde{T}^{\mu\nu}$ obeys the usual conservation law, $\nabla_\nu \tilde{T}^{\mu\nu} = 0$, which takes the form

$$\dot{\rho_m} + 3(\rho_m + p_m)\frac{\dot{a}}{a} = -\dot{\rho_\Lambda}. \tag{13}$$

In this paper, we study the model dominated by pressureless dark matter ($p_m = 0$) in BD theory. It is to be noted that the equation of state of the vacuum energy density maintains the usual form $p_\Lambda(t) = -\rho_\Lambda(t) = -\phi\Lambda(t)/8\pi$ despite the fact that $\Lambda(t)$ evolves with time. Now, from (12) and because of matter conservation, we are left with[1]

$$\left(\nabla_\nu\{\frac{8\pi}{\phi}\}\tilde{T}^{\mu\nu} + \nabla_\nu\{\frac{8\pi}{\phi}T^{\mu\nu}_{BD}\}\right) = 0, \tag{14}$$

which finally gives

$$\dot{\rho}_{BD} + 3\frac{\dot{a}}{a}(\rho_{BD} + p_{BD}) = \left(\frac{\dot{\phi}}{\phi}\right)(\rho + \rho_{BD}). \tag{15}$$

Equation (15) is indeed a consistency condition originating from the Bianchi identity $\nabla_\nu G^{\mu\nu} = 0$. It can indeed be checked that the covariant conservation laws (13) and (15) can also be obtained upon lengthy but straightforward computation by combining Eqs. (7) - (9), which are identical to that of general relativity (GR). Although the calculation is more involved than in GR, the final result turns out to be the same. Thus, we shall use (7) and (13) to obtain the solution of the model and finally we use (15) to get the consistency condition using the fitting values of the model parameters obtained from observational data (to be discussed in Sect.4).

In the framework of BD cosmology the BD scalar field ϕ one usually searches for power-law relation in terms of scale factor [78, 79, 80, 81, 82, 83], namely[2]

$$\phi = \phi_0\, a(t)^\varepsilon, \tag{16}$$

where ϕ_0 and ε are constants. A case of particular interest is that when ε is small whereas ω is large so that the product $\omega\varepsilon$ results of order unity [79, 81, 82]. This is interesting because local experiments set a very high lower bound on ω. This choice with small ε can lead to consistent results which may justify this specific choice among other possible choices [79]. The Cassini experiment [39] implies that $\omega > 10^4$. Likewise, as previously indicated, a slow fractional variation of ϕ will lead to a small fraction variation of G, consistent with observations. Therefore, it is clear that the interesting case is that one when ε is small whereas ω is large, so that the product $\omega\varepsilon$ results of order unity. As for ω, it is usually assumed large, but we do not find ω so large because we assume that the Cassini bound on this parameter only applies to the astrophysical domain, not to the cosmological one. This is admissible because of the possible existence of screening effects (chamaleon etc) which can operate in the local

[1] It may be illustrative to point out here that if one would define the energy density and pressure of the BD field in a different form, namely in such a way that they would represent the exact departure of the BD theory from GR, then the corresponding tensor associated to these new quantities $T^{\mu\nu}_{BD}$ would be locally and covariantly conserved as it does in the case of matter, namely $\nabla_\mu T^{\mu\nu}_{BD} = 0$. This alternative formulation has been used in [62, 63].

[2] While there is no a priori reason to assume that a power-law solution is viable, let us notice that the fractional variation of the effective gravitational coupling $G = \frac{1}{\phi}$ in BD theories is given by $\frac{\dot{G}}{G} = -\frac{\dot{\phi}}{\phi} = -\varepsilon H$. Therefore, for sufficiently small $|\varepsilon|$, this is consistent with the bounds on the time variation of the gravitational coupling [84]. Other partial justifications can be checked a posteriori, as we shall see.

domain. These effects do not apply at the cosmological level and permit a discussion of the BD framework free from the stringent Cassini bounds [40,41]. See e.g. [63] for a detailed discussion.

We also note that in Ref.[81,82] this kind of power-law solution is used to show that BD cosmology with a cosmological term can mimic the running vacuum model, which is very convenient in order to improve the fitting of the cosmological data [82]. We expect that this feature will also help here.

With the above assumption, the Eq.(7) is rewritten as

$$H^2 = \frac{2}{(6+6\varepsilon-\omega\varepsilon^2)}\frac{8\pi}{\phi}(\rho_m+\rho_\Lambda), \tag{17}$$

where $H = \dot{a}/a$ is the Hubble parameter and $\Lambda = 8\pi\rho_\Lambda/\phi$. It can be observed that in the limit of $\varepsilon \to 0$, the standard cosmology is recovered. To make the Eq.(17) to have physical meaning, i.e., to make $(6+6\varepsilon-\omega\varepsilon^2) > 0$, one has the following constraint on the value of ε which is given by $\frac{3-\sqrt{9+6\omega}}{\omega} < \varepsilon < \frac{3+\sqrt{9+6\omega}}{\omega}$, where $\omega > 0$.

Finally, combining equations (13) and (17), we find

$$\dot{H} + \frac{(3+\varepsilon)}{2}H^2 = \frac{3\Lambda}{(6+6\varepsilon-\omega\varepsilon^2)}. \tag{18}$$

In what follows, we investigate the cosmic evolution with a class of time evolving vacuum models. Notice that up to this point the above equations are valid for any Λ, not necessarily a constant, it can be a function of the cosmic time. Recall that the equation of state remains $p_\Lambda(t) = -\rho_\Lambda(t)$. In the next sections, however, we specify some possible forms.

3 Brans-Dicke theory with time-varying Λ

In this paper we parameterize the functional form of $\Lambda(t)$ as a combination of constant term and some multiple of the Hubble parameter, i.e.,

$$\Lambda(t) = \lambda + \sigma H, \tag{19}$$

where λ and σ are positive constants. The model with $\lambda = 0$, hence $\Lambda \propto H$, was discussed in Refs.[22,24,25,26], but only in the context of GR. In [81,82] this case was studied from the point of view of its ability to emulate the running vacuum model. In the following subsections, we study the two extreme situations $(\lambda \neq 0, \sigma = 0)$ and $(\lambda = 0, \sigma \neq 0)$ in the BD framework and perform the corresponding observational analysis. The detailed solution of the general class of models (19) for any value of λ and σ is given in an Appendix.

3.1 $\Lambda_{H1}-$ model: the standard Λ cosmology in BD theory

In this section, we consider the $\Lambda =$ const. cosmology in the context of BD theory in order to appreciate the differences with respect to the $\Lambda(t)$ model explored subsequently. Assuming $\sigma = 0$ in Eq. (19), we have $\Lambda(t) = \lambda = const.$ (hereafter $\Lambda_{H1}-$ model). Thus, the vacuum term in (19) is constant and given by

$$\Lambda_0 = \lambda = 3\Omega_\Lambda H_0^2, \tag{20}$$

where Λ_0, H_0 and Ω_Λ are the current value of vacuum energy density, Hubble parameter and density parameter at present epoch $t = t_0$, respectively.

Using (20), the evolution equation (18) reads

$$\frac{d\,h^2}{dx} + (3+\varepsilon)h^2 = \frac{18\,\Omega_\Lambda}{(6+6\varepsilon-\omega\varepsilon^2)}, \tag{21}$$

where $h = H/H_0$ is the dimensionless Hubble parameter and $x = \ln a$. Solving Eq.(21), we obtain the Hubble function in terms of redshift z as

$$H(z) = H_0\left[\frac{18\Omega_\Lambda}{(6+6\varepsilon-\omega\varepsilon^2)(3+\varepsilon)} + \left(1 - \frac{18\Omega_\Lambda}{(6+6\varepsilon-\omega\varepsilon^2)(3+\varepsilon)}\right)(1+z)^{(3+\varepsilon)}\right]^{1/2}, \tag{22}$$

where $(1+z) = a_0/a$. We can check that $\varepsilon = 0$ reduces Eq. (22) to the corresponding equation in ΛCDM model as expected. We can define the normalized Hubble expansion as a function of redshift

$$h^2(z) = \frac{H^2(z)}{H_0^2} = \tilde{\Omega_{\Lambda 1}} + \tilde{\Omega_{m1}}(1+z)^{(3+\varepsilon)}, \tag{23}$$

where we have used the following parametrization

$$\tilde{\Omega_{\Lambda 1}} = 1 - \tilde{\Omega_{m1}} = \frac{18\,\Omega_\Lambda}{(6+6\varepsilon-\omega\varepsilon^2)(3+\varepsilon)} \tag{24}$$

As expected, for $\varepsilon \to 0$, we have $\tilde{\Omega_{\Lambda 1}} \sim \Omega_\Lambda$. Thus, the traditional cosmology is a particular solution of the Λ_{H1} model with ε strictly equal to zero.

The scale factor of the Universe, normalized to unity at the present epoch, is given by

$$a_\Lambda(t) = \left(\frac{\tilde{\Omega_{m1}}}{\tilde{\Omega_{\Lambda 1}}}\right)^{1/(3+\varepsilon)} \left[\sinh\left(\frac{(3+\varepsilon)\sqrt{\tilde{\Omega_{\Lambda 1}}}}{2}\,H_0 t\right)\right]^{2/(3+\varepsilon)}. \tag{25}$$

The Hubble parameter in terms of cosmic time t is given by

$$H(t) = H_0\sqrt{\tilde{\Omega_{\Lambda 1}}}\,\coth\left(\frac{(3+\varepsilon)\sqrt{\tilde{\Omega_{\Lambda 1}}}}{2}\,H_0 t\right). \tag{26}$$

The cosmic time is related with the scale factor as

$$t_\Lambda(a) = \frac{2}{(3+\varepsilon)\sqrt{\tilde{\Omega_{\Lambda 1}}}H_0}\,\sinh^{-1}\left(\sqrt{\frac{\tilde{\Omega_{\Lambda 1}}}{\tilde{\Omega_{m1}}}}\,a^{(3+\varepsilon)/2}\right) \tag{27}$$

The current age of the Universe is given by

$$t_{0\Lambda} = \frac{2}{(3+\varepsilon)\sqrt{\tilde{\Omega_{\Lambda 1}}}H_0}\,\sinh^{-1}\left(\sqrt{\frac{\tilde{\Omega_{\Lambda 1}}}{\tilde{\Omega_{m1}}}}\right). \tag{28}$$

From (22), one can deduce the deceleration parameter q, which is defined as $q = -a\ddot{a}/\dot{a}^2$. It is given by

$$q(z) = -1 + \frac{\frac{(3+\varepsilon)}{2}\tilde{\Omega_{m1}}\,(1+z)^{(3+\varepsilon)}}{\tilde{\Omega_{\Lambda 1}} + \tilde{\Omega_{m1}}(1+z)^{(3+\varepsilon)}}. \tag{29}$$

The value of q at present time ($z=0$) is given by

$$q(z=0) = -1 + \left(\frac{3+\varepsilon}{2}\right)\tilde{\Omega_{m1}}. \tag{30}$$

The transition from deceleration to acceleration takes place for $\tilde{\Omega_{m1}} = 2/(3+\varepsilon)$. For any value $\tilde{\Omega_{m1}} < 2/(3+\varepsilon)$, the present-day cosmic expansion is accelerating. Now, it is also possible to find the transition redshift z_{tr} at which the Universe transits from deceleration to acceleration, i.e.,

$$z_{tr} = \left[\frac{2}{(1+\varepsilon)}\frac{\tilde{\Omega_{\Lambda 1}}}{\tilde{\Omega_{m1}}}\right]^{1/(3+\varepsilon)} - 1 \tag{31}$$

For $\varepsilon = 0$ we can see that we recover the corresponding value in the concordance Λ-cosmology with GR, as expected. The transition time is calculated as

$$t_{tr} = \frac{2}{(3+\varepsilon)\sqrt{\tilde{\Omega_{\Lambda 1}}}H_0}\sinh^{-1}\left(\sqrt{\frac{1+\varepsilon}{2}}\right). \tag{32}$$

Let us calculate the effective equation of state (EoS) parameter w_{eff} for the compound fluid of the model. An accelerated expansion of the Universe is possible only if the effective EoS parameter w_{eff} satisfies $3w_{eff}+1<0$. The effective EoS parameter can be obtained by

$$w_{eff} = -1 - \frac{1}{3}\frac{d(\ln h^2)}{dx}, \tag{33}$$

where $x = ln\, a$. Using (23), the EoS parameter (33) is calculated as

$$w_{eff} = -1 + \frac{(3+\varepsilon)\tilde{\Omega_{m1}}}{3h^2}(1+z)^{(3+\varepsilon)}. \tag{34}$$

The ε-dependent part represents the departure from the GR result. It is easy to check that the above EoS can be written as follows:

$$w_{eff} = -1 + \frac{3+\varepsilon}{3}\frac{1}{1+ra^{3+\varepsilon}}, \tag{35}$$

where we have defined the ratio $r = \tilde{\Omega_{\Lambda 1}}/\tilde{\Omega_{m1}}$. One can observe that $w_{eff} \to -1$ in the late time ($a \gg 1$), whereas $w_{eff} \to \frac{\varepsilon}{3}$ in the remote past ($a \to 0$). Therefore, the model corresponds to de Sitter in future time and performs a transition from a situation of essentially matter dominance ($w_{eff} \simeq 0$) into a future one of vacuum dominance ($w_{eff} = -1$). The EoS does not cross the phantom divide line $w \leq -1$ which shows that the Λ_{H2} model is free from big-rip singularity. The present value of EoS parameter is obtained by setting $a = 1$ ($z = 0$) in the above equation:

$$w_{eff}(z=0) = -1 + \frac{(3+\varepsilon)\tilde{\Omega_{m1}}}{3}, \tag{36}$$

where $h(z) = 1$ at $z = 0$. Therefore, the condition for $3w_{eff}(z=0)+1<0$ implies that $\tilde{\Omega_{m1}} < 2/(3+\varepsilon)$.

Let us check the consistency condition (15) for the solution of this model. Using (10), (11) and (17), Eq.(15) can be rewritten as[3]

$$2(\omega\varepsilon-3)\dot{H}+(\omega\varepsilon^2+6\omega\varepsilon-12)H^2=0 \tag{37}$$

Now, using the solution of H obtained in Eq.(22) into (37), we obtain

$$(\omega\varepsilon-3)(3+\varepsilon)(1-\tilde{\Omega_{\Lambda 1}})a^{-(3+\varepsilon)}-(\omega\varepsilon^2+6\omega\varepsilon-12)[\tilde{\Omega_{\Lambda 1}}+(1-\tilde{\Omega_{\Lambda 1}})a^{-(3+\varepsilon)}]=0. \tag{38}$$

The above equation gives a relation between the constants for $a=a_0=1$, i.e., at present, which is given by

$$(\omega\varepsilon-3)(3+\varepsilon)(1-\tilde{\Omega_{\Lambda 1}})-(\omega\varepsilon^2+6\omega\varepsilon-12)=0. \tag{39}$$

It is to be noted that one can use this equation for consistency checkup and not for constraining the parameters. In Section 5, we will present a detail discussion of the solutions obtained in this section by using best fit values. We will check explictly the above consistency condition.

3.2 $\Lambda_{H2}-$ model: $\Lambda=\sigma H$ in BD theory

In this Section, we assume that the vacuum term is proportional to the Hubble parameter. This kind of cosmological model is a particular case of Eq.(19) by setting $\lambda=0$ which is given by (hereafter, Λ_{H2} model)

$$\Lambda(t)=\sigma H \tag{40}$$

In Refs. [22,24,25,26], the authors have studied the FLRW model with Eq.(40) to describe the late time evolution of the Universe in the context of GR. In this paper, our aim is to study the FLRW model with this ansatz in a more dynamical framework of scalar -tensor theory as described by the Brans and Dicke theory.

Utilizing this ansatz at the present epoch $\Lambda_0=\sigma H_0$ and taking into account that the current value of the vacuum energy density is $\Lambda_0=3\Omega_\Lambda H_0^2$, the parameter σ is obtained to be

$$\sigma=3\Omega_\Lambda H_0. \tag{41}$$

Using (40), the evolution equation (18) reduces to

$$\frac{dh}{dx}+\frac{(3+\varepsilon)}{2}h=\frac{3\sigma}{(6+6\varepsilon-\omega\varepsilon^2)H_0}, \tag{42}$$

The solution of (42) is given by

$$h=\frac{6\sigma}{(6+6\varepsilon-\omega\varepsilon^2)(3+\varepsilon)H_0}+\left(1-\frac{6\sigma}{(6+6\varepsilon-\omega\varepsilon^2)(3+\varepsilon)H_0}\right)(1+z)^{\frac{(3+\varepsilon)}{2}}. \tag{43}$$

Using (41) into (43), the Hubble parameter in terms of redshift can be given by

$$H=H_0\left[\frac{18\Omega_\Lambda}{(6+6\varepsilon-\omega\varepsilon^2)(3+\varepsilon)}+\left(1-\frac{18\Omega_\Lambda}{(6+6\varepsilon-\omega\varepsilon^2)(3+\varepsilon)}\right)(1+z)^{\frac{(3+\varepsilon)}{2}}\right], \tag{44}$$

[3] It is important to keep in mind that the exact consistency condition is actually ε times Eq. (37). By dividing out the exact equation by ε we are assuming that $\varepsilon\neq 0$, as in fact it is the case in our fitting result (see Table 1). However, the presence of the additional factor of ε shows that Eq. (37) has a smooth limit to GR (for which $\varepsilon=0$ exactly) and hence no such equation remains in that limit. Recall that the primary origin of the consistency condition is Eq. (15), which disappears of course when there is no BD fluid (10)-(11) .

We can define the normalized Hubble expansion as a function of redshift

$$h = \frac{H(z)}{H_0} = \tilde{\Omega_{\Lambda 2}} + \tilde{\Omega_{m2}}(1+z)^{\frac{(3+\varepsilon)}{2}}, \tag{45}$$

where we have used the following parametrization

$$\tilde{\Omega_{\Lambda 2}} = 1 - \tilde{\Omega_{m2}} = \frac{18\,\Omega_\Lambda}{(6+6\varepsilon-\omega\varepsilon^2)(3+\varepsilon)}. \tag{46}$$

Thus, the Hubble rate of Λ_{H2} is very different from Eq. (23) of Λ_{H1} model, which could have the different behavior when we do the observation. It is to be noted that in the absence of BD theory ($\varepsilon = 0$), Eq. (45) reduces to the Eq. (70) of Ref.[28]. The scale factor in normalized unit is given by

$$a(t) = \left[\frac{e^{\frac{(3+\varepsilon)}{2}\tilde{\Omega_{\Lambda 2}}H_0\,t} - 1 + \tilde{\Omega_{\Lambda 2}}}{\tilde{\Omega_{\Lambda 2}}}\right]^{2/(3+\varepsilon)}. \tag{47}$$

From Eq.(47), it is observed that for small times (small compared to the present time), it can be approximated by

$$a(t) \sim \left(1 + \frac{(3+\varepsilon)}{2}H_0\,t\right)^{2/(3+\varepsilon)}, \tag{48}$$

which has the same time dependence as in the standard flat BD model with dust. Therefore, the Λ_{H2} model expands with decelerated rate in early time. The model predicts the Big-Bang in the past at the cosmic time: $t_{BB} = \frac{2H_0^{-1}}{(3+\varepsilon)\,\tilde{\Omega_{\Lambda 2}}}\,\ln(1-\tilde{\Omega_{\Lambda 2}})$. The varying $\Lambda(t)$ starts dominating just at present time. In the limit of large times, that is, $(3+\varepsilon)\tilde{\Omega_{\Lambda 2}}H_0\,t \gg 1$ and $a \to \infty$, Eq.(47) leads to $\exp(\tilde{\Omega_{\Lambda 2}}H_0\,t)$, that is, the model tends to a de Sitter Universe.

The Hubble parameter in terms of cosmic time t is given by

$$H(t) = \frac{H_0\,\tilde{\Omega_{\Lambda 2}}\,e^{\frac{(3+\varepsilon)}{2}\tilde{\Omega_{\Lambda 2}}H_0\,t}}{e^{\frac{(3+\varepsilon)}{2}\tilde{\Omega_{\Lambda 2}}H_0\,t} - 1 + \tilde{\Omega_{\Lambda 2}}}. \tag{49}$$

In this model, the cosmic time is related to the scale factor as

$$t(a) = \frac{2H_0^{-1}}{(3+\varepsilon)\,\tilde{\Omega_{\Lambda 2}}}\,\ln[1+\tilde{\Omega_{\Lambda 2}}(a^{2/(3+\varepsilon)}-1)]. \tag{50}$$

It is straightforward to calculate the deceleration parameter in terms of redshift which takes the following form:

$$q = -1 + \frac{(3+\varepsilon)\tilde{\Omega_{m2}}\,(1+z)^{(3+\varepsilon)/2}}{2\left[\tilde{\Omega_{\Lambda 2}} + \tilde{\Omega_{m2}}(1+z)^{(3+\varepsilon)/2}\right]}. \tag{51}$$

Note that for $z = 0$, one finds the current value of deceleration parameter

$$q_0 = -1 + \frac{(3+\varepsilon)\tilde{\Omega_{m2}}}{2}. \tag{52}$$

Now, it is easy to check that the transition redshift, defined to be the zero point of the deceleration parameter, is given by

$$z_{tr} = \left(\frac{2\tilde{\Omega_{\Lambda 2}}}{(1+\varepsilon)\tilde{\Omega_{m2}}}\right)^{2/(3+\varepsilon)} - 1. \tag{53}$$

For $\varepsilon = 0$ we can see that on this occasion we do not recover the corresponding value in the Λ-cosmology with GR. Compare, in contrast, with Eq.(31) of the $\Lambda_{H1}-$ model. This is because model Λ_{H2} does not have a smooth connection with the concordance ΛCDM owing to the absence of a rigid cosmological term in it.

The inflection point takes place at

$$t_{tr} = \frac{2H_0^{-1}}{(3+\varepsilon)\tilde{\Omega_{\Lambda 2}}} \ln\left(\frac{(3+\varepsilon)\tilde{\Omega_{m2}}}{2}\right). \tag{54}$$

The effective EoS for this model is given by

$$w_{eff} = -1 + \frac{(3+\varepsilon)\tilde{\Omega_{m2}}}{3h}\,(1+z)^{(3+\varepsilon)/2}. \tag{55}$$

We can observe that $w_{eff} \to -1$ in the late time. Therefore, the model corresponds to de Sitter in future time. The EoS does not cross the phantom divide line $w \leq -1$ which shows that the Λ_{H2} model is free from big-rip singularity. The present-day value of w_{eff} is obtained as

$$w_{eff} = -1 + \frac{(3+\varepsilon)\tilde{\Omega_{m2}}}{3} \tag{56}$$

From Eq. (56), we can observe that the condition $3w_{eff}+1 < 0$ gives $\tilde{\Omega_{m2}} < 2/(3+\varepsilon)$. In particular, if $\varepsilon = 0$, we recover the expression of all the cosmological parameters obtained in Ref.[25], where the model was treated within GR.

Now, using the solution of H obtained in Eq.(44) into consistent eq.(37), we obtain

$$(\omega\varepsilon - 3)(3+\varepsilon)(1-\tilde{\Omega_{\Lambda 2}})a^{-\frac{(3+\varepsilon)}{2}} - (\omega\varepsilon^2 + 6\omega\varepsilon - 12)[\tilde{\Omega_{\Lambda 2}} + (1-\tilde{\Omega_{\Lambda 2}})a^{-\frac{(3+\varepsilon)}{2}}] = 0. \tag{57}$$

The above equation gives a relation between the constants for $a = a_0 = 1$, i.e., at present, which reads

$$(\omega\varepsilon - 3)(3+\varepsilon)(1-\tilde{\Omega_{\Lambda 2}}) - (\omega\varepsilon^2 + 6\omega\varepsilon - 12) = 0. \tag{58}$$

Again, in this model, one can use this equation for consistency checkup and not for constraining the parameters.

4 Parameter estimation

In this section, we present the cosmic observations on the free parameters of Λ_{H1} and Λ_{H2} models. To this end we will use two joint observational set of data, as described below. We perform the goodness-of-fit of the models using Markov Chain Monte Carlo (MCMC) method by employing EMCEE python package [85]. We also perform the model selection criteria to determine favoured model. In what follows, we discuss the observational data which are to be used to constraint the parameters of the models.

4.1 Hubble data

We use the Hubble data comprising of 36 measurements which includes 31 measurements from cosmic chronometric (CC) method [86], three correlated measurements from BAO signal in galaxy distribution [87], and lastly two measurements from BAO signal in Ly-α forest distribution alone or cross-correlated with quasistellar objects (QSOs) [88,89].

Thus, the chi-square function corresponding to 33 measurements of CC and Ly-α is defined as

$$\chi^2_{CC+Ly\alpha} = \sum_{i=1}^{33} \frac{[H_{obs}(z_i) - H_{th}(z_i)]^2}{\sigma_i^2}, \tag{59}$$

where $H_{th}(z_i)$ and $H_{obs}(z_i)$ represents theoretical and observed values, respectively, and σ_i^2 is the standard deviation of each $H_{obs}(z_i)$ as given in Table 2 of Ref.[90].

Further, the chi-squared corresponds to the 3 galaxy distribution measurements is given by

$$\chi^2_{gal} = A^T C^{-1} A \tag{60}$$

where C is the covariance matrix given by [87]

$$C = \begin{bmatrix} 3.65 & 1.78 & 0.93 \\ 1.78 & 3.65 & 2.20 \\ 0.93 & 2.20 & 4.45 \end{bmatrix}$$

and

$$A = \begin{bmatrix} H_{obs}(0.38) - H_{th}(0.38) \\ H_{obs}(0.51) - H_{th}(0.51) \\ H_{obs}(0.61) - H_{th}(0.61) \end{bmatrix}$$

Thus, the combined χ^2 function for Hubble data is given by

$$\chi^2_H = \chi^2_{CC+Ly\alpha} + \chi^2_{gal}. \tag{61}$$

4.2 Type Ia supernovae (Pantheon data)

We use the Pantheon sample, the latest compilation of Type Ia supernovae (SNe) comprising of 40 binned data points in the redshift region $z \in [0.014, 1.62]$ [91].

The χ^2 function of the Pantheon SNe data is given by

$$\chi^2_{SNe(Pan)} = \Delta\mu^T \cdot C^{-1} \cdot \Delta\mu \tag{62}$$

in which $\Delta\mu = \mu_i^{obs} - \mu^{th}$, where μ_i^{obs} is the observed distance modulus defined in Ref.[92] and μ^{th}, the theoretical distance modulus that depends on redshift and the cosmological parameters, is given by

$$\mu^{th} = 5\ \log_{10}[d_L(z)/10\ pc] + \mathscr{M}, \tag{63}$$

where $\mathscr{M}$ is the nuisance parameter. The quantity d_L, known as the dimensionless luminosity distance, is given by [91]

$$d_L(z) = (1+z)c \int_0^z \frac{dz'}{H(z',\theta)}, \tag{64}$$

where θ represents the set of model parameters and c is the speed of light.

It should be noted that the covariance matrix C in (62) is the sum of the systematic covariance C_{sys} and statistical matrix D_{stat} having a diagonal component [91,93].

4.3 BAO/CMB data set

We use the combined baryon acoustic oscillation and cosmic microwave background (BAO/CMB) data from different observational missions [94]. We have taken the sample of BAO distances measurements from SDSS(R) [95], the 6dF Galaxy survey [96], BOSS CMASS [97] and three parallel measurements from WiggleZ survey [98]. We combine theses results with the Planck 2015 [99].

We use measurements derived from the product of the CMB acoustic scale, and from the ratio of the BAO dilation scale to the sound horizon scale at the drag epoch. Thus, we can write the χ^2 function as [94]

$$\chi^2_{BAO/CMB} = A^T C^{-1} A \tag{65}$$

where C^{-1} is the inverse of the covariance matrix [94] and A is the matrix

$$A = \begin{bmatrix} \frac{d_A(z_*,\theta)}{D_v(0.106,\theta)} - 30.84 \\ \frac{d_A(z_*,\theta)}{D_v(0.35,\theta)} - 10.33 \\ \frac{d_A(z_*,\theta)}{D_v(0.57,\theta)} - 6.72 \\ \frac{d_A(z_*,\theta)}{D_v(0.44,\theta)} - 8.41 \\ \frac{d_A(z_*,\theta)}{D_v(0.6,\theta)} - 6.66 \\ \frac{d_A(z_*,\theta)}{D_v(0.73,\theta)} - 5.43 \end{bmatrix}$$

Here, $D_v(z,\theta)$ represents the dilation scale which is given by $D_v(z,\theta) = \left(\frac{d_A^2(z,\theta)\, zc}{H(z,\theta)}\right)^{1/3}$. The comoving angular diameter, $d_A(z,\theta)$ is defined as

$$d_A(z_*,\theta) = \int_0^{z_*} \frac{dz'}{H(z',\theta)} \tag{66}$$

where z_* indicates the photon decoupling redshift and hold the value $z_* = 1090$ as per the Planck 2015 results [99]. We have taken the correlation coefficient from Ref. [100].

4.4 Local Hubble constant

We use $H_0 = 73.5 \pm 1.4\ km\ s^{-1} Mpc^{-1}$ which is locally measured by SH0ES as reported in [101] in our analysis.

In order to constrain the model parameters with the above data sets we perform a Bayesian Markov Chain Monte Carlo (MCMC) method. This method is based on the publicly available EMCEE package [85] for analysing and plotting the contours. In our calculation, we have minimized the chi-square for two combinations of data set, which we believe are helpful for better fit values. The first one is labeled $DS1$ and contains $SNe(Pan) + H(z) + BAO/CMB + H_0$. The chi-square function for DS1 reads as $\chi^2_{DS1} = \chi^2_{SNe(Pan)} + \chi^2_H + \chi^2_{BAO/CMB} + H_0$. The second is $DS2$ which contains $SNe(Pan) + H(z) + BAO/CMB$ and the total chi-square function reads as $\chi^2_{DS2} = \chi^2_{SNe(Pan)} + \chi^2_H + \chi^2_{BAO/CMB}$.

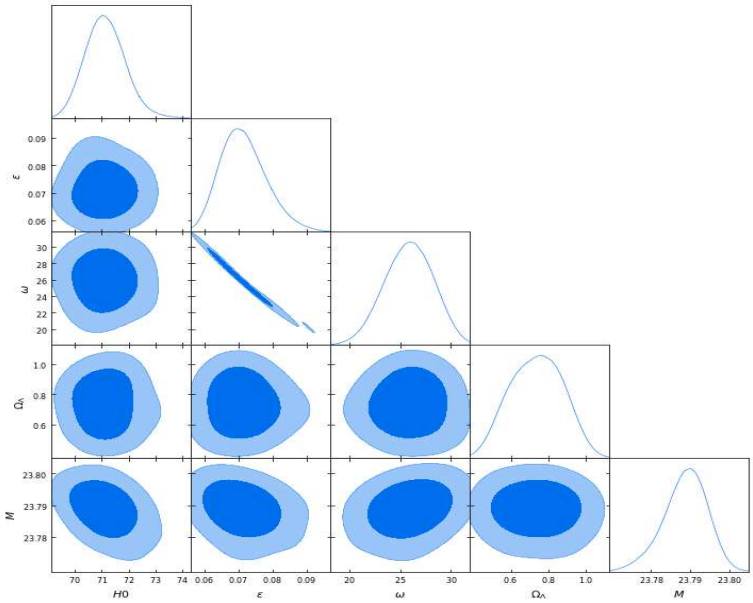


Fig. 1 One dimensional and two-dimensional marginalized confidence regions for the model parameters of Λ_{H1} from DS1 data set

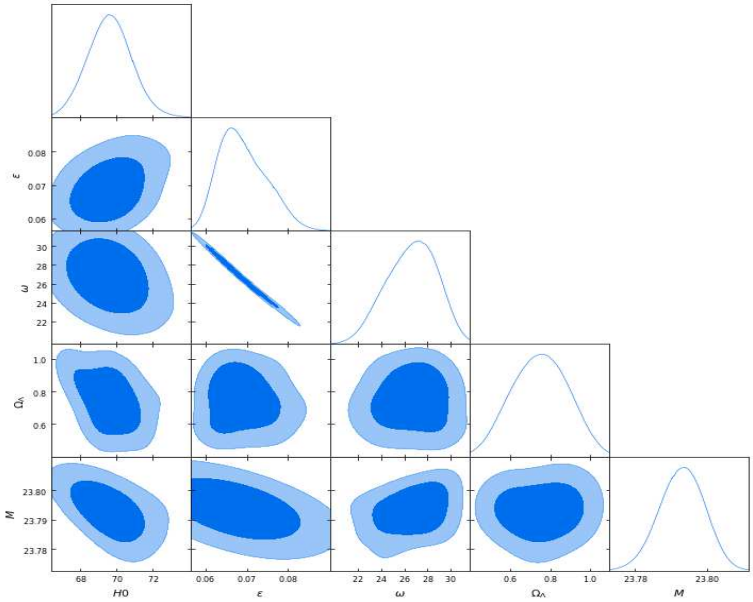


Fig. 2 One dimensional and two-dimensional marginalized confidence regions for the model parameters of Λ_{H1} from DS2 data set

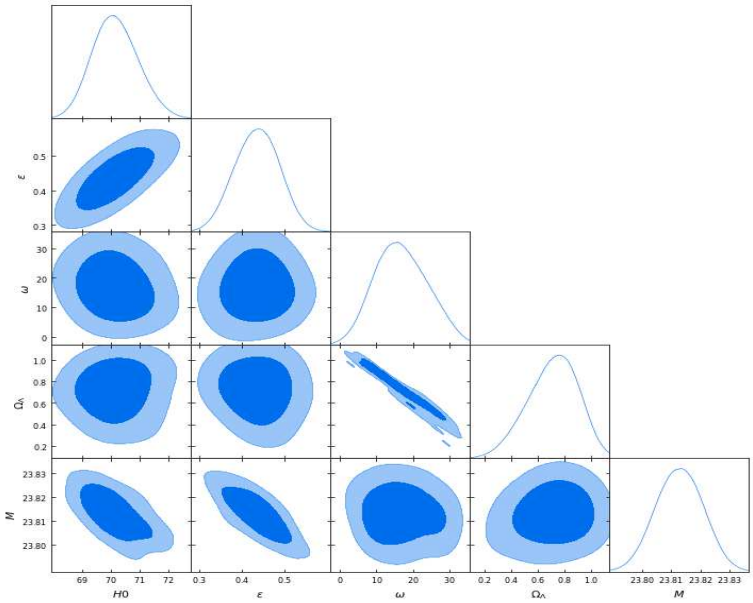


Fig. 3 One dimensional and two-dimensional marginalized confidence regions for the model parameters of Λ_{H2} from DS1 data set

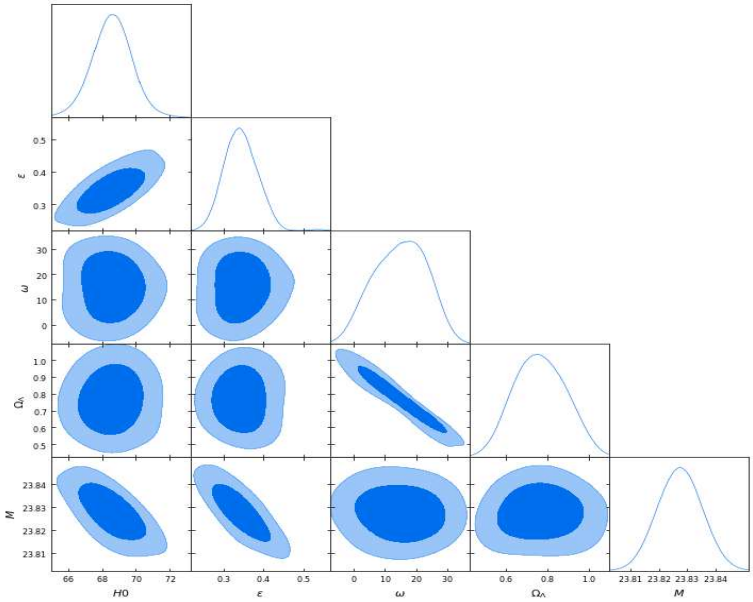


Fig. 4 One dimensional and two-dimensional marginalized confidence regions for the model parameters of Λ_{H2} from DS2 data set

Table 1 The fitting values for the considered models (Λ_{H1}, Λ_{H2} and ΛCDM) obtained from two different joint analysis of data sets, namely DS1:$SNe(Pan)+H(z)+BAO/CMB+H_0$ and DS2:$SNe(Pan)+H(z)+BAO/CMB$.

	DS1			DS2		
Parameter	Λ_{H1}	Λ_{H2}	ΛCDM	Λ_{H1}	Λ_{H2}	ΛCDM
H_0	$71.090^{+0.743}_{-0.627}$	$70.858^{+0.729}_{-0.969}$	$71.545^{+1.175}_{-0.820}$	$69.603^{+1.081}_{-1.100}$	$68.643^{+0.942}_{-1.111}$	$68.545^{+2.102}_{-1.742}$
ε	$0.070^{+0.005}_{-0.005}$	$0.406^{+0.053}_{-0.055}$	–	$0.068^{+0.005}_{-0.004}$	$0.344^{+0.051}_{-0.044}$	–
ω	$25.40^{+0.004}_{-0.003}$	$15.783^{+7.744}_{-7.968}$	–	$26.231^{+0.002}_{-0.003}$	$15.801^{+8.492}_{-10.684}$	–
Ω_Λ	$0.73^{+0.133}_{-0.183}$	$0.74^{+0.102}_{-0.229}$	$0.68^{+0.015}_{-0.010}$	$0.72^{+0.142}_{-0.169}$	$0.76^{+0.148}_{-0.123}$	$0.69^{+0.032}_{-0.028}$

Table 2 The numerical values of a_{tr}, z_{tr}, q_0, $w_{eff}(z=0)$ and t_0 using best-fit results of model parameters

	DS1			DS2		
Values	Λ_{H1}	Λ_{H2}	ΛCDM	Λ_{H1}	Λ_{H2}	ΛCDM
z_{tr}	$0.574^{+0.360}_{-0.324}$	$0.821^{+0.201}_{-0.103}$	$0.701^{+0.024}_{-0.020}$	$0.601^{+0.339}_{-0.343}$	$0.763^{+0.245}_{-0.167}$	$0.672^{+0.028}_{-0.025}$
q_0	$-0.580^{+0.210}_{-0.160}$	$-0.421^{+0.386}_{-0.288}$	$-0.594^{+0.014}_{-0.018}$	$-0.610^{+0.170}_{-0.190}$	$-0.388^{+0.294}_{-0.630}$	$-0.554^{+0.024}_{-0.030}$
$w_{eff}(z=0)$	$-0.675^{+0.173}_{-0.123}$	$-0.615^{+0.257}_{-0.259}$	$-0.729^{+0.009}_{-0.012}$	$-0.680^{+0.172}_{-0.120}$	$-0.592^{+0.196}_{-0.435}$	$-0.703^{+0.016}_{-0.020}$
t_0 (in Gyrs)	$13.69^{+1.829}_{-1.548}$	$14.18^{+1.419}_{-1.421}$	$13.48^{+0.450}_{-0.230}$	$13.73^{+1.941}_{-1.419}$	$14.14^{+1.502}_{-1.513}$	$13.69^{+0.09}_{-0.09}$

5 Results and discussion

In this section, we report the fitting results of the Λ_{H1}, Λ_{H2} and ΛCDM models using the two data sets DS1 and DS2 defined in the previous section, and discuss the implications of these results. Figures 1-4 show the constrained parameter space for the Λ_{H1} and Λ_{H2} models under consideration at 68.3% and 95.4% confidence level (CL) using DS1 and DS2 data sets, respectively. The mean fitting results obtained for both models with ΛCDM using DS1 and DS2 joint analysis are summarized in Table 1. We report uncertainties corresponding to 1σ CL. The transition redshift z_{tr}, the present values of deceleration parameter q_0, effective EoS parameter $w_{eff}(z=0)$, and the present age of the Universe, t_0, for these models are given in Table 2. The χ^2, reduced χ^2_{red} ($=\chi^2/(N-d)$, where N is the number of observational data and d is the number of free parameters), the model selection criterion (AIC, BIC, ΔAIC and ΔBIC) of different models are listed in Table 3. It is to be noted that we have taken $N=83$ for DS1 and $N=82$ for DS2(40 bin data points of Pantheon, 36 data points of H(z), 06 of BAO/CMB and 01 of H_0) and $d=4$ for our joint observational analysis: $(H_0, \varepsilon, \omega, \Omega_\Lambda)$. In what follows, we present the analysis of data in two parts: the cosmological parameters and the model selection criterion.

5.1 Cosmological parameters

The evolution of the deceleration parameter, q, with the redshift for the best-fit values of the parameters is shown in Fig. 5. It is observed that there is a sign change in each tra-

jectory of $q(z)$ from positive to negative showing that the Universe transits from decelerated phase to accelerated phase (positive values of q indicate decelerating expansion while negative values indicate an accelerating evolution). We find that the Λ_{H1}-model transits at around $z_{tr} = 0.574^{+0.360}_{-0.324}$ with DS1 data and $z_{tr} = 0.601^{+0.339}_{-0.343}$ with DS2 data, which are little smaller than ΛCDM model. However, model Λ_{H2} transits at around $z_{tr} = 0.821^{+0.201}_{0.103}$ with DS1 data and $0.763^{+0.245}_{-0.167}$ with DS2 data, which are higher than the transition value of ΛCDM model. The present-day values of q_0 and the transition redshift z_{tr} are listed in Table 2. It is found that the present value of q for Λ_{H1} is $q_0 = -0.580^{+0.210}_{-0.160}$ using DS1 data and $q_0 = -0.610^{+0.176}_{-0.190}$ using DS2 data. However, the present values of q for Λ_{H2} model are $q_0 = -0.421^{+0.386}_{-0.288}$ and $-0.388^{+0.294}_{-0.630}$ using DS1 and DS2 data, respectively. We observe that both the values of q_0 for data set DS1 and DS2 in Λ_{H1} model are very close to observational constraint $q_0 \simeq -0.63 \pm 0.12$ [102, 103], and are smaller than these values in Λ_{H2} model. It is to be noted that q tends to -1 in late times for both the models.

The evolution of the effective EoS parameter w_{eff} with redshift z is plotted in Fig.6 for the different models with their respective best fit values. The present values of ω_{eff} are listed in Table 2 for data sets DS1 and DS2. The present values of w_{eff} for Λ_{H1} model are $w_{eff}(z=0) = -0.675^{+0.173}_{-0.123}$ and $w_{eff}(z=0) = -0.680^{+0.172}_{-0.120}$ whereas for Λ_{H2} model, we have $w_{eff}(z=0) = -0.615^{+0.257}_{-0.259}$ and $w_{eff}(z=0) = -0.592^{+0.196}_{-0.435}$ for the data sets DS1 and DS2, respectively. These values are comparatively higher than the ΛCDM model. It is also observed from Fig.6 that w_{eff} becomes positive at high redshifts, which represents the early decelerated phase. In late times, w_{eff} approaches to -1 for all these models, thus leading to Einstein-de-Sitter behavior. These models do not cross the phantom-divide line $w_{eff} = -1$, which shows that they are free from big-rip singularity. Thus, w_{eff} can easily accommodates both phases of the cosmic evolution, i.e., early decelerated phase and late-time accelerated phase.

The evolutions of the age of the Universe with redshift for the best estimates of model parameters using DS1 and DS2 data sets are given in Table 2. The age of the Universe for

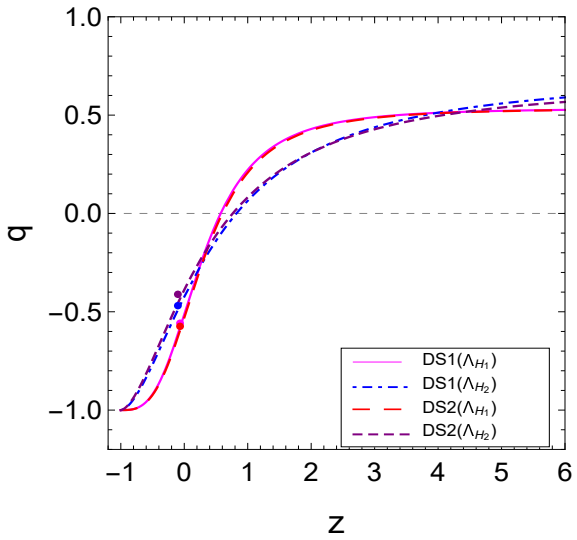


Fig. 5 The redshift evolution of the deceleration parameter for Λ_{H1} and Λ_{H2} models obtained from observational data sets DS1 and DS2. A dot denotes the current value of q (hence q_0).

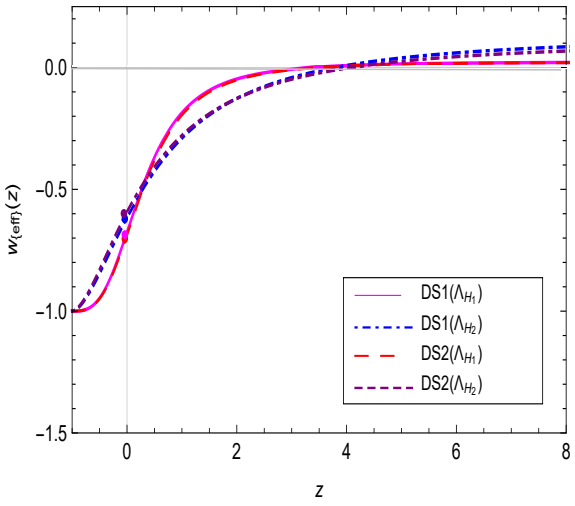


Fig. 6 The redshift evolution of effective EoS parameter for Λ_{H1} and Λ_{H2} models using DS1 and DS2 data sets. A dot denotes the present value of the EoS parameter

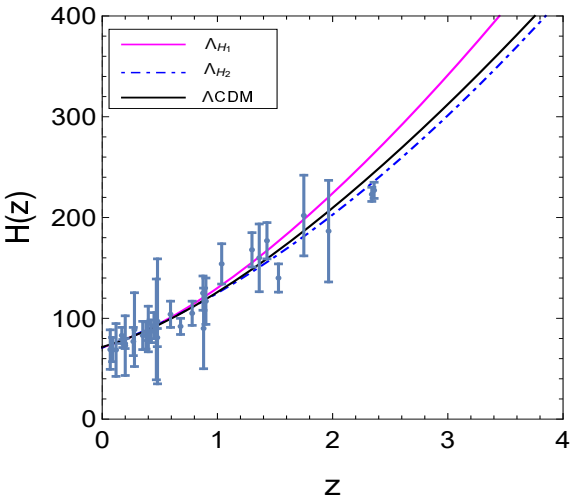


Fig. 7 Variation of the Hubble function as a function of the redshift z for the best-fit value of the models using DS1 data set. The observational 36 $H(z)$ points are shown with error bars (grey colour). The variation of the Hubble function in the standard ΛCDM model is also represented as the solid curve

Λ_{H1} and Λ_{H2} models are $t_0 = 13.69^{+1.829}_{-1.548}$ Gyrs and $t_0 = 14.18^{+1.419}_{-1.421}$ Gyrs for DS1 data set, and $t_0 = 13.73^{+1.941}_{-1.419}$ Gyrs and $t_0 = 14.14^{+1.502}_{-1.513}$ Gyrs for DS2 data set, respectively. In our finding, the large errors on the ages of the Universe for models are reported.

From Table 1, we observe that the current Hubble constant H_0 for Λ_{H1}, Λ_{H2} and ΛCDM models constrained from DS2 are a bit higher than the constrained observed from the Planck mission ($H_0 = 67.8 \pm 0.9\ Km\ s^{-1}\ Mpc^{-1}$)[104]. However, the addition of local H_0, namely DS1 data set, makes the constraint on H_0 bigger for these models. It is to be noted that the improved local measurement $H_0 = 73.5 \pm 1.4\ km\ s^{-1} Mpc^{-1}$ reported by SH0ES [101]

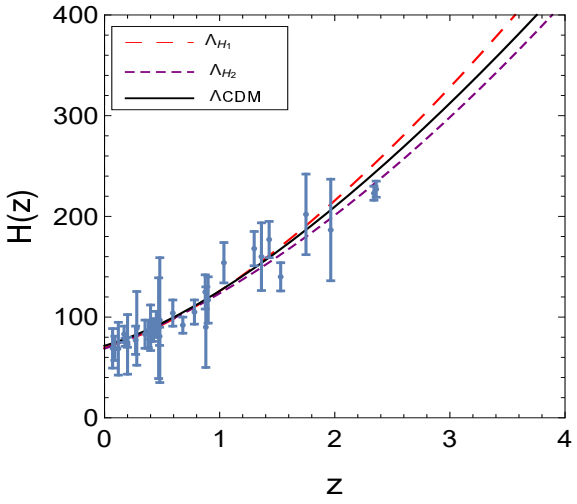


Fig. 8 Variation of the Hubble function as a function of the redshift z for the best-fit value of the models using Ds2 data set. The observational 36 $H(z)$ points are shown with error bars (grey colour). The variation of the Hubble function in the standard ΛCDM model is also represented as the solid curve

exhibits a strong tension with the Planck mission data [105]. Let us note that in the paper [106] it has been shown that true quintessence models show a preference for lower values of H_0 relative to the ΛCDM model; and this is true even for coupled quintessence, as recently shown in [107]. Let us, however, emphasize that there are dynamical DE models in the market which show an effective quintessence behavior (i.e. they mimic quintessence in that the DE density diminishes with the expansion) but they are nevertheless very different from true quintessence models based on scalar fields. Some of these models mimicking quintessence behavior can have an impact on the H_0 tension (and even on the σ_8) one. Such is the case e.g. for the running vacuum models (RVMs), see the recent work [74]. It is also interesting to remark that BD cosmology can help to relax these tensions, as shown in [62,63], the reason being that BD cosmology with a CC term can mimic the RVM behavior [81]. Our present study further reinforces such welcome property of BD models possessing vacuum energy.

Figures 7 and 8 display the Hubble diagram with the error bar of Hubble data set in the range $z \in (0,4)$ for DS1 and DS2 data sets. For the sake of comparison, the flat ΛCDM scenario is also shown. The evolutions of $H(z)$ of Λ_{H1} and Λ_{H2} model are comparatively similar to the ΛCDM model. At low redshifts, the cosmological evolutions of models Λ_{H1} and Λ_{H2} are consistent with the Hubble data.

5.2 Model selection

Taking into account that models Λ_{H1} and Λ_{H2} have the same number of extra parameters (λ and σ, respectively), we could directly compare them on the basis of computing the minimum χ^2 values for each model. But when we compare them with the concordance ΛCDM model, the χ^2 comparison becomes unfair because in the context of BD theory the number of parameters is different. In our approach two more parameters, ε and ω, have to be considered which are not involved in the concordance model. For this reason we employ the Akaike

Table 3 Summary of χ^2, χ^2_{red}, *AIC* and *BIC* values and their differences from the reference model of ΛCDM obtained from joint analysis of DS1 and DS2 data sets

Data	Model	χ^2	χ^2_{red}	*AIC*	*BIC*	ΔAIC	ΔBIC
DS1	ΛCDM	34.78	0.43	40.78	48.04	0	0
	Λ_{H1}	34.94	0.44	42.94	52.62	2.16	4.58
	Λ_{H2}	36.94	0.47	44.94	54.62	4.16	6.58
DS2	ΛCDM	34.69	0.44	40.69	47.91	0	0
	Λ_{H1}	35.28	0.45	43.28	52.90	2.59	4.99
	Λ_{H2}	39.42	0.51	47.42	57.04	6.73	9.13

Information Criterion (AIC)[108] and the Bayesian Information Criterion (BIC)[109] so as to do a fairer model comparison, see e.g. [110] for a review. The AIC parameter is defined through the relation

$$AIC = \chi^2_{min} + 2d, \tag{67}$$

where d is the number of free parameters in the model and χ^2_{min} is the minimum value of the χ^2 function. We calculate $\Delta AIC_i = AIC_i - AIC_j$, where i, j denote respectively the model i and model j. This is interpreted as "evidence in favour" of the model i compared to the model j. The preferred model for this criterion is one with the smaller value of AIC.

On the other hand, the BIC is defined through the relation

$$BIC = \chi^2_{min} + d\ ln\ N, \tag{68}$$

where N is the number of data points. Similar to ΔAIC, we have $\Delta BIC = BIC_i - BIC_j$. This can be interpreted as "evidence against" the model i compared to the model j.

To be more precise, a model having $0 \leq \Delta AIC < 2$ and $0 \leq \Delta BIC < 2$ receives "strong evidence in favour". In contrast, for $2 < \Delta AIC < 4$ and for $2 \leq \Delta BIC < 6$, the model has "average evidence in favour", whereas for $4 < \Delta AIC < 7$ and $6 \leq \Delta BIC < 10$, the model is considered to have "less evidence in favour"; and, finally, for $\Delta AIC > 10$ or $\Delta BIC > 10$, the model receives no significant support since it has "no evidence in favour" [110].

In Table 3, we present the values of χ^2, AIC, BIC and their differences for the discussed models. From Table 3, we find that Λ_{H1} has ΔAIC~ 2.16 and ΔAIC~ 2.59, whereas it has ΔBIC~ 4.58 and ΔBIC~ 4.99 from DS1 and DS2 data sets, respectively. We see that the Λ_{H1} model is in the range of $2 \leq \Delta AIC < 4$ and $2 \leq \Delta BIC < 6$. Thus, this model shows average evidence in favour. However, the Λ_{H2} model has ΔAIC~ 4.16 and ΔAIC~ 6.73, and ΔBIC~ 6.58 and ΔBIC~ 9.13 from DS1 and DS2 data sets, respectively. Since, this model shows the differences in the range of $4 < \Delta AIC < 7$ and $6 \leq \Delta BIC < 10$, therefore, this model has "less evidence in favour". The AIC and BIC impose a strict penalty against the presence of additional parameters.

In Table 3, we give the values of χ^2_{min}. We find, among the models, the ΛCDM model is still the best one in fitting the current observational data. The ΛCDM has least number of parameters, but it gets the smallest χ^2_{min} value in this fit. The Λ_{H1} and Λ_{H2} models have one more parameter than the ΛCDM model. However, Λ_{H1} yields very close χ^2_{min} values to that of ΛCDM from the two datasets ,DS1 and DS2, whilst Λ_{H2} renders corresponding

higher values. The reduced χ^2_{red} of both the models is less than one (cf. Table 3), so overall these two models can be considered in good agreement with the ΛCDM model and data are consistent with the considered models. Model Λ_{H1} seems to be the closest. Therefore, our analysis suggests that the BD version of the Λ-cosmology is on an essentially equal footing position as compared to the concordance model in the light of our fits.

6 Conclusion

Among the many proposals to describe the late time acceleration of the Universe, the cosmological constant (CC) is the simplest candidate to provide an explanation. It defines the standard or concordance ΛCDM model, and it is referred to also as the Λ-cosmology. In this paper, we have studied if the Λ-cosmology, which in its standard version is implemented through general relativity (GR), can be realized too in the context of Brans-Dicke (BD) gravity and with a similar or better level of achievement. We have assumed, as in many other studies in the literature, that the local constraints imposed on BD gravity can be avoided by resorting to the presence of screening forces, which do not affect the study of cosmology at the level of the large scales. To make our study of BD cosmology more complete, we have explored the possibility to add some dynamical component to the vacuum energy in the BD framework. For such purpose we have explored theoretical and observational features of a simple class of cosmological models driven by a time-varying vacuum energy density for a spatially flat FRW spacetime in BD theory, hence beyond the GR paradigm underlying the standard ΛCDM model. Such class of models is characterized by a time-evolving CC of the form $\Lambda = \lambda + \sigma H$. We have solved these models searching for power-law solutions and checked the consistency of the obtained solutions. We have separately solved the two particular cases $\Lambda = \lambda$ (model Λ_{H1}) and $\Lambda = \sigma H$ (model Λ_{H2}) as well as the general case with arbitrary λ and σ. For the numerical analysis we have used the latest observational measurements of SnIa (Pantheon), $H(z)$, BAO/CMB and local H_0. The corresponding results are presented in Table 1 while Figs. 1-4 show the confidence contours for the different parameters. The two models Λ_{H1} and Λ_{H2} have been analyzed using two data sets DS1 and DS2, where the local H_0 value is only included in the first set. Upon using the model selection criteria AIC and BIC, we find that for both data sets the phenomenological performance of model Λ_{H1} is better than that of Λ_{H2}, but both models are acceptable for the description of the data. The more general model $\Lambda = \lambda + \sigma H$ only interpolates between the two former ones. Since the description of the data does not improve with a nonvanishing value of σ, we conclude that the BD version of the Λ-cosmology is the preferred option and it proves comparable to the conventional ΛCDM in light of our fitting results.

The observational analysis shows that both the models Λ_{H1} and Λ_{H2} exhibit the same characteristics concerning the evolution of the Universe, i.e., they describe the transition from a decelerated to an accelerated phase. In both cases the effective equation of state (EoS) performs an evolution from $w_{eff} = 0$ in the remote past to to $w_{eff} = -1$ in the remote future without crossing the phantom divide $w = -1$. At late-time, the deceleration parameter q tends also to -1, showing that these models predict de Sitter behavior in the future. In the case of Λ_{H1}, the current value is $q_0 \simeq -0.6$, thus similarly to the concordance model, whereas for model Λ_{H2} it is smaller in absolute value ($q_0 \simeq -0.4$), see Table 2 for detailed results.

Using the fitting values of the parameters listed in Table 1 into the consistency relation (39), it has been found that the Λ_{H1} model has $\Omega_\Lambda = 0.728^{+0.213}_{-0.173}$ with DS1 data set and $\Omega_\Lambda = 0.733^{+0.222}_{-0.146}$ with DS2 data set. The errors are sizeable, but even at the level of the best

fit values these results are perfectly consistent with those obtained from the observations using DS1 and DS2 data sets, viz., $\Omega_\Lambda = 0.73^{+0.133}_{-0.183}$ and $\Omega_\Lambda = 0.72^{+0.142}_{-0.169}$, respectively (cf. Table 1). As for model Λ_{H2}, using the fitting values of that table into the consistency relation (58) we find $\Omega_\Lambda = -1.658^{+0.826}_{-11.361}$ with DS1 data and $\Omega_\Lambda = -2.025^{+0.985}_{-2.722}$ with DS2 data set. These numerical values are inconsistent with the fitting values of Ω_Λ obtained for DS1 and DS2. The latter remain nonetheless in the approximate range $\Omega_\Lambda = 0.72 - 0.77$ (using the errors) for both data sets (cf. Table 1). Therefore, we find that in the context of our analysis the performance of model Λ_{H1} is very similar to that of the standard ΛCDM model. At the level of information criteria it is at the border line of not implying any significant difference with the standard Λ-cosmology. On the other hand, Λ_{H1} satisfies remarkably well the consistency equation (39). In stark contrast, despite the quality fit of the Λ_{H2} is lesser, it is still a reasonable one for the DS1 data while it is not so good for DS2. In addition, the model does not adapt to the consistency condition (58). We should emphasize that this is not caused by any analytical inconsistency in our study, the discrepancy is only numerical because of assuming a power-law relation between the scalar field and the scale factor. Such relation may not be a perfect choice for the solutions of the BD field equations in the case of the Λ_{H2} model, and this means that a more general family of solutions is needed. Let us, however, note that it is difficult to explore other kind of analytical solutions for the complicated system of BD equations and one may be forced to go fully numerical in this case. The dynamics of model Λ_{H1}, instead, adapts well to the power-law solution since the numerical consistency is manifest. Recall from the footnote on page 9 that model Λ_{H1} has a smooth analytic limit to GR for $\varepsilon \to 0$. In the case of model Λ_{H2} we do not expect such limit to hold since its effective cosmological term is not constant at any time and hence there is no smooth connection with the concordance ΛCDM model. This can also explain why this model does not adapt equally well to the same power-law family of solutions as for model Λ_{H1}.

The main conclusion of our study is in our opinion significant. We have shown that model Λ_{H1}, namely the Λ-cosmology in the context of the BD theory, is more favored than Λ_{H2} and is comparable to the concordance ΛCDM model within GR. Since for model Λ_{H1} the consistency relation is fully realized also at the numerical level we can say that this claim is robust. In the case of model Λ_{H2} it is only indicative. This does not preclude, however, the possibility that other forms of dynamical Λ can improve the performance of BD theory as compared to the standard ΛCDM model. In the meantime our analysis shows that of all the possible dynamical models $\Lambda = \lambda + \sigma H$ within the BD paradigm, the most promising ones are those with $\sigma \simeq 0$. Notwithstanding, this conclusion should not be interpreted as saying that a rigid cosmological term in BD theory is equivalent to the effect of a rigid cosmological term in GR. As previously noted, model Λ_{H1} despite it being associated to a rigid cosmological constant term in the BD context, it is perceived as a running vacuum model from the point of view of GR. This fact is helpful since it is known that the running vacuum model performs a fit to the overall cosmological data which is competitive with that of the concordance model [74]. Only future studies can reveal if Brans-Dicke gravity with a rigid cosmological term can be fully competitive with GR in all aspects of the observational cosmological data, and to which extent it may be necessary to introduce a dynamical component in it. Here we have shown that the Λ-cosmology in such BD context is not second rate as compared to the GR version, and that the addition of the simplest possible dynamical component to the vacuum energy does not perturb exceedingly this conclusion.

Acknowledgements One of the authors, JSP, acknowledges partial support by projects PID2019-105614GB-C21 and FPA2016-76005-C2-1-P (MINECO, Spain), 2017-SGR-929 (Generalitat de Catalunya) and CEX2019-

000918-M (ICCUB). JSP also acknowledges participation in the COST Association Action CA18108 "*Quantum Gravity Phenomenology in the Multimessenger Approach (QG-MM)*".

Appendix A: Time-varying model $\Lambda = \lambda + \sigma H$ within the BD theory

In this appendix, we briefly provide the analytical solution of the time varying $\Lambda(t)$ model $\Lambda = \lambda + \sigma H$, as defined in Eq.(19) of the main text, in its general form, i.e. for arbitrary values of $\lambda \neq 0$ and $\sigma \neq 0$, and within the Brans-Dicke context.

Using this form of $\Lambda(t)$ into Eq. (18), the evolution equation for the Hubble function can be rewritten as

$$\dot{H} + \frac{(3+\varepsilon)}{2}H^2 - \frac{3\sigma}{(6+6\varepsilon-\omega\varepsilon^2)}H - \frac{3\lambda}{(6+6\varepsilon-\omega\varepsilon^2)} = 0. \tag{69}$$

The solution to this differential equation reads as follows:

$$H = H_0\left[\frac{3\sigma}{(6+6\varepsilon-\omega\varepsilon^2)(3+\varepsilon)} + \alpha\left(\frac{e^{(3+\varepsilon)\alpha\, t}+1}{e^{(3+\varepsilon)\alpha\, t}-1}\right)\right], \tag{70}$$

where $\alpha = \frac{\sqrt{9\sigma^2+6\lambda(6+6\varepsilon-\omega\varepsilon^2)(3+\varepsilon)}}{(6+6\varepsilon-\omega\varepsilon^2)(3+\varepsilon)}$.
The solution of the scale factor is given by

$$a(t) = \left(e^{(3+\varepsilon)\alpha\, t}-1\right)^{\frac{1}{3+\varepsilon}}\; e^{\left(\frac{3}{(6+6\varepsilon-\omega\varepsilon^2)(3+\varepsilon)}-1\right)\alpha\, t}. \tag{71}$$

In this case, the deceleration parameter is obtained as

$$q = -1 + \frac{2\alpha^2(3+\varepsilon)\, e^{(3+\varepsilon)\alpha\, t}}{\left(\frac{3\sigma}{(6+6\varepsilon-\omega\varepsilon^2)(3+\varepsilon)} + \alpha\left(\frac{e^{(3+\varepsilon)\alpha\, t}+1}{e^{(3+\varepsilon)\alpha\, t}-1}\right)\right)^2 \left(e^{(3+\varepsilon)\alpha\, t}-1\right)^2}. \tag{72}$$

Lastly, the effective EoS parameter can be worked out with the following result:

$$w_{ef} = -1 + \frac{2}{3}\frac{2\alpha^2(3+\varepsilon)\, e^{(3+\varepsilon)\alpha\, t}}{\left(\frac{3\sigma}{(6+6\varepsilon-\omega\varepsilon^2)(3+\varepsilon)} + \alpha\left(\frac{e^{(3+\varepsilon)\alpha\, t}+1}{e^{(3+\varepsilon)\alpha\, t}-1}\right)\right)^2 \left(e^{(3+\varepsilon)\alpha\, t}-1\right)^2}. \tag{73}$$

The above solution of cosmological parameters are found in terms of exponential form which show that the model can accommodate the late time acceleration. It can also be shown that Λ_{H1} and Λ_{H2} models are particular solutions of the above general vacuum model.

References

1. S. Perlmutter et al., Astrophys. J. **517**, 565 (1999)
2. A.G. Riess et al., Astrophys. J. **607**, 665 (2004)
3. P. Astier et al., Astron. Astrophys. **447**, 31 (2006)
4. D.N. Spergel. et al., Astrophys. J. Suppl. Ser. **170**, 377 (2007)
5. H.A. Feldman et al., Astrophys. J. **596**, L131 (2003)
6. E. Komatsu et al., Astrophys. J. Suppl. Ser. **180**, 330 (2009)
7. E. Komatsu et al., Astrophys. J. Suppl. Ser. **192**, 11 (2011)
8. A.G. Sanchez et al., Mon. Not. R. astron. Soc. **425**, 415 (2011)
9. P.A.R. Ade et al., Aston. Astrophys. **517**, A16 (2014)

10. P.A.R. Ade et al., Aston. Astrophys. **594**, 13 (2016)
11. S. Weinberg, Reviews of Mod. Phys. **61**, 1 (1989)
12. J. Solà, J.Phys.Conf.Ser. **453**, 012015 (2013)
13. E. J. Copeland, M. Sami and S. Tsujikawa, Int. J. Mod. Phys. D **15**, 1753 (2006)
14. M. Ozer and O. Taha, Phys. Lett. B **171**, 363 (1986)
15. P.L.E. Peebles and B. Ratra, Astrophys. J. **325**, L17 (1988)
16. J.C. Carvalho, J.A.S. Lima and I. Waga, Phys. Rev. D **46**, 2404 (1992)
17. J.A.S. Lima, Phys. Rev. D **54**, 2571 (1996)
18. J. M. Overduin and S. Cooperstock, Phys. Rev. D **58**, 043506 (1998).
19. I.L. Shapiro and J. Solà, JHEP **02**, 006 (2002)
20. J. Solà, J.Phys.A **41**, 164066 (2008)
21. C. Moreno-Pulido and J. Solà Peracaula, Eur.Phys.J.C **80**, 692 (2020)
22. R. Schützhold, Phys. Rev. Lett. **89**, 081302 (2002)
23. S. Carneiro, Int. J. Mod. Phys. D **12**, 1669 (2003)
24. H.A. Borges and S. Carneiro, Gen. Relativ. Grav. **37** 1385 (2005)
25. S. Carneiro, C. Pigozzo, H.A. Borges and J.S. Alcaniz, Phys. Rev. D **74**, 023532 (2006)
26. S. Carneiro, M.A. Dantas, C. Pigozzo and J.S. Alcaniz, Phys. Rev. D **77**, 083504 (2008)
27. S. Basilakos, Mon. Not. R. Astron. Soc. **395**, 2347 (2009)
28. S. Basilakos, M, Plionis and J. Solà, Phys. Rev. D **80**, 083511 (2009)
29. E.L.D. Perico, et al., Phys. Rev. D **88**, 063531 (2013)
30. D. Bessada and O.D. Miranda, Phys. Rev. D **88**, 083530 (2013)
31. J.A.S. Lima, S. Basilakos and J. Solà, Mon. Not. R. Astron. Soc. **431**, 923 (2013)
32. M. Szydlowski and A. Stachowski, J. Cosmol. Astropart. Phys. **066**, 10 (2015)
33. A.P. Jayadevan, et al., Astrophys. Space Sci. **364**, 67 (2019)
34. J. Solà and A. Gómez-Valent, Int.J.Mod.Phys. D **24**, 1541003 (2015)
35. F.R. Urban and A. R. Zhitnitsky, Phys.Lett.B 688, 9 (2010) ; Nucl. Phys. B**835** 135 (2010)
36. N. Ohta, Phys.Lett. B**695** 41 (2011)
37. R-G Cai et al. Phys.Rev. D**86**, 023511 (2012)
38. C. Brans and R.H. Dicke, Phys. Rev. **124**, 925 (1961)
39. B. Bertotti, L. Iess and P. Tortora, Nature **425**, 374 (2003)
40. A. Avilez and C. Skordis, Phys.Rev.Lett. **113**, 011101 (2014)
41. T. Clifton et al., Phys.Rept. **513**, 1 (2012)
42. L.O. Pimental, Astrophys. Space Sci. **112**, 175 (1985)
43. V.P. Johri and D. Kalyani, Gen. Relativ. Grav. **26**, 1217 (1994)
44. ShriRam and C.P. Singh, Nuovo Cim. B **114**, 245 (1999)
45. A.A. Sen, S. Sen and S. Sethi, Phys. Rev. D **63**, 107501 (2001)
46. N. Banerjee and D. Pavon, Phys. Rev. D **63**, 043504 (2001)
47. N. Banerjee and D. Pavon, Class. Quantum Grav. **18**, 593 (2001)
48. S. Sen and A. A. Sen, Phys. Rev. D. **63**, 124006 (2001).
49. D. F. Mota and J. D. Barrow, Mon. Not. Roy. Astron. Soc. **349**, 291 (2004)
50. S. Das, P. S. Corasaniti and J. Khoury, Phys. Rev. D. **73**, 083509 (2006)
51. M. Arik and M. Çalik, Mod. Phys. Lett. A **21**, 1241 (2006)
52. M. Arik, M. Çalik and M.B. Sheftel, Int. J. Mod. Phys. D **17**, 225 (2008)
53. L. Xu et al., Mod. Phys. Lett. A **25**, 1441 (2010)
54. C.P. Singh, Astrophys. Space Sci. **338**, 411 (2012)
55. A.P.K. Karchi and H. Shojaie, Int. J. Mod. Phys. D **25**, 1650045 (2016)
56. P. Kumar and C.P. Singh, Astrophys. Space Sci. **362**, 52 (2017)
57. C.P.Singh and P. Kumar, Int. J. Theor. Phys. **56**, 3297 (2017)
58. M. Srivastava and C.P. Singh, Int. J. Geom. Methods Mod. Phys. **15**, 1850124 (2018)
59. C.P. Singh and S. Kaur, Phys. Rev. D **100**, 084057 (2019)
60. M. Sharif and Syed Asit Ali Shah, Mod. Phys. Lett. A **34**, 1950083 (2019)
61. C.P. Singh and S. Kaur, Astrophys. Space Sci. **365**, 2 (2020)
62. J. Solà Peracaula, A. Gómez-Valent., J. de Cruz Pérez and C. Moreno-Pulido, Astrophys.J.Lett. **886**, L6 (2019)
63. J. Solà Peracaula, A. Gómez-Valent., J. de Cruz Pérez and C. Moreno-Pulido, Class. Quantum Grav. **37**, 245003 (2020)
64. J. Solà Peracaula, J. de Cruz Pérez and A. Gómez-Valent, Eur. Phys. Lett. **121**, 39001 (2018)
65. J. Solà Peracaula, J. de Cruz Pérez and A. Gómez-Valent, Mon. Not. R. Astron. Soc. **478**, 4357 (2018)
66. J. Solà, J. de Cruz Pérez and A. Gómez-Valent, Astrophys.J. **836**, 43 (2017)
67. M. Rezaei, M. Malekjani and J. Solà Peracaula, Phys.Rev.D**100** 023539 (2019)

68. M. Rezaei, J. Solà Peracaula and M. Malekjani, *Cosmographic approach to Running Vacuum dark energy models: new constraints using BAOs and Hubble diagrams at higher redshifts*, arXiv: 2108.06255
69. L. Verde, T. Treu and A. G. Riess, Nat. Astron. **3** 891, (2019)
70. J. Solà Peracaula, Int.J. Mod. Phys. A**33**, 1844009 (2018)
71. E. Di Valentino, *et al.*, *Cosmology Intertwined II: The Hubble Constant Tension*, arXiv:2008.11284; *Cosmology Intertwined III: $f\sigma_8$ and S_8*, arXiv:2008.11285
72. J. Solà, A. Gómez-Valent, J. de Cruz Pérez, Phys.Lett. B**774**, 317 (2017).
73. A. Gómez-Valent and J. Solà Peracaula, Mon. Not. Roy. Astron. Soc. **478**, 126 (2018).
74. J. Solà Peracaula, A. Gómez-Valent., J. de Cruz Pérez and C. Moreno-Pulido, Eur. Phys. J. **134**, 19001 (2021)
75. K. Uhera and C. W. Kim, Phys. Rev. D **26**, 2575 (1982)
76. H. Kim, Mon. Not. Roy. Astron. Soc. **364**, 813 (2005)
77. S. Weinberg, Gravitation and Cosmology (2002)
78. L.O. Pimentel, Astrophys. Space Sci. **112**, 175 (1985)
79. N. Banerjee and D. Pavon, Phys.Lett.B**647**, 477 (2007)
80. A. Sheykhi, Phys. Rev.D **81**, 023525 (2010)
81. J. Solà Peracaula, Int. J. Mod. Phys. D **27** 1847029 (2018)
82. J. de Cruz Pérez and J. Solà Peracaula, Mod. Phys. Lett. A **33**, 1850228 (2018)
83. E. Karimkhani and A. Khoadam-Mohammadi, Astrophys.Space Sci. **364**, 177 (2019)
84. J-P. Uzan, Living Rev.Rel. **14**, 2 (2011)
85. D. Foreman-Mackey, D. Hogg, D. Lang and J. Goodman, Publ. Astron. Soc. Pac. **125**, 306 (2013)
86. M. Moresco et al., J. Cosmol. Astropart. Phys. **08**, 006 (2012)
87. S. Alam et al., Mon. Not. R. Astron. Soc. **470**, 2617 (2017)
88. T. Delubac et al., Astron. Astrophys. **574**, A59 (2015)
89. A. Font-Ribera et al., J. Cosmol. Astropart. Phys. **05**, 027 (2014)
90. O. Akarsu, S. Kumar, S. Sharma and L. Tedesco, Phys. Rev. D **100**, 023532 (2019)
91. D.M. Scolnic et al., Astrophys. J. **859**, 101 (2018)
92. M. Betoule, et al., Astron. Astrophys. **568**, A22 (2014)
93. A. Conley et al., Astrophys. J. Suppl. Ser. **192** 1 (2011)
94. M.V. de Santos, R.R.R. Reis, I. Waga, J. Cosmol. Astropart. Phys. **02**, 66 (2016)
95. N. Padmanabhan, et al., Mon. Not. R. Astron. Soc. **427**, 2132 (2012)
96. F. Beutler, et al., Mon. Not. R. Astron. Soc. **416**, 3017 (2011)
97. L. Anderson, et al., Mon. Not. R. Astron. Soc. **441**, 24 (2014)
98. C. Blake, et al., Mon. Not. R. Astron. Soc. **425**, 405 (2012)
99. P. A. R. Ade, et al., Astron. Astrophys. **594**, A13 (2016)
100. G. Hinshaw, et al., Astrophys. J. Suppl. **208**, 19 (2013)
101. M. J. Reid, D. W. Pesce and A. G. Riess, Astrophys. J. Lett. **886 (2)**, L27 (2019)
102. N. Aghanim, et al., Astron. Astrophys. **596**, A107 (2016)
103. R. Adam, et al., Astron. Astrophys. **596**, A108 (2016)
104. A. Dominguez, et al., Astrophys. J. **885**, 137 (2019)
105. P.R.A. Ade, et al. Planck Collaboration XIII 2015 Cosmological Parameters, Astron. Astrophys. **594**, 13 (2016), ArXiv: 1502.01589
106. A. Banerjee, et al., Phys.Rev.D **103**, L081305 (2021)
107. A. Gómez-Valent, V. Pettorino and L. Amendola, Phys.Rev.D **101**, 123513 (2020)
108. H. Akaike, IEEE Transactions on Automatic Control. **19**, 716 (1974)
109. G. Schwarz, Ann. Statist. **6**, 461 (1978)
110. A.R. Liddle, Mon. Not. R. Astron. Soc. **377**, L74 (2007)

2021 8th International Conference on Signal Processing and Integrated Networks (SPIN) | 978-1-6654-3564-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SPIN52536.2021.9566076

Grounded-Capacitor Analog Inverse Active Filters using CMOS OTAs

Ajishek Raj
Department of Electrical Engineering
Delhi Technological University
Delhi, India
ajishekaerotek@gmail.com

Ram Bhagat
Department of Electrical Engineering
Delhi Technological University
Delhi, India
rambhagat@dtu.ac.in

Pragati Kumar
Department of Electrical Engineering
Delhi Technological University
Delhi, India
pragatikumar@dtu.ac.in

D. R. Bhaskar
Department of Electrical Engineering
Delhi Technological University
Delhi, India
drbhaskar@dtu.ac.in

***Abstract*—Five new electronically tunable voltage-mode analog inverse active filters using single output operational transconductance amplifiers (OTAs) have been presented. The new proposed configurations can realize four different inverse filter structures using four/five OTAs along with two grounded capacitors. An exemplary OTA with 0.18 μm TSMC CMOS technology parameters has been used to corroborate the workability of these inverse filter topologies. PSPICE simulations with MATLAB evaluation results have been presented to validate the theoretical propositions. Temperature analysis and Monte-Carlo simulations have also been carried out to justify the robustness of the proposed inverse filters.**

***Index Terms*—OTA, Inverse Filter design, PSPICE, Voltage Mode**

I. Introduction

At present, analog active inverse filter circuits have received considerable attention due to their applications in communication, control and instrumentation systems where these filters are needed to correct the distortion of the signal caused by the transmission medium or the signal processor [1]. In general, an inverse active filter with a characteristic, which is the inverse of the loss frequency characteristics of the transmission medium, is required. Over the decades, different analog active building blocks (ABBs) have been employed to design active inverse analog filters [2] - [41].

Operational transconductance amplifier, because of its simple structure, easy integrability (both Bipolar and CMOS technology), versatility, has gained widespread popularity amongest various ABBs used for the realization of VM and CM analog active filter circuits, oscillators and inverse analog filters [23]. A comprehensive review of the work done on the realization of analog inverse active filters was presented in [21] from which it has been observed that only a single publication [23] has appeared in open literature so far, for the realization of electronically controllable inverse active filters employing OTAs. The inverse filters presented therein, employ two/four dual-output OTAs along with two/one single output OTA and two grounded capacitors (GCs) for the realization of various inverse filter responses. It may also be pointed out that the fractional order inverse active filters employing three OTAs, one floating fractional order capacitor and one grounded fractional order capacitor have been presented in [34]. The fractional order inverse filters reported in [34] can also be converted into second order inverse analog filters just by replacing both the fractional order capacitors with normal capacitors.

Therefore, the aim of this communication is to investigate some new voltage mode analog inverse filter circuits that employ single output OTAs and two grounded capacitors. For the validation of proposed inverse filter topologies, CMOS OTA with 0.18 μm CMOS parameters has been used. Sample results of the proposed inverse active filters have been included using PSpice simulation tool along with MATLAB evaluations.

II. Proposed voltage mode inverse active filter configurations

Fig. 1 shows the symbol of an OTA.

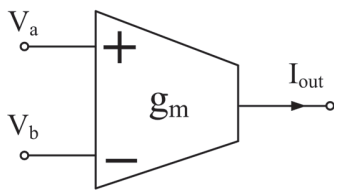


Fig. 1: Symbol of an OTA

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:33:21 UTC from IEEE Xplore. Restrictions apply.

The output current of OTA in terms of input voltages is given by:

$$I_{out} = g_m(V_a - V_b) \tag{1}$$

where g_m is the transconductance of the OTA.

The proposed active filter configurations with inverse characteristic viz. high pass (IHP), low pass (ILP), band pass (IBP), and band reject (IBR) filters are displayed in Fig. 2.

The different transfer functions of the inverse filters shown in Fig. 2 have been derived assuming ideal OTAs and are given by:

$$\frac{V_{01}(s)}{V_{IN}(s)} = \frac{1}{\frac{s^2}{s^2+s\frac{g_{m1}g_{m4}}{C_1g_{m3}}+\frac{g_{m1}g_{m2}g_{m4}}{C_1C_2g_{m3}}}} \tag{2}$$

$$\frac{V_{02}(s)}{V_{IN}(s)} = \frac{1}{\frac{\frac{g_{m1}g_{m2}g_{m4}}{C_1C_2g_{m3}}}{s^2+s\left(\frac{g_{m1}}{C_1}+\frac{g_{m1}}{C_1}\right)+\frac{g_{m1}g_{m2}}{C_1C_2}}} \tag{3}$$

$$\frac{V_{03}(s)}{V_{IN}(s)} = \frac{1}{\frac{\frac{g_{m1}g_{m2}}{C_1C_2}\left(1+\frac{g_{m3}}{g_{m4}}\right)}{s^2+s\left(\frac{g_{m1}}{C_1}+\frac{g_{m1}}{C_1}\right)+\frac{g_{m1}g_{m2}}{C_1C_2}}} \tag{4}$$

$$\frac{V_{04}(s)}{V_{IN}(s)} = \frac{1}{\frac{s\left(\frac{g_{m1}g_{m3}}{C_1g_{m4}}\right)}{s^2+s\left(\frac{g_{m1}}{C_1}+\frac{g_{m2}}{C_2}\right)+\frac{g_{m1}g_{m2}}{C_1C_2}}} \tag{5}$$

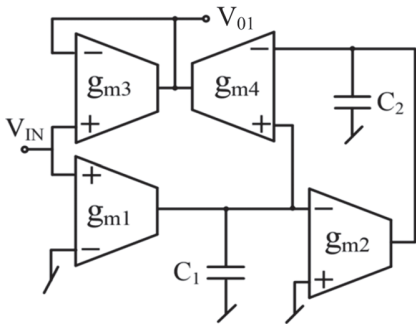
$$\frac{V_{05}(s)}{V_{IN}(s)} = \frac{1}{\frac{s^2+\frac{g_{m1}g_{m2}g_{m4}}{C_1C_2g_{m5}}}{s^2+s\frac{g_{m3}g_{m4}}{C_1g_{m5}}+\frac{g_{m1}g_{m2}g_{m4}}{C_1C_2g_{m5}}}} \tag{6}$$

From equations (2) - (6), we have obtained the pole frequency (ω), quality factor (Q) and gain (H) of proposed inverse filters which are given in Table I.

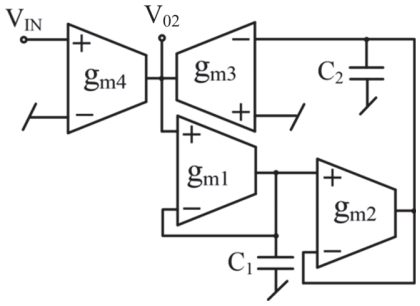
TABLE I: ω, Q and $gain$ of the various inverse filters shown in Fig. 2

Fig.	ω	H	Q
2(a)	$\sqrt{\frac{g_{m1}g_{m2}g_{m4}}{C_1C_2g_{m3}}}$	1	–
2(b)	$\sqrt{\frac{g_{m1}g_{m2}}{C_1C_2}}$	$\frac{g_{m3}}{g_{m4}}$	–
2(c)	$\sqrt{\frac{g_{m1}g_{m2}}{C_1C_2}}$	$1+\frac{g_{m3}}{g_{m4}}$	–
2(d)	$\sqrt{\frac{g_{m1}g_{m2}}{C_1C_2}}$	$\frac{g_{m1}g_{m3}C_2}{g_{m4}(g_{m1}C_2+g_{m2}C_1)}$	$\frac{\sqrt{g_{m1}g_{m2}C_1C_2}}{g_{m1}C_2+g_{m2}C_1}$
2(e)	$\sqrt{\frac{g_{m1}g_{m2}g_{m4}}{C_1C_2g_{m5}}}$	1	$\frac{1}{g_{m3}}\sqrt{\frac{C_2g_{m1}g_{m2}g_{m5}}{C_1g_{m4}}}$

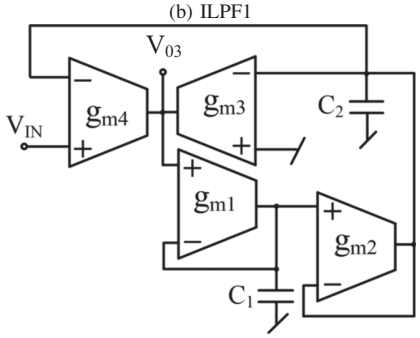
It may be noted that the non-inverting IBPF configuration shown in Fig. 2d can easily be converted into an oscillator by connecting the output and input terminals. From the Table I, it is observed that ω and Q are orthogonally adjustable; Q can be adjusted through the transconductance g_{m3} while ω is kept fixed. The gain of both ILPFs and IBPF can be tuned electronically independently through g_{m3} and g_{m4} respectively.



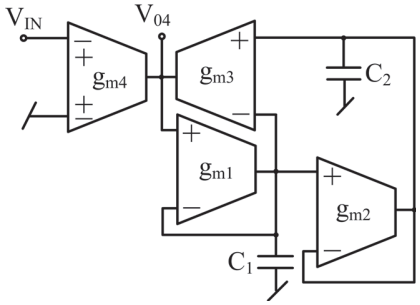
(a) IHPF



(b) ILPF1



(c) ILPF2

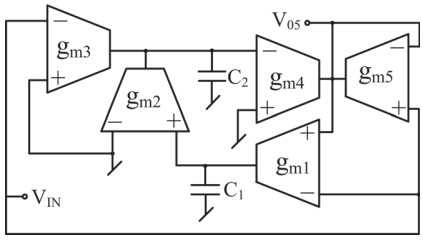


(d) IBPF

III. Non-ideal analysis

We have used a non-ideal equivalent circuit of the CMOS OTA employed in the proposed circuits in which the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:33:21 UTC from IEEE Xplore. Restrictions apply.



(e) IBRF

Fig. 2: Proposed inverse active filter configurations

parasitic impedance ($R_O \parallel \frac{1}{sC_O}$) of the output port has been considered. The characteristic polynomial of all the inverse filter configurations has been found to be of second order even under the influence of these parasitics. The following non-ideal expressions have been obtained for the cut-off frequency of the various inverse active filter configurations.

$$\omega_{ONIHP} = \sqrt{\frac{g_{m1}g_{m4}\left(g_{m2}+G_O\right)}{C_{N1}C_{N2}g_{m3}}} \tag{7}$$

$$\omega_{ONILP} = \sqrt{\frac{\left(g_{m1}+G_O\right)\left(g_{m2}+G_O\right)}{C_{N1}C_{N2}}} \tag{8}$$

$$\omega_{ONIBP} = \sqrt{\frac{\left(g_{m1}+G_O\right)\left(g_{m2}+G_O\right)}{C_{N1}C_{N2}}} \tag{9}$$

$$\omega_{ONIBR} = \sqrt{\frac{g_{m1}g_{m4}\left(g_{m2}+G_O\right)}{C_{N1}C_{N2}g_{m5}}} \tag{10}$$

where $C_{N1} = C_O + C_1$, $C_{N2} = C_O + C_2$.
We have used the values of $G_O = \frac{1}{R_O}$ and C_O from [42]. Using these values, the % error in ω for a nominal designed value of 1.59 MHz has been found to be 0.6284 x 10^{-5}, 1.257 x 10^{-5}, 1.257 x 10^{-5}, 0.6284 x 10^{-5} and which are very small. Hence, the effect of parasitic is not notable.

IV. PSpice Simulations And Matlab Evaluations

This section presents simulation results of the inverse filters with the help of PSpice simulation tool and MATLAB evaluations. For such verifications, we have used a CMOS OTA presented in Fig. 3 and its aspect ratios have been taken from [43]. For the operation of an OTA, power supplies of value ±0.9V were used. For different values of transconductance, bias voltage was varied from – 0.6V to – 0.78V.

A. Frequency Responses

The component values $g_{m1} = g_{m2} = g_{m3} = g_{m4} = g_{m5}$ 1mʊ and $C_1 = C_2$ =100 pF were used to design the inverse filters for nominal frequency 1.59 MHz. The input signal value was kept at 20 mV (pp). We have simulated all the circuits presented in Fig. 2 and found them to work satisfactorily. However to conserve space, in the following,

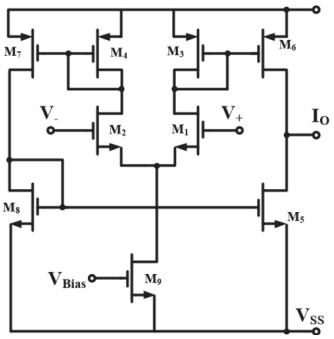
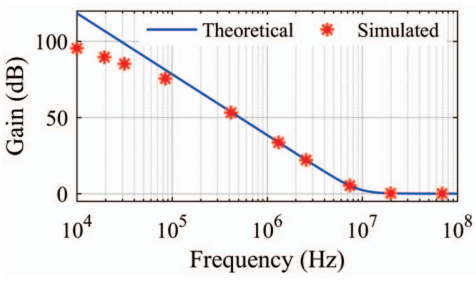
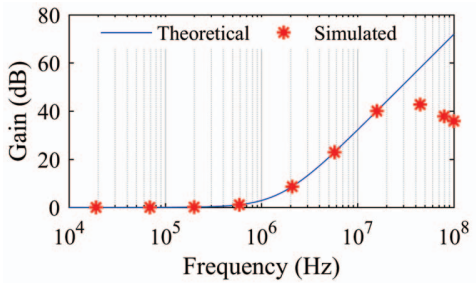


Fig. 3: Symbolic representation of an OTA [43].

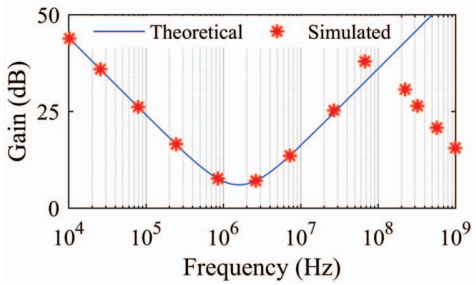
we have presented a few exemplary simulation results. The simulated and theoretical frequency responses of proposed inverse filters have been demonstrated in Fig. 4a - Fig. 4d respectively.



(a)

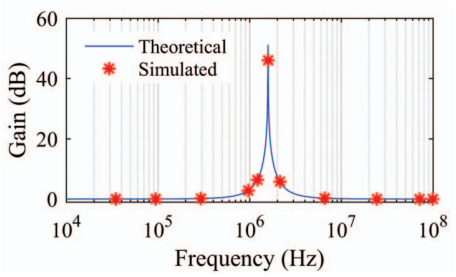


(b)



(c)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:33:21 UTC from IEEE Xplore. Restrictions apply.



(d)

Fig. 4: Frequency responses of (a) IHP, (b) ILP, (c) IBP and (d) IBR filters

B. Tunability of ω and BW

The orthogonal tunability of ω and Q for the IBR filter circuit was simulated to have a pole frequency of 1.59 MHz while Q was varied (by varying g_m of OTA_3). The simulated results are shown in Fig. 5.

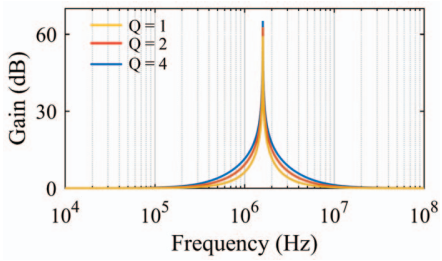
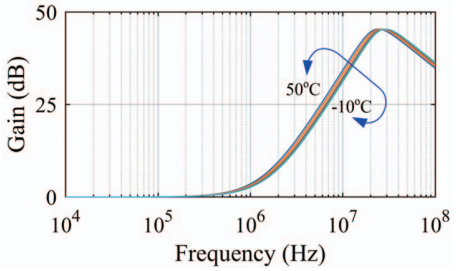


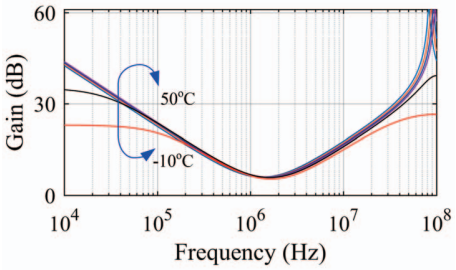
Fig. 5: Variation in Q for constant ω for IBRF

C. Temperature Analysis and Monte-Carlo Simulations

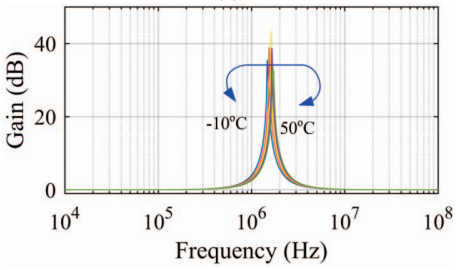
We have also simulated the circuits shown in Fig. 2 at different temperatures (-10°, 0, 10°, 20°, 30°, 40°, 50°) and the simulated results have been displayed in Fig. 6a - Fig. 6c. From these responses, the variation in frequency responses due to temperature variations is not significant. In Fig. 7,



(a)



(b)



(c)

Fig. 6: Frequency responses (a) ILP, (b) IBP, (c) IBR filters at different temperatures

we have shown the histogram of the Monte-Carlo simulations which has been carried out for IBR filter with variation in capacitors with a nominal tolerance of 5% for 100 runs. The deviation in frequency due to variation in tolerance capacitors is approximately 38 kHz which is not considerable.

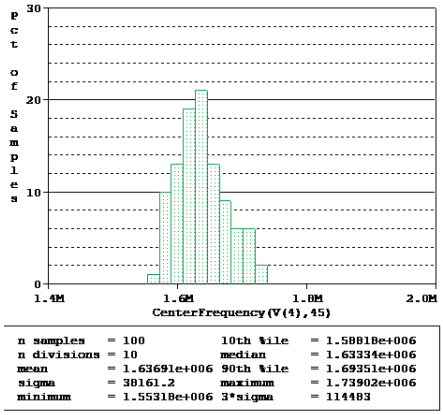


Fig. 7: Monte-Carlo simulations of IBRF of Fig. 2(d)

D. Sensitivity Analysis

The sensitivities of ω with respect to transconductors and capacitors have been evaluated using equation (11) and these

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:33:21 UTC from IEEE Xplore. Restrictions apply.

sensitivities are tabulated in Table II.

$$S_X^{\omega} = \left(\frac{X}{\omega}\right)\frac{\partial \omega}{\partial X} \tag{11}$$

where X is the parameter of interest (g_{mi} and C_i)

TABLE II: Active and passive sensivities of ω of various inverse filters

IHPF	$S_{g_{m1}}^{\omega}$=$S_{g_{m2}}^{\omega}$=$S_{g_{m4}}^{\omega}$ = 0.5, $S_{g_{m3}}^{\omega}$=$S_{C_1}^{\omega}$=$S_{C_2}^{\omega}$ = -0.5
ILPF	$S_{g_{m1}}^{\omega}$=$S_{g_{m2}}^{\omega}$ = 0.5, $S_{C_1}^{\omega}$=$S_{C_2}^{\omega}$ = -0.5
IBPF	$S_{g_{m1}}^{\omega}$=$S_{g_{m2}}^{\omega}$ = 0.5, $S_{C_1}^{\omega}$=$S_{C_2}^{\omega}$ = -0.5
IBRF	$S_{g_{m1}}^{\omega}$=$S_{g_{m2}}^{\omega}$=$S_{g_{m4}}^{\omega}$ = 0.5, $S_{g_{m5}}^{\omega}$=$S_{C_1}^{\omega}$=$S_{C_2}^{\omega}$ = -0.5

Table II, shows that the proposed inverse active filter circuits have low active and passive sensitivities.

V. Conclusions

Five new VM analog inverse active filter configurations have been introduced employing single output OTAs and two grounded capacitors.Different inverse filter functions viz. IHP, ILP, IBR and IBP using four/five OTAs have been presented. The IBR filter enjoys orthogonal electronic controllability between ω and Q. Non-ideal analysis with parasitic impedances has been carried out to justify the theoretical propositions and found small error between ideal and non-ideal pole frequencies. The sensitivity of the pole frequency with respect to transconductors and capacitors have also been evaluated and found to be less than 0.5. PSpice simulation tool and MATLAB evaluations have been used to validate the workabilty of presented inverse active filter circuits. Various simulation results such as frequency responses, Monte-Carlo results, temperature analysis have been included and these results were found to be in good agreement with the theoretical results. These new inverse active filters, thus, further enrich the repertoire of the OTA- based inverse active filters available in the open literature.

References

[1] A. Leuciuc, "Using nullors for realisation of inverse transfer functions and characteristics." Electron. Lett. vol. 33, pp. 949–950, 1997.

[2] D. R. Bhaskar, M. Kumar, P. Kumar, "Fractional order inverse filters using operational amplifier." Analog Integrated Circuits and Signal Processing, vol. 97, no. 1, pp. 149-158, 2018.

[3] T. J. Freeborn, A. S. Elwakil, B, Maundy, "Approximated fractional-order inverse Chebyshev lowpass filters", Circuits, Systems, and Signal Processing, vol. 35, no. 6, pp. 1973-1982, 2016.

[4] S. S. Gupta, D. R. Bhaskar, R. Senani, A. K. Singh, "Inverse active filters employing CFOAs." Electr. Eng. vol. 91, pp. 23-26, 2009.

[5] S. S. Gupta, D. R. Bhaskar, R. Senani, "New analogue inverse filters realised with current feedback op-amps." Int. J. Electron. vol. 98, no. 8, pp. 1103–1113, 2011.

[6] V. N. Patil, S. R. Desai, "Inverse Active Filters Employing CFOAs", LAP LAMBERT Academic Publishing, 2017, ISBN-10:3330063653

[7] H. Y. Wang, S. H. Chang, T. Y. Yang, P. Y. Tsai, "A novel multifunction CFOA-based inverse filter." Circuits Syst. vol. 2, pp. 14–17, 2011.

[8] V. N. Patil, R. K. Sharma, "Novel Inverse Active Filters employing CFOAs." International Journal for Scientific Research & Development. vol. 3, no. 7, pp. 359–360, 2015.

[9] K. Garg, R. Bhagat, B. Jaint, "Novel multifunction modified CFOA based inverse filter." Proc. IEEE Int. Conf. Power Electronics pp. 1–5, 2012.

[10] N. Herencsar, A. Lahiri, J. Koton, K. Vrba, "Realization of second-order inverse active filter using minimum passive components and DDCC." Proc. 33rd Int. Conf. Telecommunications and Signal Processing, pp. 38–41, 2010.

[11] R. Pandey, N. Pandey, T. Negi, V. Garg, "CDBA based universal inverse filter." ISRN Electron., 181869, 2011.

[12] N. A. Shah, M. F. Rather, "Realization of voltage-mode CCII-based all pass filter and its inverse version." Indian J. Pure Appl. Phys. vol. 44, pp. 269–271, 2006.

[13] N. A. Shah, M.A. Malik, "FTFN based dual inputs current-mode allpass inverse filters", Indian J Radio & Space Physics, vol. 34, pp. 206-209, 2005.

[14] A. R. Nasir, S. N. Ahmad, "A new current-mode multifunction inverse filter using CDBAs", International Journal of Computer Science and Information Security, vol. 11, no. 12, pp. 50-53, 2013.

[15] A. Sharma, A. Kumar, P. Whig, "On the performance of CDTA based novel analog Inverse Low Pass filter using 0.35 um CMOS parameter." Int J Science, Technology & Management. vol. 4, no. 1, pp. 594–601, 2015.

[16] N. A. Shah, M. Quadri, S. Z. Iqbal, "High output impedance current-mode all pass inverse filter using CDTA." Indian J. Pure Appl. Phys. vol. 46, pp. 893–896, 2008.

[17] A. Leuciuc, "The realization of inverse system for circuits containing nullors with applications in chaos synchronization." International Journal of Circuit Theory and Applications. vol. 26. no. 1, pp. 1–12, 1998.

[18] B. Chipipop, W. Surakampontorn, "Realisation of current mode FTFN-based inverse filter." Electron. Lett. vol. 35, no. 9, pp. 690–692, 1999.

[19] H. Y. Wang, C. T. Lee, "Using nullors for realization of current-mode FTFN-based inverse filters." Electron. Lett. vol. 35, no. 22, pp. 1889–1890, 1999.

[20] M. T. Abuelma'atti, "Identification of cascadable current-mode filters and inverse-filters using single FTFN." Frequenz, vol. 54, no. 11-12, pp. 284–289, 2000.

[21] A. K. Singh, A. Gupta, R. Senani, "OTRA-Based Multi-Function Inverse Filter Configuration." Advances in Electrical and Electronic Engineering, vol. 15, no. 5, pp. 846-856, 2018.

[22] P. Kumar, N. Pandey, S. K. Paul, "Realization of Resis-

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:33:21 UTC from IEEE Xplore. Restrictions apply.

torless and Electronically Tunable Inverse Filters Using VDTA." J Circuits, Syst Computers, 1950143, 2018.
[23] T. Tsukutani, Y. Sumi, N. Yabuki, "Electronically tunable inverse active filters employing OTAs and grounded capacitors." Int J Electron. Lett., vol. 4, no. 2, pp. 166–176, 2014.
[24] T. Tsukutani, Y. Kunugasa, N. Yabuki, "CCII-Based Inverse Active Filters with Grounded Passive Components", J. Electr. Eng., vol. 6, pp. 212-215, 2018.
[25] R. Bhagat, D. R. Bhaskar, P. Kumar, "Inverse Band Reject and All Pass Filter Structure Employing CMOS CDBAs", Int. J Engineering Research and Technology, vol. 08, no. 09, 2019.
[26] R. Bhagat., D. R. Bhaskar, P Kumar, "Multifunction Filter/Inverse Filter Configuration Employing CMOS CDBAs." Int. J. Recent Technol. Eng., vol. 8, no. 4, pp. 8844-8853, 2019
[27] A. Pradhan, R. K. Sharma, "Generation of OTRA-Based Inverse All Pass and Inverse Band Reject Filters", Proceedings of the National Acad. of Sci. India Section A: Physical Sciences, pp. 1–11, 2019
[28] S. Banergee, S. S. Borah, M. Ghosh, P. Mondal, "Three novel Configurations of Second Order Inverse Band Reject Filter using a Single Operational Transresistance Amplifier", TENCON-19-2019 IEEE Region 10 Conference, pp. 17-20, Oct.2019, Kochi, India
[29] D. Prasad, D. Tayal, A. Yadav,L. Singla, Z. Haseeb, "CNTFET Based OTRA and its Application as Inverse Low Pass Filter." International Journal of Electronics and Telecommunications, vol. 65, no. 4, pp. 665-670, 2019.
[30] D.V. Kamat, "New Operational Amplifier based Inverse Filters", In 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1177-1181, 2019.
[31] S. S. Borah, A. Singh, M. Ghosh, "CMOS CDBA Based 6th Order Inverse Filter Realization for Low-Power Applications." In 2020 IEEE REGION 10 CONFERENCE (TENCON), pp. 11-15, 2020
[32] E. M. Hamed, L. A. Said, A. H. Madian, A. G. Radwan, "On the approximations of cfoa-based fractional-order inverse filters." Circuits, Systems, and Signal Processing, vol. 9, no. 1, pp. 2-29, 2020.
[33] D.R. Bhaskar, M. Kumar, P. Kumar, "Minimal Realization of Fractional-Order Inverse Filters", IETE Journal of Research, pp. 1-14, 2020.
[34] J. Srivastava, R. Bhagat, P. Kumar, "Analog Inverse Filters Using OTAs", In 2020 6th International Conference on Control, Automation and Robotics (ICCAR), pp. 627-631, 2020.
[35] T. K Paul, S. Roy, R. R. Pal, "Realization of Inverse Active Filters Using Single Current Differencing Buffered Amplifier." Journal of Scientific Research, vol. 13, no. 1, pp. 85-99, 2021.
[36] P. Kumar, N. Pandey, S. K. Paul, "Electronically Tunable VDTA-Based Multi-function Inverse Filter." Iranian Journal of Science and Technology, Transactions of Electrical Engineering, vol. 45, no. 1, pp. 247-257, 2021.
[37] S. K. Jha, S. Aggarwal, S. Yadav, R. Bhagat, "Design of inverse filters using CFOAs." Advances in Electrical and Computer Technologies, Lecture Notes in Electrical Engineering, vol. 711, pp. 1153-1163, 2021.
[38] M. Kumar, D. R. Bhaskar, P. Kumar, "CFOA-Based New Structure of Fractional Order Inverse Filters", International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 5, pp. 2277-3878, 2020.
[39] T. K. Paul, S. Roy, R. R. Pal, "Realization of Inverse Active Filters Using Single Current Differencing Buffered Amplifier." Journal of Scientific Research, vol. 13, no. 1, pp. 85-99, 2021.
[40] M. A. Al-Absi, "Realization of inverse filters using second generation voltage conveyor (VCII)." Analog Integrated Circuits and Signal Processing, pp. 1-4, 2021.
[41] S. M. Al-Shahrani, M. A. Al-Absi, "Efficient Inverse Filters based on Second-Generation Voltage Conveyor (VCII)." Arab J Sci Eng, https://doi.org/10.1007/s13369-021-05775-5, 2021.
[42] A. Raj, D. R. Bhaskar, P. Kumar, "Two New Third-Order Quadrature Sinusoidal Oscillators." IETE Journal of Research pp. 1-14, 2021.
[43] D. R. Bhaskar, A. Raj, P. Kumar, "Mixed-mode universal biquad filter using OTAs." Journal of Circuits, Systems and Computers, vol. 29, no. 10, p.2050162, 2020.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:33:21 UTC from IEEE Xplore. Restrictions apply.

2021 Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 28-29, 2021

Heart Failure Prediction using Voting Ensemble Classifier

Shubham Kaushik
M.Tech (Signal Processing and Digital Design)
Delhi Technological University
Delhi, India
shubhamkaushik.sk08@gmail.com

Rajesh Birok
Associate Professor
Electronics and Communication Engineering
Delhi Technological University
Delhi, India
rbirok@gmail.com

***Abstract*- Heart is one of the most crucial organs present inside the human body which is responsible for providing our body with oxygen rich blood. Without a functioning heart human body cannot survive. This paper focuses on prediction of heart failure using machine learning algorithms and improving the accuracy of algorithms by using voting ensemble classifier. The experimental results show that the accuracy achieved by using voting ensemble classifier is more than that of individual machine learning algorithms.**

Keywords – Decision Tree, K Nearest Neighbor, Gaussian Naïve Bayes, voting classifier, accuracy.

I. INTRODUCTION

Cardiovascular Diseases (CVDs) are one of the leading causes of deaths worldwide. We refer CVDs to as a group of heart and blood vessel disorders, for example heart attack and stroke. Heart attack is basically an acute event which caused by the blockages in the blood vessels which prevents blood to flow to brain or heart[1]. Fatty deposits present inside the wall of blood vessels are the most common cause of the blockage. There are multiple causes of CVDs primary being tobacco use, alcohol consumption, hypertension, physical inactivity, and obesity[2]. An approximate of 18 million people died from CVDS in the year of 2016 alone, which represents an estimated of 31% of total global deaths[3]. Out of these, 85% were because of stroke or heart attack. Majority of deaths take place in developing and lower income nations. Out of 63% deaths in India because of CVDs in 2016, 27% deaths were because of CVDs. Identifying people at higher risk of CVDs in time can help in early diagnosis preventing millions of people from dying.

Prediction of heart ailments using current clinical techniques have unable to reach high prediction accuracy. Along with difficulty of achieving high accuracy, current clinical techniques are expensive. To overcome these drawbacks, researchers have been using various computational techniques, such as data mining and machine learning algorithms for disease prediction. Machine learning techniques are one of the fastest growing domains of artificial intelligence. Using machine learning algorithms in medical industry has proven to be very effective in early diagnosis of diseases accurately[4]. These techniques not only help in accurate prediction of disease but also help medical professionals to extract hidden underlying information from the data. The exponential increase in computational power over the years has allowed machine learning algorithms to work over large datasets efficiently. These techniques use datasets to train the different models and then use a trained model to evaluate the performance of the model by using various performance metrics.

The dataset used in this work was released by Ahmad et al. in July 2017[5]. They applied various biostatical techniques such as Cox Regression for survival prediction and as well as for feature selection from the dataset. They identified that older people, increase in serum creatinine, high blood pressure, increased level of anemia and lesser value of ejection fraction as the features which were responsible for increase in risk of mortality. They also concluded that gender played no significant role in identifying mortality among the patients.

Work done by Zaid et al.[6], they created two different gender-based models using the same dataset. One model was created for males and second model was created for the females. They applied biostatistics techniques on two models, and they concluded that survival prediction for males and females is significantly different. Likelihood test was used for testing goodness of fit for two different gender-based models.

David Chicco and Giuseppe[7] employed several machine learning algorithms for survival prediction using the same dataset. They also used algorithms along with biostatics techniques for evaluating feature importance. They concluded that using all the features from the dataset Decision Tree algorithm outperformed all the algorithms for obtaining the results on true positive. The Random Forest algorithm achieved top ROC. The linear support vector machine achieved low score on negative (sensitivity) elements but a higher score on positive (specificity) elements. They employed various other algorithms and compared the results. In the end they concluded ejection fraction and serum creatinine to be the most important features for accurately predicting the mortality among the patients.

In the first half of implementation in this paper we have employed Decision Tree, Gaussian Naïve Bayes and K Nearest Neighbor algorithm for survival prediction among the patients. While implementing K Nearest Neighbor algorithm, we have used elbow method for evaluating the value of K which results in the best accuracy. In the latter half of the implementation, we have used hard voting ensemble classifier by combing the algorithms used in the first half of the implementation. The voting ensemble classifier results in better accuracy than using the algorithms individually. As the stochastic algorithm generates a slightly different model every time we train it, we

2021 Asian Conference on Innovation in Technology (ASIANCON) | 978-1-7281-8402-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ASIANCON51346.2021.9544871

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:44:47 UTC from IEEE Xplore. Restrictions apply.

have taken average of 100 different simulations for comparing the performance.

The succeeding paper is compiled as follows; the dataset used in the implementation is explained briefly in the section 2. In section 3 proposed methodology is explained. Section 4 contains the results. Finally, the conclusion and the future work drawn has been written in section 5 and section 6 respectively.

II. DATASET

We analyzed the dataset which was collected between the months of April and December in the year of 2017 from the Faisalabad Institute of Cardiology and Allied Hospital, both in Faisalabad, Pakistan[5]. Medical record of 299 different patients is present in the dataset. The dataset comprises 194 males and 105 females. All the patients are of age between 40 and 95 years old. Every patient comes under the category III or IV of New York Heart Association (NHYA) stages of heart failure as all the patients had previously suffered heart failure[8] and had left ventricular systolic dysfunction. The dataset comprises 13 features out of which some features are numerical: age, Creatinine Phosphokinase (CPK), ejection fraction, platelets, serum creatinine, follow up time and serum sodium, rest of them are binary: blood pressure, sex, smoking, anemia, and smoking. Follow up time for patients was between 4 and 285 days, on an average it was 130 days. A patient with hematocrit level of less than 36% was having anemia. When a muscle tissue gets damaged, it can cause CPK to flow into the bloodstream of the body, which increases the level of CPK in the blood. The increased measure of CPK in the blood may show presence of heart injury[9]. Serum creatine value of greater than 1.5 is an indicator of renal dysfunction. Further on they divided ejection fraction values and platelets into three different categories, depending on their numerical value. The last attribute of the dataset, death event, denotes the mortality among the patients. The patient has survived if the value of death event is 0 and not survived if it is 1. Overall, data comprises 96 mortalities and 203 survived patients. In the implementation we have divided our dataset into 33% of test data and rest into training data. We have set the random state to a constant integer while splitting the dataset, and its value is same throughout the implementation so that same data is divided into training and test data for all the algorithms.

TABLE I. ATTRIBUTES IN THE DATASET

S. No.	Feature	Description of feature	Type
1	age	Age of patient (40-95)	Numerical
2	anemia	Decrease in hemoglobin	Binary
3	creatinine phosphokinase	Level of CPK in the bloodstream	Numerical
4	diabetes	If patient has diabetes or not	Binary
5	ejection fraction	Amount of blood leaving (in %)	Numerical
6	blood pressure	If patient has hypertension or not	Binary
7	platelets	Platelet count in blood	Numerical
8	serum creatine	Creatinine level in the bloodstream	Numerical
9	serum sodium	Sodium level in the blood	Numerical
10	sex	Male or female	Binary
11	smoking	If patient smoke's or not	Binary
12	time	Follow up time (4-285)	Numerical
13	death event	Survived or not	Binary

III. PROPOSED METHODOLOGY

The flow-chart for the proposed methodology is shown below:

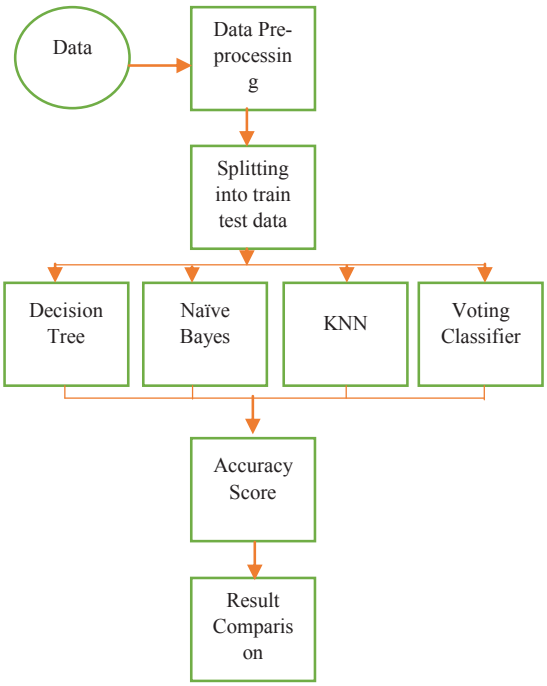


Fig. 1. Flowchart of methodology

In this work we have applied Decision Tree, K Nearest Neighbor algorithm and Naïve Bayes on our dataset for survival prediction among the patients. After that we have combined all the three algorithms using hard voting ensemble classifier and have compared the accuracy of ensemble classifier with the rest of the algorithms.

A. Data Collection and Pre-processing

Dataset used is this work is balanced enough to be used for classification purpose. Any two categories attribute is converted to binary before the implementation of algorithms by assuming dummy value of "0" or "1". Dataset is divided into 67% training data and 33% test data. Table 1 explains about each, and every feature used in the implementation.

B. Decision Tree Machine Learning Algorithm

As the name suggests, Decision Tree is a tree-based algorithm which starts from parent (root) node and ends at child (leaf) node[10]. It comes under the category of supervised machine learning algorithm as the model is trained using true labels which are given in the dataset. The parent node splits into multiple internal nodes which represents various features of the data. The decision made by the tree is denoted by the child nodes.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:44:47 UTC from IEEE Xplore. Restrictions apply.

Decision tree is very easy to interpret as compared to many other supervised algorithms. It does not require much of data pre-processing before the implementation.

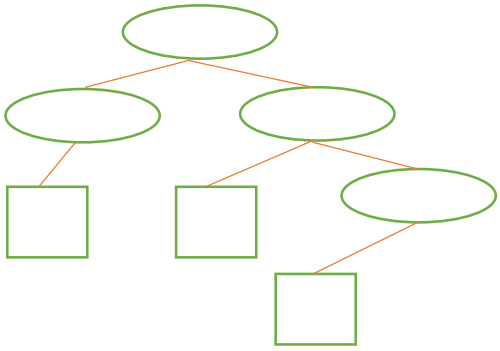


Fig. 2. Decision Tree representation

The decision tree is split based on the feature that results into maximum information gain after the split. Entropy is used to measure the purity of split in the tree. Entropy is given as:

$$H(s)= -P(q)\log_2 P(q)-P(r)\log_2 P(r) \quad (1)$$

Here P(q) and P(r) denotes the percentage of class q and class r respectively. Information gain denotes the decrease in the entropy after the splits has been done. It measures the reduction in entropy. More the information gain better will be our result. It is given as:

$$\text{Info. Gain}=H(R_n)- \sum (w_i)H(f_i) \quad (2)$$

$H(R_n)$ denotes the root entropy, w_i denotes the weighted average and $H(f_i)$ denotes the entropy of each feature.

C. K- Nearest Neighbor Machine Learning Algorithm

K-Nearest Neighbor algorithm (KNN) is one the most elegant, easy to implement, machine learning algorithm. Like decision tree it also uses true label to train the model and comes under the category of supervised algorithm. Unlike other machine learning algorithm KNN does not generates a discriminant function. It classifies the new data based on similarity with the previously loaded data[11], hence it is also known as a lazy learner algorithm. It assumes that data that is similar exist close to each other. It works very well when dataset is small.

K represents the number of neighbors. It is the parameter which denotes the number of nearest neighbors that algorithm includes during the classification process. For e.g., if K=3, KNN classifies new data point based on majority vote of its three closest neighbor.

In Fig. 3 we have two classes of circle (let A) and triangles (let B), each having three data points. KNN(K=3) will classify the new data point to class A as majority of the closest data point belongs to class A.

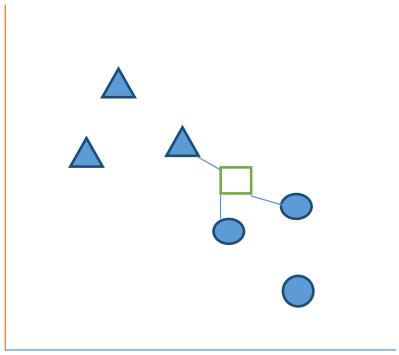


Fig. 3. KNN classification for K=3

K should be selected very wisely. Small values of K results in noise and inaccurate results whereas large value of K results in computationally expensive model. Elbow method can be used to compute the value of K which results in least error rate.

D. Naïve Bayes Machine Learning Algorithm

Naïve Bayes algorithm is a probabilistic, supervised machine learning algorithm. It is based on Bayes theorem. It assumes that all the features in the dataset are independent of each other[12], and each feature makes an equal contribution in prediction of output. The principle of Naïve Bayes is represented mathematically as:

$$P(r|s)= \frac{P(s|r)P(r)}{P(s)} \quad (3)$$

Posterior probability is denoted by P(r|s), P(s|r) denotes the likelihood probability, P(r) and P(n) denotes the class prior probability and marginal probability respectively. The classes are denoted by r and the features by s.

Naïve Bayes is widely used in prediction of multi class probabilities. With the assumption that each feature is independent, it performs exceptionally well as compared to other algorithms when training data is very small.

E. Voting Ensemble Classifier

In Ensemble learning we combine different models together to achieve a particular computational task such as regression or classification. Instead of using a single model, it combines various models together to achieve better performance.

Voting ensemble classifier is an ensemble model that combines the prediction from various other algorithms. It is used to improve the model by achieving better performance than any other algorithm used in voting ensemble. It works by combing the prediction from various other models[13]. We can use it for both regression and classification. In case of regression, it evaluates the average of the predictions from other single models. For classification, the label which has most vote is predicted. It is very useful when we have equally performing machine learning models, where it helps in balancing out their individual weaknesses. We can also combine multiple fit of a single algorithm with different hyper-parameters. One condition should be kept in mind while using the voting classifier is that it considers the contribution to

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:44:47 UTC from IEEE Xplore. Restrictions apply.

prediction from each model equally, which can cause some problem when one model performs well in one condition but poor in another.

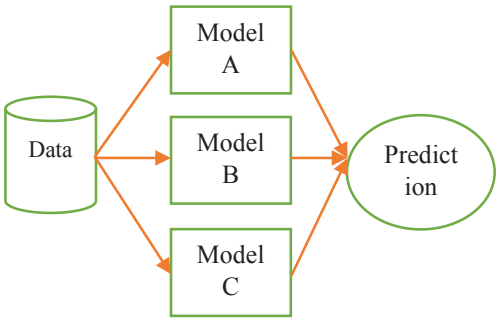


Fig. 4. Voting Ensemble Model

Classification using voting ensemble can be achieved using two different ways:

- Hard Voting: It is also known as majority voting. In Hard voting, each individual model votes for a particular class and the class with most votes wins[14]. In this work we have used hard voting for classification.
- Soft Voting: probability of data belonging to a particular class is evaluated by the given models in the soft voting[15]. In the end predictions are summed up. The class with the largest sum wins the vote.

We have used Python's machine learning library, Scikit-Learn for implementing different algorithms. Voting classifier is provided by 'VotingClassifier ()' class from the library[16]. In the voting classifier class, there are number of parameters that need to specify before implementation. The first and the most important parameter is the list of estimators that we have used, namely Decision Tree, KNN, and Naïve Bayes. Voting ensemble works well when we combine different stochastic models that result in different model every time it is trained on the dataset. The second most important parameter that needs to specified is the type of voting to be used while building the model. The type of voting could be hard voting or soft voting. If the type of voting is not mentioned, it takes hard voting by default. We can also weight the occurrence of predicted label before voting is performed by specifying the weight parameter, it takes up uniform weight if it is not None. After specifying the estimators in the voting classifier object, the object is trained on the training dataset using '.fit()' method. It trains the model by using the correct labels present inside the dataset.

IV. RESULTS

We first split the dataset into training and testing data then we trained three different models using Decision Tree, KNN and Naïve Bayes. Same dataset is used for training and testing for all the algorithms to maintain the uniformity by using the random state argument during splitting. Using test data, we evaluated the performance of each algorithm. In case of KNN algorithm we used elbow method to find out the optimal value of K by comparing the error rate for different values of K. It works by evaluating error rate for different values of K, generally up to 40. Value of K is selected carefully as large values can make our model computationally expensive.

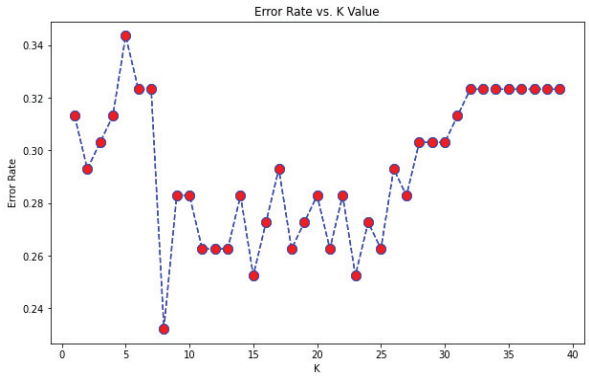


Fig. 5. Elbow Method Plot

From Fig. 5 we can see the error rate for different values of K as evaluated by the elbow method. Using this plot, we have selected value of K that gives use the least error rate.

We have used accuracy score as our performance metrics. We have implemented algorithms 100 different times and have taken an average of the accuracy score for each algorithm. Accuracy score shown in the result in rounded up to two decimal places.

In Table 2, we have shown a comparison between the mean accuracy score for all the algorithms. We can clearly see that Decision Tree, KNN (for optimal K value) and Naïve Bayes perform equally well for the given dataset with accuracy score of 0.77, 0.76 and 0.77 respectively, which make them an ideal candidate for voting ensemble classifier. Voting ensemble classifier combining all the individual algorithm performs significantly well with an accuracy score of 0.80.

TABLE II. COMPARISON BETWEEN THE ACCURACY OF DIFFERENT MODELS

Algorithm	Mean Accuracy Score
Decision Tree	0.77
K Nearest Neighbor	0.76
Naïve Bayes	0.77
Hard Voting classifier	0.80

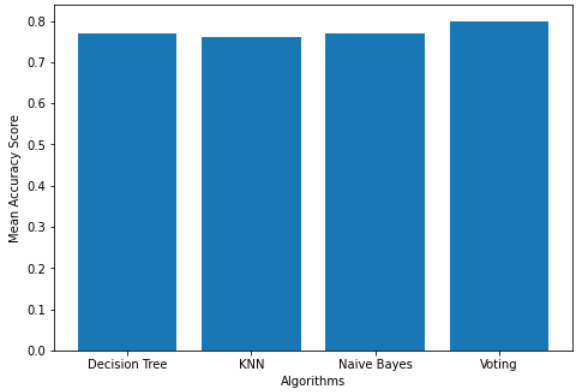


Fig. 6. Mean accuracy score plot

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:44:47 UTC from IEEE Xplore. Restrictions apply.

In Fig. 6, we have plotted bar plot for the accuracy. We can see the increase in accuracy for voting ensemble classifier. By combining similar performing algorithms, we have observed that accuracy of resultant ensemble model improves significantly.

V. CONCLUSION

Early prediction and diagnosis of heart disease is very important to prevent the deaths among the people. With a rise in heart related illness over the years, it is important to detect these diseases accurately. Machine learning techniques have proven to be effective in early detection of diseases. In this work, we have used machine learning algorithms for survival prediction among the heart failure patients. From the results we can clearly see the rise in accuracy for voting classifier. Hence, we can conclude that voting classifier for heart failure prediction comprising decision tree, KNN and naïve bayes as models improves the accuracy of the model. Voting ensemble classifier improves the accuracy of the model when estimators used perform equally well when used for classification.

VI. FUTURE WORK

In future work, we plan to implement data mining and machine learning techniques on a larger dataset. Larger dataset would enable to us achieve more reliable results. We will also implement other machine learning models and will compare the performance of those models with various other performance metrics. We will also use different ensemble techniques like bagging and stacking for further improvement in performance of our model.

REFERENCES

[1] M. B. Liu, "Cardiovascular diseases," *Chin. Med. J. (Engl).*, vol. 127, no. May 2017, pp. 6–7, 2014.

[2] A. Heart, "Heart failure projected to increase dramatically, according to new statistics," *Am. Hear. Assoc. News*, pp. 2018–2020, 2017, [Online]. Available: https://www.heart.org/en/news/2018/05/01/heart-failure-projected-to-increase-dramatically-according-to-new-statistics.

[3] K. J. Ho, "Cardiovascular diseases," *Nutr. Asp. Aging Vol. 2*, pp. 75–100, 2018, doi: 10.1201/9781351075145.

[4] S. J. Al'Aref *et al.*, "Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging," *Eur. Heart J.*, vol. 40, no. 24, pp. 1975–1986, 2019, doi: 10.1093/eurheartj/ehy404.

[5] T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab, and M. A. Raza, "Survival analysis of heart failure patients: A case study," *PLoS One*, vol. 12, no. 7, pp. 1–8, 2017, doi: 10.1371/journal.pone.0181001.

[6] F. M. Zahid, S. Ramzan, S. Faisal, and I. Hussain, "Gender based survival prediction models for heart failure patients: A case study in Pakistan," *PLoS One*, vol. 14, no. 2, pp. 1–10, 2019, doi: 10.1371/journal.pone.0210602.

[7] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, pp. 1–16, 2020, doi: 10.1186/s12911-020-1023-5.

[8] American Heart Association, "Classes of Heart Failure," *Am. Hear. Assoc.*, pp. 1–2, 2013, doi: 10.1109/IEMBS.2010.5627149.

[9] T. H. E. Conservation and O. F. T. H. E. W. S, "/ At H n-s ^ O," no. June 1992, pp. 1–4, 2021.

[10] J. Gehrke, "Classification and Regression Trees," *Encycl. Data Warehous. Min.*, pp. 246–280, 2011, doi: 10.4018/9781591405573.ch027.

[11] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2888, pp. 986–996, 2003, doi: 10.1007/978-3-540-39964-3_62.

[12] P. Langley, W. Iba, and K. Thompson, "Analysis of Bayesian classifiers," *Proc. Tenth Natl. Conf. Artif. Intell.*, pp. 223–228, 1992.

[13] F. Leon, S. A. Floria, and C. Badica, "Evaluating the effect of voting methods on ensemble-based classification," *Proc. - 2017 IEEE Int. Conf. Innov. Intell. Syst. Appl. INISTA 2017*, pp. 1–6, 2017, doi: 10.1109/INISTA.2017.8001122.

[14] M. Beyeler, "Understanding different voting schemes," pp. 7–9, 2021, [Online]. Available: https://www.oreilly.com/library/view/machine-learning-for/9781783980284/47c32d8b-7b01-4696-8043-3f8472e3a447.xhtml.

[15] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *Int. J. Cogn. Comput. Eng.*, vol. 2, no. January, pp. 40–46, 2021, doi: 10.1016/j.ijcce.2021.01.001.

[16] S. S. Edition, "Gorz ¡ fls Theory of the New W ork in g C la ss," pp. 2–5, 2010.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:44:47 UTC from IEEE Xplore. Restrictions apply.

2021 Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 28-29, 2021

High-Capacity Reversible Data Hiding using Deep Learning

Mohd Asif
Dept. of Electrical Engineering
Delhi Technological University
Delhi, India
imaasif28@gmail.com

Lokesh Kumar
Dept. of Electrical Engineering
Delhi Technological University
Delhi, India
imlokesh190@gmail.com

Gaurav Swami
Dept. of Electrical Engineering
Delhi Technological University
Delhi, India
gauravswami0999@gmail.com

Ankita Arora
Dept. of Electrical Engineering
Delhi Technological University
Delhi, India
arora.ankita09@gmail.com

***Abstract*— In today's time, it has become imperative to lay more focus on the security problems such as privacy, copyright protection, and tamper detection which become relevant with the transfer of digital data. A strategy that has recently gained popularity is reversible data hiding (RDH), which helps the recipient on the other end to retrieve confidential information that the sender had concealed in digital data such as photographs or videos. This is done with the help of image encryption and Deep Neural Network (DNN). Through this study we have tried to dive deeper into the domain of reversible data hiding to try and gauge its effect when used with deep learning techniques on the decryption end. Images containing secret data were encrypted and then a deep learning-based model was utilized at the receiver end to classify which of the images had been decrypted accurately along with extraction of the hidden data bits. We have achieved higher embedding capacity with DNN when compared to the existing Support Vector Machine (SVM) scheme.**

Keywords*— *Reversible data hiding; deep learning; Deep Neural Network; image encryption; embedding capacity.

I. INTRODUCTION

The process where message bits are embedded into digital contents like images or videos for protected data transmission is called data hiding. Traditional encryption processes transform simple plaintexts into incomprehensible data for the purpose of privacy preservation. Reversible Data Hiding (RDH) techniques, on the other hand, enable a sender to secure communication by inserting sensitive information into videos, photographs, or other types of digital material. The original material in which the data is hidden can then be extracted and recovered by the recipient of the coded message [1].

Lately, the mixture of data encryption and data hiding schemes has gained the attention of a considerable number of researchers. The Reversible Data Hiding techniques have an extensive amount of uses to it, such as in medical imagery where no deformation of original cover is permitted or military imagery where it is necessary to hide the data that is being transmitted. A selection of multimedia devices that hold the data of various users can be developed with the aid of encryption techniques for the purpose of tracing piracy. This method allows for the safe transfer of additional information by embedding the data into the digital content rather than sending it as a separate file.

The RDH techniques that currently exist are broadly divided into three verticals: Lossless Compression based RDH, Difference Expansion based RDH and Histogram Shifting based RDH. The procedure of lossless compression based RDH uses a lossless image compression technique to compress a part of the cover image and due to the compression, some unengaged space is produced which is then used for data concealment. The difference expansion process is used to insert a single bit of data using the difference between the pixel intensity values in the technique based on difference expansion. The additional bits of data are embedded in the peak pixel strength in the digital material using the histogram shifting technique. One of the major concerns of RDH is the visual standards after the process of data hiding is done in the image. The image quality evaluation method can be applied so that the visual quality can be measured after the hiding of the classified data is done in an image [2].

In images for the purpose of RDH, other completely different methods are also available. Confidentiality of images when they are being transferred are ensured with the help of image encryption which is an efficient tool for this purpose. These RDH procedures that encrypt images are broadly analyzed by researchers. Most of the RDH schemes that are currently in existence carry out the image encryption and the data hiding in two distinct phases.

In this scheme, we have integrated RDH, as well as, image encryption into one single process. The process of data hiding takes place when the image encryption procedure is done. In the scheme that is proposed here, any confidential image encryption can be utilized as there is no limitation to the image encryption process that is being used here. Data hiding by means of image encryption scheme is achieved due to the suggested reversible data hiding using an image encryption method Few of the most accepted RDH procedures for encrypted images have been seen in [2, 3].

J. Tian demonstrated a novel reversible data-embedding process for digital images and investigated digital image redundancy to achieve high embedding capability while minimizing distortion. He takes the advantage of the correlation between neighboring pixels which makes the low difference values between neighboring pixels, to hide the data in the pixels reversibly [3].

2021 Asian Conference on Innovation in Technology (ASIANCON) | 978-1-7281-8402-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ASIANCON51346.2021.9544626

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:37:44 UTC from IEEE Xplore. Restrictions apply.

W. Puech et al. demonstrated that data compression is needed to reduce transmission time, and for the past few years, various methods to combine compression, encryption, and data hiding in a single phase have been investigated, but a new challenge has emerged: embedding data in encrypted photos. Since the pixels in an encrypted image have no association, the he proposed analyzing the local standard decoder because the pixels in an encrypted image have no relation [4].

P. Tsai et al. proposed a reversible image hiding scheme for medical images dependent on histogram shifting the data was hidden using the histogram's peak in this technique. The hidden data was masked using the residual histogram of the predicted errors of the host image, and the similarity of neighboring pixels in the images was investigated using the prediction technique [5].

Xinpeng discussed that for encrypted images, a novel reversible data hiding scheme was suggested. They tried encrypting all of the data in an uncompressed image with a stream cypher and then embedding additional data into the image by modifying a small portion of the encrypted data. Using the data-hiding key, the embedded data can be successfully removed, and the original image can be perfectly retrieved using spatial similarities in a natural image [6].

K. Ma et al. proposed a novel approach in this paper to make it easier for the data hider to reversibly insert data in the encrypted image, the data hider had to reserve room before using a traditional RDH algorithm. Previous approaches used embedded data by reversibly vacating room from encrypted files, which could lead to data extraction and/or image restoration errors. The proposed method has the ability to achieve true reversibility, which means error-free data extraction and image recovery [7].

N. Nasrullah et al. proposed the Kd-tree used to build a joint and separable scheme for reversible data hiding (RDH) in compressed and encrypted photos. The hidden data is embedded in reserved spaces before being encrypted with several chaotic maps in the joint RDH scheme [8].

B. Ma et al. discussed the linear regression model to predict the value of pixels using neighboring pixels so that data can be hidden in any bit plane of the pixel and pixel recovery is done by using a regression model. They also tried to improve this technique by using multiple linear regression models for a single image instead of using just a single model. This ensures that the person who hides the data doesn't need access to the original image to do so. Also, the receiver can decrypt the image without getting secret data and get the secret data without decrypting the image which allows us to provide separate roles for the receiver. The problem with this technique is that with every image we have to train multiple regression models which can only be applicable only on a single image hence increasing time complexity. This model becomes very tedious when training for a large number of images and subsequent hidden data in them thus increasing the computational cost as well [9].

V. M. Manikandana et al. deals with the topic of reversible data hiding using machine learning. In this method, the authors hid messages in the form of bits by embedding them which were then extracted by the receiver on the other end using a Support Vector Machine (SVM) classifier to determine if a particular image had been accurately decrypted or not. In our approach we draw a comparison between the technique used by the authors and our technique where on the receiver end, we have tried to make use of a deep learning model instead of an SVM model. Through the course of the study, we find that the deep learning model proves to have better performance measures in comparison to the SVM model [10].

Hiding of the content that is being transmitted is ensured with the help of the scheme that is proposed here. This is done with the help of the image encryption and it enables the classified message to be embedded into the digital content. The principal idea to the suggested procedure is that non-overlapping portions are made of the cover image such that image blocks that are formed are then used to embed a single bit of the message that is needed to be transmitted. The encryption key that is to be used to embed a data bit into a particular image block is decided upon according to the data bit itself. This proposed scheme uses three different encryption keys. The receiver can decode the secret message as well as recover the original picture that has the actual authorization to access this data. A trained Deep Neural Network model used in this paper is being used to retrieve the original image as well as the data to be extracted from the obtained image. Experiments show that our method can embed nearly ten times as large payloads as previous methods while maintaining the same image quality. Advantage of our scheme over other methods is that we can embed or hide huge amount of data without compromising the image quality and proposed scheme also save effort and time by generating the features by its own instead of hand crafting them like in previous methods.

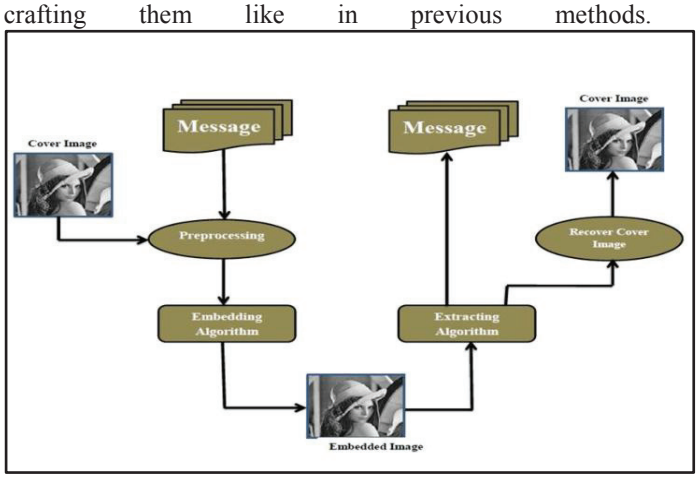


Fig. 1. Basic flow of data hiding scheme

II. PROPOSED METHODOLOGY

The overview of the proposed model is as follows. First the sender divides the image in blocks and starts encrypting the blocks according to the secret data. On completion, the sender sends the complete encrypted image to the receiver. The receiver divides the encrypted image into blocks and tries to decrypt the image. To identify the correct bit hidden in each block, sender uses a classifier. To get secret data and the original image, the receiver needs to decrypt all the blocks in the encrypted image. A quick rundown of the proposed RDH using DNN via encryption is shown in Fig. 2.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:37:44 UTC from IEEE Xplore. Restrictions apply.

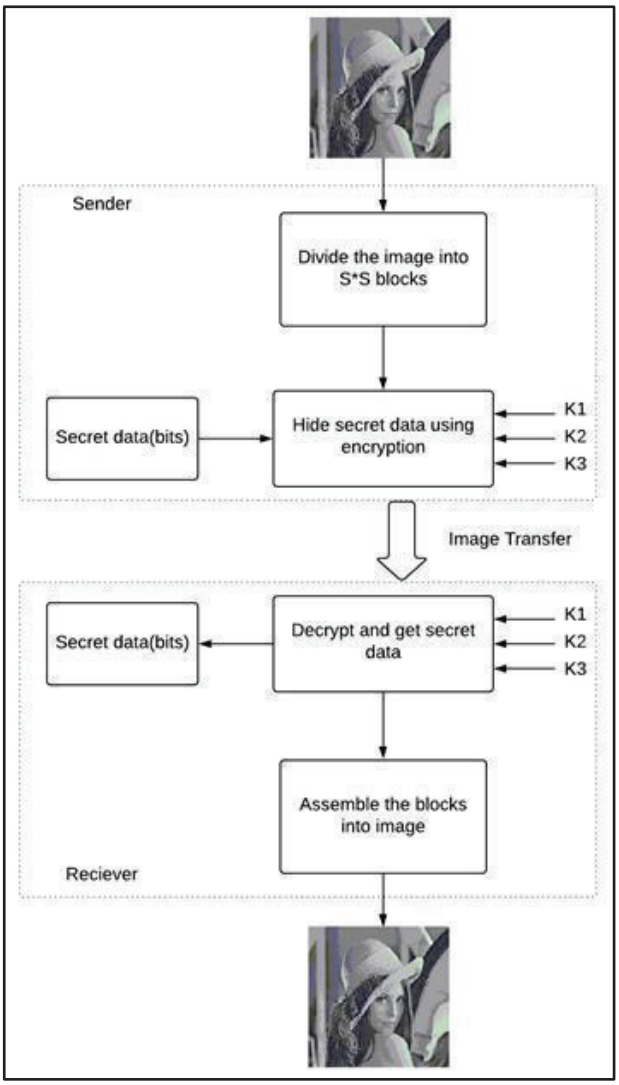


Fig. 2. Overview of the proposed RDH using DNN through encryption

A. Hiding Secret Data within Image

Sender needs an original image (IO), secret data (S), and three encryption keys (K0, K1 and K2) in order to insert the hidden data inside the image. The original picture IO is divided by the sender of size m × n pixels into blocks of size s × s pixels as shown in Fig. 3.

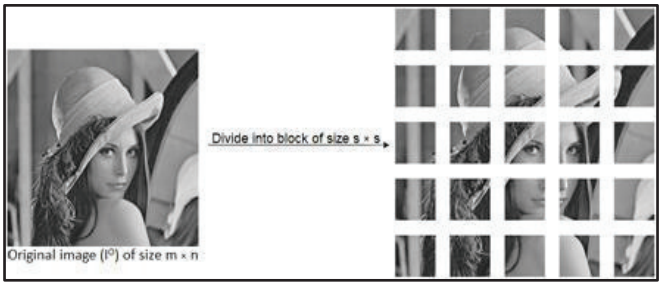


Fig. 3. Dividing the original image into blocks

Now, each block can be used to embed single bit of secret data using the following rule:

Rules for Encryption are as follows:

- For embedding 0 bit, use K0 Key to encrypt.
- For embedding 1 bit, use K1 Key to encrypt.
- For no embedding bit, use K2 Key to encrypt.

Sender applies this rule to each block of the image as shown in Fig. 4. To make this process more secure, the sender assessed the block in a pseudo-random sequence determined by the Data hiding key (KH). Sometimes the sender does not need to hide the secret bit within the block in that case K2 will be used for encryption.

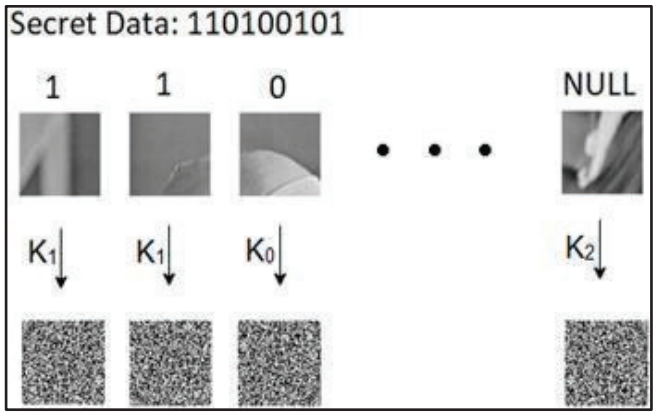


Fig. 4. Encryption of image blocks for embedding secret data

After the encryption process, the sender reassembles the encrypted block to get the single encrypted image (IE). Finally, the encrypted image is ready to send.

B. Recovery of Secret Data and Image

On receiving the encrypted image (IE), the receiver first divides it into blocks of size s × s pixels, the same as the sender. Each block embeds single bit of information which can be extracted out of the image by decrypting each block with their corresponding encryption key, but receiver does not know out of 3 keys (K1, K2 and K3) which key to use, so receiver tries to decrypt the block by all given keys, after decryption receiver will get three blocks out of which only one block is decrypted. Now as soon as the receiver finds out which block is decrypted, he can easily predict bit information embedded within the block as shown in Fig. 5.

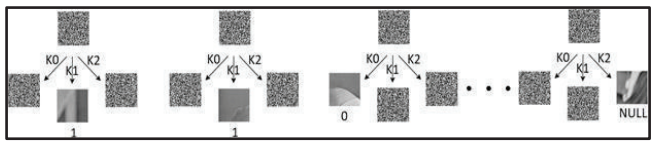


Fig. 5. Extracting embedded information from each block

By assembling all the information bits extracted from the image and image blocks, secret message and original image can be recovered.

C. Significance of Deep Neural Network classifier

While recovering the image and secret message, the receiver needs to classify out of three blocks (generated while decrypting the encrypted block with all three keys), which block is actually decrypted. A classifier is needed to solve this problem which can take a block of image and predict whether the block is decrypted (1) or not (0).

For the classification in this scheme, a Deep Neural Network is selected as shown in Fig. 6, since it can easily extract features from the image by itself while training.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:37:44 UTC from IEEE Xplore. Restrictions apply.

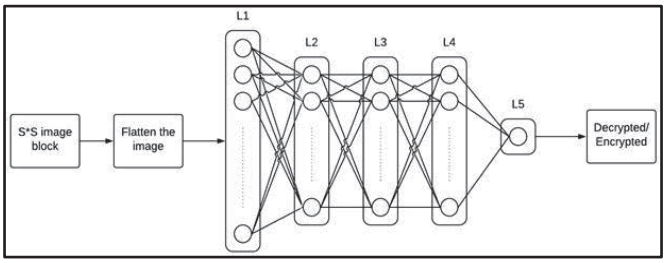


Fig. 6. DNN architecture for predicting image block is decrypted or not

In the proposed classifier, we have used L2 normalization to reduce overfitting. In the beginning of training, we have used a low learning rate so that it can explore a broader range of solutions and after a few iterations we rapidly increase the learning rate and train it for some iteration and finally start decreasing it. The details for DNN classifier are given below:

TABLE I. DETAIL FOR DNN CLASSIFIER

Layer	No. of Neurons	Activation function
L1	64	ReLU
L2	20	ReLU
L3	7	ReLU
L4	5	ReLU
L5	1	Sigmoid

III. RESULTS AND DISCUSSIONS

For the validation of results of this experiment study, we have used BossBase 1.01 image dataset with 10000 diverse grayscale images of size 512×512 pixels. In this section, the Confusion matrix and embedding ability of the proposed method are presented in table 2,3,4,5,6, and 7 for evaluation of the study's results. The comparison of the efficiency of the proposed method with the existing scheme in terms of the accuracy depicting the performance at block size as 4,5,6,7 and 8 is shown in Fig. 6.

A. Confusion Matrix

The confusion matrices of the DNN based extraction and recovery scheme introduced in this paper for block sizes 4,5,6,7 and 8 are shown in the following Tables II-VII.

TABLE II. BLOCK SIZE = 3

Actual \ Predicted	Not Decrypted	Decrypted
Not Decrypted	TN = 0.49578	FP = 0.00411
Decrypted	FN = 0.00389	TP = 0.49222

TABLE III. BLOCK SIZE = 4

Actual \ Predicted	Not Decrypted	Decrypted
Not Decrypted	TN = 0.05011	FP = 0.00375
Decrypted	FN = 0.00417	TP = 0.49097

TABLE IV. BLOCK SIZE = 5

Actual \ Predicted	Not Decrypted	Decrypted
Not Decrypted	TN = 0.4982	FP = 0.0036
Decrypted	FN = 0.004	TP = 0.4942

TABLE V. BLOCK SIZE = 6

Actual \ Predicted	Not Decrypted	Decrypted
Not Decrypted	TN = 0.4953	FP = 0.0042
Decrypted	FN = 0.0037	TP = 0.49679

TABLE VI. BLOCK SIZE = 7

Actual \ Predicted	Not Decrypted	Decrypted
Not Decrypted	TN = 0.50245	FP = 0.00571
Decrypted	FN = 0.00204	TP = 0.48979

B. Accuracy Comparison

Performance in terms of Accuracy of the method is shown in the following graph in Fig. 7. Here, Accuracy represents the ratio of wrongly extracted bits by the respective method to the originally embedded bits in the image.

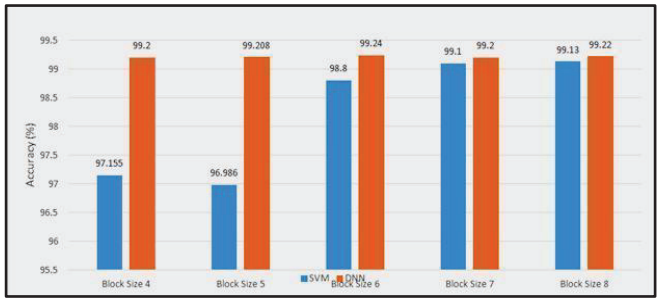


Fig. 7. Accuracy (%) comparison chart for SVM and DNN

As depicted by the graph for accuracy comparison, our approach with Deep Neural Network classification achieved consistent accuracy of above 99 percent which is significantly higher than the study using Support Vector Machine classification. While the SVM method achieved lower accuracy at a smaller block size because it wasn't able to capture the required features, our method efficiently captures the required features at all block sizes.

C. Embedding Capacity

Embedding capacity or Embedding rate of a data hiding scheme refers to the number of bits hidden or embedded into a single pixel of the digital image. Generally, the unit used for Embedding rate or Embedding capacity is bits per pixels. Embedding rate for our method is shown in the following table for different block sizes.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:37:44 UTC from IEEE Xplore. Restrictions apply.

TABLE VII. EMBEDDING CAPACITY(BPP)

BLOCK SIZE	4	5	6	7	8
EMBEDDING CAPACITY(BPP)	0.0625	0.04	0.0278	0.0204	0.0156

Through this approach, we have achieved higher embedding capacity with Deep Neural Networks when compared to the existing SVM scheme even for the low block size. It can be seen that our model outperforms the existing methods as it is able to attain higher embedding capacity without hampering the performance in terms of accuracy.

IV. CONCLUSIONS

In this paper, a scheme to extract data and recover the original image with it has been proposed. This is done with the help of image encryption and deep neural networks. This procedure proposes a way to ensure transmission of image in a secure manner. Rather than sending information as a distinct file, which can provide a way toward easy decryption by unauthorized groups, this scheme embeds the additional data into the image itself. Three separate encryption/decryption keys, that is, K1, K2 and K3, are required to be agreed upon by the sender, as well as the receiver and then a block-wise encoding procedure is done on the cover image. The encryption keys are ascertained by the message bits that are needed to be embedded into any particular block. The extraction of the confidential message along with the initial image that is to be recovered with the message are extracted by the receiver with the assistance of a trained Deep Neural Network model. The preliminary study of the proposed method on the BossBase 1.01 Image Dataset displays that this methodology surpasses the currently available methods with respect to its embedding rate and bit error rate. The assumption made in this RDH procedure is that the images are transferred through lossless channels, whereas there are quite a few instances where the transmission of data is done through lossy channels. Sturdier RDH methodologies can be proposed in the future to improve secure data transfer.

REFERENCES

[1] X. Xie, C. Chang, , C. Lin et al. “A Turtle Shell based RDH scheme with two-dimensional histogram shifting”, Multimedia Tools Appl 78, 19413–19436, 2019.

[2] J. Fridrich, M. Goljan, R. Du, “Invertible Authentication”, in Security and Watermarking of Multimedia Contents International Society for Optics and Photonics, pp. 197-209.

[3] J. Tian, “Reversible data embedding using a difference expansion," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no. 8, pp. 890-896, Aug. 2003.

[4] W. Puech, M. Chaumont, O. Strauss "A reversible data hiding method for encrypted images", Proc. SPIE 6819, Security, Forensics, Steganography, and Watermarking of Multimedia Contents X, 68191E , 18 March 2008.

[5] P. Tsai, Y.C. Hu, H.L Yeh, “Reversible image hiding scheme using predictive coding and histogram shifting”, Signal Processing, vol.89, no. 6, pp 1129-1143, 2009.

[6] X. Zhang, "Reversible Data Hiding in Encrypted Image," in IEEE Signal Processing Letters, vol. 18, no. 4, pp. 255-258, April 2011.

[7] K. Ma, W. Zhang, X. Zhao, N. Yu and F. Li, "Reversible Data Hiding in Encrypted Images by Reserving Room Before Encryption," in IEEE Transactions on Information Forensics and Security, vol. 8, no. 3, pp. 553-562, March 2013.

[8] N. Nasrullah, J. Sang, M. Mateen et al. “Reversible data hiding in compressed and encrypted images by using Kd-tree”, Multimedia Tools Appl 78, 17535–17554, 2019.

[9] B. Ma, X. Wang, Q. Li et al. “Adaptive error prediction method based on multiple linear regression for reversible data hiding”, J Real-Time Image Proc 16, 821–834, 2019.

[10] V. M. Manikandana, V. Masilamani, “Reversible Data Hiding Scheme During Encryption Using Machine Learning”, International Conference on Robotics and Smart Manufacturing (RoSMa2018).

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:37:44 UTC from IEEE Xplore. Restrictions apply.

arXiv:2109.13767v1 [cs.CL] 28 Sep 2021

Identifying and Mitigating Gender Bias in Hyperbolic Word Embeddings

Vaibhav Kumar* **Tenzin Singhay Bhotia*** **Vaibhav Kumar***
Delhi Technological University, India
{kumar.vaibhav1o1, tenzinbhotia0, vaibhavk992}@gmail.com

Tanmoy Chakraborty
Indraprastha Institute of Information Technology Delhi
tanmoy@iiitd.ac.in

Abstract

Euclidean word embedding models such as GloVe and Word2Vec have been shown to reflect human-like gender biases. In this paper, we extend the study of gender bias to the recently popularized hyperbolic word embeddings. We propose **gyrocosine bias**, a novel measure for quantifying gender bias in hyperbolic word representations and observe a significant presence of gender bias. To address this problem, we propose **Poincaré Gender Debias** (**PGD**), a novel debiasing procedure for hyperbolic word representations. Experiments on a suit of evaluation tests show that PGD effectively reduces bias while adding a minimal semantic offset.

1 Introduction

Word embeddings are often used for word representation in a wide variety of tasks. However, they have been shown to acquire gender stereotypes from the data which they are trained on (Bolukbasi et al., 2016; Caliskan et al., 2017). For instance, Bolukbasi et al. (2016) showed that a Word2Vec (Mikolov et al., 2013a) model trained on Google news corpus[1], generates analogy such as "man is to computer programmer as woman is to homemaker". Therefore, given their pervasive usage, a large number of debiasing methods (Bolukbasi et al., 2016; Kaneko and Bollegala, 2019; Bordia and Bowman, 2019; Zhao et al., 2018; Maudslay et al., 2019) have recently been proposed.

The existing literature on mitigating gender bias exclusively focuses on word vectors which are embedded in an Euclidean space. However recently, non-Euclidean space has also been investigated for embedding words. For instance, motivated by the latent hierarchical structure of words, Tifrea et al. (2018) generalize GloVe (Pennington et al., 2014) to the hyperbolic space. As a result, they obtain Poincaré GloVe embedding, which achieves superior performance w.r.t GloVe in analogy, hypernymy detection, and word similarity tasks simultaneously. Despite the performance gain achieved, we observe that the issue of gender bias persists, as indicated by the biased analogies shown in Table 1.

Gender Biased Analogies	
man → doctor	woman → nurse
male → supervisor	female → assistant
man → programmer	woman → storyteller
he → surgeon	she → nurse
man → chairman	woman → secretory

Table 1: Gender biased analogies generated by Poincaré GloVe (Tifrea et al., 2018) trained on the English Wikipedia dump containing 1.4 billion tokens.

In order to address this issue, we extend the study of gender bias to the hyperbolic space. We introduce **gyrocosine bias**, a metric for quantifying the differential association of a given gender neutral word with a set of male and female definitional words using *gyrovector-based* (Ungar, 2008) *formalism*. Further, we propose **Poincaré Gender Debias** (PGD), a post-processing method for debiasing words embedded on the hyperbolic space. Experiments on a suite of evaluation tests show that PGD successfully mitigates gender bias while introducing minimal semantic and syntactic offset. *To the best of our knowledge, this is the first work that addresses the issue of gender bias in hyperbolic word embeddings.*

2 Preliminaries

Hyperbolic Space and Poincaré Ball. An n-dimensional Hyperbolic space $\mathbb{H}^n$ is a non-Euclidean space with a constant negative curvature. Since a hyperbolic space cannot be embedded into an Euclidean space (Krioukov et al., 2010), several isometric models have been proposed. Fol-

*Authors have contributed equally.

[1] https://code.google.com/archive/p/word2vec/

lowing Tifrea et al. (2018), we use the Poincaré ball model of hyperbolic space. Formally, a Poincaré ball $\left(\mathbb{D}_c^n, g^{\mathbb{D}}\right)$ of a radius $1/\sqrt{c}, c > 0$ is defined by an n-dimensional manifold:

$$\mathbb{D}_c^n = \{x \in \mathbb{R}^n : c\|x\| < 1\}$$

having the Riemannian metric $g_x^{\mathbb{D}}$ as $g_x^{\mathbb{D}} = \lambda_x^2 g^E$, where $\lambda_{\mathbf{x}} = 2/\left(1 - c\|\mathbf{x}\|^2\right)$ is the conformal factor, and $g^E = \mathbf{I}_n$ is the Euclidean metric tensor.

Gyrovector Space. Gyrovector space provides the algebraic setting for hyperbolic geometry (analogous to vector space for Euclidean geometry) (Ungar, 2008). We use gyrovector operations to quantify the bias in hyperbolic space. A brief description of the concepts used is given below[2].

Möbius Addition and Subtraction. The Möbius addition for $x, y \in \mathbb{D}_c^n$ is defined as:

$$x \oplus_c y = \frac{\left(1+2c\langle x,y\rangle+c\|y\|^2\right)x+\left(1-c\|x\|^2\right)y}{1+2c\langle x,y\rangle+c^2\|x\|^2\|y\|^2}$$

where $\|\cdot\|$ denotes the Euclidean norm, and $\langle\cdot,\cdot\rangle$ is the Euclidean dot product. The Möbius subtraction $\ominus$ is then defined as $a \ominus z = a \oplus (-z)$. Note that in all cases, $c > 0$ can be reduced to $c = 1$ without any loss of generality. We refer to $\oplus_1$ as $\oplus$.

Rooted Gyrovectors. Any two ordered points $x, y \in \mathbb{D}^n$ give rise to a unique rooted gyrovector $v = \ominus x \oplus y$, which has its tail as x and head as y. Additionally, any $z \in \mathbb{D}^n$ can be identified as a rooted gyrovector with its tail at origin O and head at z, which we represent as: $z' = O \oplus z$.

Gyrocosine Function. The gyrocosine is a measure of the gyroangle α ($0 \leq \alpha \leq \pi$) between two non-zero rooted gyrovectors $\ominus A_1 \oplus B_1$ and $\ominus A_2 \oplus B_2$, as given by:

$$\cos\alpha = \frac{\ominus A_1 \oplus B_1}{\|\ominus A_1 \oplus B_1\|} \cdot \frac{\ominus A_2 \oplus B_2}{\|\ominus A_2 \oplus B_2\|}$$

Riemannian Optimization. Let $f : \mathcal{M} \to \mathbb{R}$ be a smooth function to be optimized over the Riemannian manifold $(\mathcal{M}, p)$. The Riemannian stochastic gradient descent update (Bonnabel, 2013) is then defined as:

$$x_{t+1} \leftarrow exp_{x_t}(-\alpha g_t)$$

where $g_t \in T_{x_t}\mathcal{M}$ is the Riemannian gradient of f_t at $x_t \in \mathcal{M}$, and $\alpha > 0$ is the learning rate. In this work, we use the Riemannian ADAM optimizer (Bécigneul and Ganea, 2018)[3].

[2]For a complete discussion, refer to (Ungar, 2008).

[3]We use the implementations provided by the open source library: geoopt (Kochurov et al., 2020)

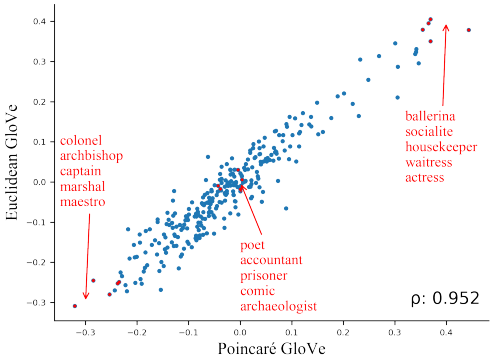


Figure 1: Comparison of gender bias across GloVe and Poincaré GloVe (vectors of size $100D$ and trained on the English Wikipedia dump); x-axis denotes gyrocosine bias, and y-axis denotes direct bias. Each point in the scatter plot indicates a profession-related word.

3 Gender Bias and its Mitigation

3.1 Bias in Hyperbolic Word Embeddings

In this section, we first propose a pair of rooted gyrovectors to capture the differential gender information in the embedding space and then use them to quantify bias for any arbitrary word vector.

Gender Gyrovectors. Given a set of words $\mathbb{W}$ embedded in an n-dimensional Poincaré ball $\mathbb{D}^n$, the gender gyrovectors are defined as follows:

$$g_{mf} = \ominus\mu_M \oplus \mu_F$$
$$g_{fm} = \ominus\mu_F \oplus \mu_M$$

where, μ_F and μ_M denote the intrinsic mean (Karcher, 1977; Fréchet, 1948) of a set of male definitional and female definitional hyperbolic word representations[4] respectively (see Appendix 1.1 for details). We propose two gyrovectors for capturing the gender information because of the non-commutative property of the binary Möbius addition operator.

Gyrocosine Bias. For a gender-neutral word w embedded in an n-dimensional Poincaré ball $\mathbb{D}^n$, the gyrocosine bias is defined as follows:

$$\gamma(w) = \frac{\cos(w', g_{mf}) - \cos(w', g_{fm})}{2}$$

where $\cos(.)$ is the gyrocosine function and w' is the rooted gyrovector for w as explained in Section 2. Here, a positive value of $\gamma(w)$ implies that w is female biased, while a negative value indicates male bias. To validate our metric w.r.t

[4]We use the same set of gender definitional words as (Bolukbasi et al., 2016).

its Euclidean counterpart, we compare the gender bias for a set of profession words [5] across GloVe and Poincaré GloVe in Figure 1, using direct bias (Bolukbasi et al., 2016) and gyrocosine bias respectively. We obtain a high Pearson correlation ($\rho = 0.952$) which indicates the competence of our metric in quantifying gender bias.

3.2 Proposed Bias Mitigation Method

We propose Poincaré Gender Debias (PGD), a multi-objective optimization based method for reducing gyrocosine bias in hyperbolic word embeddings while keeping the embeddings usable. To obtain the debiased counterpart w_d of a given word w, PGD solves the following multi-objective optimization problem:

$$argmin_{w'_d}\big(F_s(w'_d), F_g(w'_d)\big)$$

We use the weighted sum method to solve this optimization problem. More formally, we minimize the single objective function $F(w'_d)$ which can be expressed as follows:

$$\begin{aligned} F(w'_d) \;&=\; \lambda_1.F_s(w'_d) + \lambda_2.F_g(w'_d) \\ \text{such that } \lambda_i \in [0,1] \text{ and } &\sum_i \lambda_i = 1 \end{aligned} \tag{1}$$

$F(w'_d)$ is minimized using Riemannian optimization as described in Section 2. We use the learning rate of 3×10^{-4} and equal objective weights, $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$. We now explain each component of Equation 1; the range of each can be observed to be $[0, 1]$.

$F_g(w'_d)$: This objective function aims to equalize word gyrovectors with the notion of male and female gender in the embedding space. It does so by minimizing the absolute difference of the gyrocosine values of the given word gyrovector and the two gender gyrovectors. The objective function is defined as follows:

$$F_g(w'_d) = \big|\cos(w'_d, g_{mf}) - \cos(w'_d, g_{fm})\big|\,/2$$

$F_s(w'_d)$: This objective function aims to minimize the semantic offset caused by the debiasing procedure by maximizing the gyrocosine value between the rooted gyrovectors for a given word and its debias counterpart, thereby ensuring that the gyroangle between the two is minimum. The objective is defined as follows:

$$\begin{aligned} F_s(w'_d) &= |\cos(w'_d, w') - \cos(w', w')|/2 \\ &= |\cos(w'_d, w') - 1|/2 \end{aligned}$$

[5] We use the list of professions provided by Bolukbasi et al. (2016).

4 Experiments and Results

4.1 Dataset and Baselines

We use the following pre-trained word embedding models provided by Tifrea et al. (2018) as our baseline, each of which is trained on the English Wikipedia dump and consists of 189,533 tokens: $100D$ Poincaré GloVe, $h(x) = \cosh^2(x)$, with init trick[6] – referred to as **P-GloVe**; $100D$ Euclidean GloVe, with init trick – referred to as **E-Glove**. We debias Poincaré GloVe using PGD to obtain **PGD-GloVe**. Similiar to Bolukbasi et al. (2016), we split the vocabulary into a set of gender specific S and gender neutral words N. Only N, consisting of $126,065$ words, is debiased.

4.2 Evaluation Metrics and Results

Word Embedding Association Test (WEAT): Caliskan et al. (2017) introduce WEAT, a hypothesis testing method for analyzing stereotypical biases present in word embeddings. WEAT comprises two sets of equal-sized target words – X (like 'engineer', 'warrior') and Y (like 'nurse', 'receptionist'), and two sets of equal sized attribute words – A (like 'he', 'male') and B (like 'she', 'female') [7]. The aim of the test is to determine if the set X or Y is more biased towards one gender than the other. We report Cohen's d and p-value for three categories of target words in Table 2(a). Evidently, PGD-GloVe achieves consistent performance for all sets of target words, reflecting the efficient reduction in bias (see Appendix 2.2 for more details). Further, we observe that Poincaré GloVe is more biased than Euclidean GloVe, having a lower p-value and a higher Cohen's d for all the target word categories. Since, the training data and variable configurations like the dimensions and initialization were same, a higher bias could be due to the negative curvature of hyperbolic space and would remain as an interesting future direction.

SemBias: SemBias (Zhao et al., 2018) is a word analogy dataset used for measuring the ability of word embeddings to generate biased analogies. Each instance of SemBias consists of four types of word pairs – gender-definition word pair (**Def**; e.g., 'wizard-witch'), gender-stereotype word pair (**Ster**; e.g., 'doctor-nurse') and two other word pairs having similar meanings but no gender-based relation (**None**; e.g., 'salt-pepper'). For each instance, we

[6] Refer Tifrea et al. (2018).

[7] Refer Appendix 2.2 for the complete list of words.

Embedding	(a) WEAT						(b) SemBias		
	B_1: Career vs Family		B_2: Math vs Art		B_3: Science vs Art		Def ↑	Ster ↓	None ↓
	$p\uparrow$	$d\downarrow$	$p\uparrow$	$d\downarrow$	$p\uparrow$	$d\downarrow$			
E-Glove	0.0773	0.7423	0.4186	0.1716	0.5443	-0.1131	79.5	14.8	**5.7**
P-GloVe	0.0418	0.8929	0.3045	0.2929	0.2138	0.4568	75.3	7.0	17.7
PGD-GloVe	**0.2827**	**0.3054**	**0.8292**	**-0.5379**	**0.6015**	**-0.1984**	**83.0**	**0.0**	17.0

Table 2: Comparison of different embedding models on (a) WEAT for different sets of target words - B_i and (b) SemBias dataset. ↑ (↓) indicates that higher (lower) value is better. In (a), a p-value > 0.5 denotes insignificant bias, and a lower Cohen's d shows lesser bias. We use (negative) Poincaré distance as a measure of similarity for P-GloVe and PGD-GloVe (see Appendix 2.1 for details).

Embedding	(a) Semantic						(b) Analogy		
	RW	WordSim	SimLex	SimVerb	MC	RG	Google-Sem	Google-Syn	MSR
E-Glove	0.3787	0.5668	0.2964	0.1639	0.6562	0.6757	0.6427	0.5950	0.4826
P-GloVe	**0.4190**	**0.6207**	**0.3210**	**0.1915**	**0.7883**	**0.7596**	0.6629	**0.6086**	**0.4958**
PGD-GloVe	0.4170	0.6152	0.3215	0.1914	0.7873	0.7569	**0.6635**	0.6069	0.4938

Table 3: Comparison of different embedding models in (a) semantic and (b) analogy tests. The performance of PGD-GloVe is at par with other baselines, indicating the addition of minimal semantic disturbance.

choose the most probable analogy estimated by the embedding model. The analogy computation for word embeddings in Poincaré ball model uses a hyper-parameter [8] $t \in [0, 1]$ (Tifrea et al., 2018). Following Levy et al. (2015); Tifrea et al. (2018), we use 2-fold cross-validation over the gender-definition word pair analogies to estimate the value of t as 0.3. From Table 2(b), we can observe that **Ster** analogies are completely inhibited for PGD-GloVe along with a 10.2 % relative improvement in **Def** w.r.t P-GloVe. This implies that PGD dramatically mitigates the preference of a gender-neutral word between male and female words, thereby improving its efficacy to form correct analogies. We further observe that P-GloVe has a lesser (**Ster**) score as compared to E-GloVe. However, this result is attained by a lower performance in **Def** and **None**, indicating the superiority of E-GloVe.

Semantic Tests: Word semantic similarity tests determine the correlation between the similarity of word pairs obtained through the word embeddings and the similarity ratings given by humans. We report Spearman's correlations for the following benchmark datasets: RW (Luong et al., 2013), WordSim (Agirre et al., 2009), SimLex (Hill et al., 2015), SimVerb (Gerz et al., 2016), MC (Miller and Charles, 1991), and RG (Rubenstein and Goodenough, 1965). Following Tifrea et al. (2018), we use the negative Poincaré distance as the similarity metric for P-GloVe and PGD-GloVe while for E-GloVe, we use the regular cosine similarity. Table 3(a) shows that PGD-GloVe successfully retains most of semantic information. Therefore, PGD produces minimal semantic offset.

Analogy Tests: Analogy tests assess the efficacy of a word embedding model to answer the following question : 'w_1 is to w_2 as w_3 is to ?'. Here, unknown word w_4 is estimated as the closest word vector to the arithmetic: $w_2 - w_1 + w_3$. Following Tifrea et al. (2018), we estimate the closest word vector w_4 using gyro-translations and the cosine similarity metric for the following benchmarks: Google Sem, Syn (Mikolov et al., 2013b) and MSR (Mikolov et al., 2013c). Table 3(b) shows that PGD-GloVe closely follows P-GloVe in all the analogy measures. The offsets are so minimal that PGD-GloVe retains the semantic and analogy advantages of P-GloVe over E-GloVe.

5 Conclusion

In this paper, we presented the first study of gender bias in a non-Euclidean space. We proposed **gyrocosine bias**, a metric for quantifying gender bias in hyperbolic embedding. We observed that stereotypical gender biases permeate the Poincaré GloVe model. We also proposed **Poincaré Gender Debias**, a post-processing method for debiasing hyperbolic word embeddings. Experimental results indicated that our method successfully minimizes gender bias while retaining the practical usability of word embeddings.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A

[8] Refer Appendix 2.3 for more details.

study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27, Boulder, Colorado. Association for Computational Linguistics.

Gary Bécigneul and Octavian-Eugen Ganea. 2018. Riemannian adaptive optimization methods. *arXiv preprint arXiv:1810.00760*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Silvere Bonnabel. 2013. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. *CoRR*, abs/1904.03035.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Maurice Fréchet. 1948. Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l'institut Henri Poincaré*, volume 10, pages 215–310.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simverb-3500: A large-scale evaluation set of verb similarity. *arXiv preprint arXiv:1608.00869*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 1641–1650.

Hermann Karcher. 1977. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541.

Max Kochurov, Rasul Karimov, and Sergei Kozlukov. 2020. Geoopt: Riemannian optimization in pytorch. *arXiv preprint arXiv:2005.02819*.

Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguná. 2010. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Minh-Thang Luong, Richard Socher, and Christopher D Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 5267–5275.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013,Workshop Track Proceedings*, pages 1–12.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Herbert Rubenstein and John B Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. 2018. Poincare glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*.

Abraham Albert Ungar. 2008. A gyrovector space approach to hyperbolic geometry. *Synthesis Lectures on Mathematics and Statistics*, 1(1):1–194.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. *Proceedings of the 2018 Conference*

on Empirical Methods in Natural Language Processing, page 4847–4853.

arXiv:2109.13767v1 [cs.CL] 28 Sep 2021

Appendix

1 Poincare Gender Debiasing

1.1 Intrinsic Mean Calculation

To compute gender direction gyrovectors, we calculate the the intrinsic means μ_F and μ_M for a set of male and female definitional words which are obtained from the list provided by Bolukbasi et al. (2016) [1]. We use the following definition of intrinsic mean (Karcher, 1977; Fréchet, 1948) for points in a Riemannian Manifold.

Definition 1. *Let $(\mathcal{M}, d)$ be a Riemannian Manifold then the intrinsic mean μ for a set of points $x_1, ..., x_n \in \mathcal{M}$ is defined as the minimizer of the squared sum ditances to each point*

$$\mu = argmin_{x \in \mathcal{M}} \sum_{i=1}^{n} d(x, x_i)^2 \tag{1}$$

The Equation 1 is optimized using the Riemannian ADAM optimizer (Bécigneul and Ganea, 2018) provided by the geoopt python library (Kochurov et al., 2020). To speed up the optimization process, we initialize μ with the arithmetic mean of the given points and use a learning rate of 0.0003.

1.2 PGD Details

We use the pre-trained Poincaré GloVe embeddings provided by Tifrea et al. (2018) [2]. For `PGD`, we optimize the proposed objective function using the Riemannain ADAM optimizer (Bécigneul and Ganea, 2018) following the implementations provided by the geoopt (Kochurov et al., 2020) [3] python library. We set the objective function weights equal to each other ($\lambda_1 = 0.5$ and $\lambda_2 = 0.5$). The learning rate of $\alpha = 0.0003$ is used and we run the PGD optimization process for 350 epochs.

[1] https://github.com/tolga-b/debiaswe

[2] https://github.com/alex-tifrea/poincare_glove

[3] https://github.com/geoopt/geoopt/

2 Results

2.1 Poincaré Distance

We use (negative) Poincaré distance to evaluate similarity of vectors in analogy tasks. The Poincaré distance between two points $x, y \in \mathbb{D}_c^n$ is defined using the following expression:

$$d_{\mathbb{D}}(\mathbf{x}, \mathbf{y}) = \frac{2}{\sqrt{c}} \tanh^{-1}\left(\sqrt{c} \| -\mathbf{x} \oplus_c \mathbf{y} \|\right)$$

where, $\oplus_c$ denotes the *Möbius Addition* operation. Without any loss of generality we use $c = 1$.

2.2 WEAT

A description of the different set of words (Chaloner and Maldonado, 2019) used for conducting WEAT is given in Table 1. For `PGD` GloVE and Poincaré Glove, we can use both negative Poincaré distance (-pdistance) and cosine distance as a measure of similarity, Table 2 shows a comparison using both of them. Table 2 further corroborates the debiasing efficacy of `PGD`.

2.3 Analogy tests

The aim of analogy tests is to find a word d, such that d is to a word c, as the word b is to the word a. In Euclidean space, $d = c + (b - a) = b + (c - a)$. However, in the hyperbolic space, because of the non-zero curvature, there are two possible solutions to d, created using gyro-translations as follows (Tifrea et al., 2018):

$$d_1 = c \oplus gyr[c, \ominus a](\ominus a \oplus b)$$

$$d_2 = b \oplus gyr[b, \ominus a](\ominus a \oplus c)$$

Where gyr is the gyro operator (Ungar, 2008) defined as follows:

$$gyr[a, b]c = \ominus(a \oplus b) \oplus \{a \oplus (b \oplus c)\}$$

We combine the two solutions into a single point, $d^t_{d_1 d_2}$, using a hyper-parameter $t \in [0, 1]$ as follows:

$$d^t_{d_1 d_2} = d_1 \oplus ((-d_1 \oplus d_2) \otimes t)$$

	Attribute Words		
		M	brother, father, uncle, grandfather, son, he, his, him
		F	sister, mother, aunt, grandmother, daughter, she, hers, her
Target Words	B_1 : Career vs Family	X	executive, management, professional, corporation, salary, office, business, career
		Y	home, parents, children, family, cousins, marriage, wedding, relatives
	B_2 : Maths vs Arts	X	math, algebra, geometry, calculus, equations, computation, numbers, addition
		Y	poetry, art, Shakespeare, dance, literature, novel, symphony, drama
	B_3 : Science vs Arts	X	science, technology, physics, chemistry, Einstein, NASA, experiment, astronomy
		Y	poetry, art, Shakespeare, dance, literature, novel, symphony, drama

Table 1: List of attribute and target words for each category used for our experiments.

Embedding	B_1**: Career vs Family**		B_2**: Math vs Art**		B_3**: Science vs Art**	
	p	d	p	d	p	d
Euclidean GloVe	0.0773	0.7423	0.4186	0.1716	0.5443	-0.1131
Poincaré GloVe (cosdist)	0.0329	0.9423	0.2422	0.4169	0.1733	0.5267
Poincaré GloVe (-pdistance)	0.0418	0.8929	0.3045	0.2929	0.2138	0.4568
PGD GloVe (cosdist)	0.2456	0.3628	**0.8759**	**-0.6310**	0.5830	-0.1559
PGD GloVe (-pdistance)	**0.2827**	**0.3054**	0.8292	-0.5379	**0.6015**	**-0.1984**

Table 2: Compairison of word embedding on WEAT for different categories - B_i. A higher p and lower d is better. Cos dist represents the use of cosine similarity as the similarity metric while -pdistance represents the (negative) Poincaré distance

Here, t is a hyper-parameter, $\oplus$ is the Möbius addition operator as described in the manuscript while $\otimes$ is the Möbius scalar multiplication defined as follows:

Definition 2. *Möbius Scalar Multiplication of a* $x \in \mathbb{D}^n \setminus \{\mathbf{0}\}$ *by a real number* $r \in \mathbb{R}$ *is defined as*

$$r \otimes x = tanh(rtanh^{-1}(\|x\|))\frac{x}{\|x\|}$$

We follow the same approach as Tifrea et al. (2018) for estimating t, using 2-fold cross-validation and vary t across 11 values in $\{0, 0.1, ..., 1\}$ to find the best value corresponding to highest analogy score.

References

Gary Bécigneul and Octavian-Eugen Ganea. 2018. Riemannian adaptive optimization methods. *arXiv preprint arXiv:1810.00760*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32.

Maurice Fréchet. 1948. Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l'institut Henri Poincaré*, volume 10, pages 215–310.

Hermann Karcher. 1977. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541.

Max Kochurov, Rasul Karimov, and Sergei Kozlukov. 2020. Geoopt: Riemannian optimization in pytorch. *arXiv preprint arXiv:2005.02819*.

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. 2018. Poincare glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*.

Abraham Albert Ungar. 2008. A gyrovector space approach to hyperbolic geometry. *Synthesis Lectures on Mathematics and Statistics*, 1(1):1–194.

2021 8th International Conference on Signal Processing and Integrated Networks (SPIN) | 978-1-6654-3564-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SPIN52536.2021.9566081

Implementation and Comparison of Direct mapped and 4-way Set Associative mapped Cache Controller in VHDL

Dr. Gurjit Kaur
Department of Electronics and Communication Engineering
Delhi Technological University
New Delhi, India
gurjeet_kaur@rediffmail.com

Riyanshi Arora
B.Tech, Electronics, and Communication Engineering
Delhi Technological University
New Delhi, India
arorariyanshi03@gmail.com

Sakshi Shailesh Panchal
B.Tech, Electronics, and Communication Engineering
Delhi Technological University
New Delhi, India
skshpanchal@gmail.com

***Abstract*— This paper presents the implementation of design of a non-pipelined processor that generates memory read/write requests to a direct-mapped and a 4 - way set associative mapped cache controller frequently and examines their performances by calculating Cache hits versus misses. The paper also briefly explains cache memory, the Least Recently Used (LRU) algorithm, cache controller, memory mapping techniques. Following that, a cache controller based on Finite State Machine has been implemented for direct-mapped as well as Four-way set-associative synchronous Cache (buffer) memory of 256 bytes whose size of the block is 128 bits or 16 bytes (by default). A multi-banked interleaved main memory system comprising of 4 memory banks of 1K byte each has been designed and used. The implementation is done using VHDL, and the simulation has been performed using Eda playground, an online simulation tool. The maximum clock frequency is 110 MHz.**

Keywords – Cache Controller, Cache Memory, hit, latency, Mapping, Miss, primary memory, read, write

I. INTRODUCTION

Cache memory is a class of memory which is utilized by the CPU to lessen the time taken to access blocks of information from the primary memory. This type of memory is high-speed. The speed of cache memory is much more than the primary memory. It is used for storing copies of the information present in the primary memory. These copies are made in a way that the most probable portions of the main memory can be accessed at a faster rate [1]. It exists between the processor and the primary memory, facilitating the transfer between these two units at a speed that keeps up with the speed of the processor, as shown in Figure 1. The cache controller acts as a passage for the transfer of data as it steers the instructions given by the processor to main memory or cache memory and vice-versa. Among CPU and primary memory, data is moved as memory words. Among buffer memory and primary memory, data is traded as memory blocks. In the cache memory, the information is stored as blocks, each of them containing a certain number of words. In order to keep a record of which blocks are presently stored in the cache memory and how they are associated with the remaining architecture, the controller keeps a record of identifiers for the blocks which are in Cache presently. Such identifiers are the index, tag, and valid bits, which are related to a complete data block. Through these identifier bits, the cache controller is able to acknowledge the read and write requests made by the processor [2]. This is followed by the execution of these requests to specific blocks or by collecting or writing complete blocks to the bigger yet slower memory, i.e., the main memory.

The benefit of reserving data on Cache, in comparison to RAM, is that it provides much quicker retrieval times, the demerit being its on-chip energy dissipation. This paper reviews the design of an efficient cache controller along with memory.

II. CACHE MEMORY

Cache memory is a more modest, quicker type as compared to other types of memory. The copies of information from repeatedly understandable used main (primary) memory zones are stored in the cache memory. The CPU uses a cache to lessen the access time of retrieving data from the primary memory. It is a very quick memory category that behaves as a buffer between CPU and RAM (Random Access Memory). Moreover, this type of memory also has the highest level of the memory hierarchy. The main memory of a computer system is divided into various levels, of which Level II is the cache memory. The other levels of the main memory are inclusive of Level I, that is, Register, Level III, that is, Main Memory, and Level IV, that is, Secondary Memory. Inside the cache memory, data is saved for the time being, just for faster access [1].

Firstly, a corresponding entry in the Cache is verified when the processing unit has to write or read any data or location in the main memory. Now, if the processing unit has found that location or data in the memory (Cache), it is said that a cache hit has taken place, and finally, the data is retrieved into the primary memory from the buffer memory. Now, if the opposite occurs, i.e., the required data in the Cache is not found by the processing unit, it's said that a cache miss has

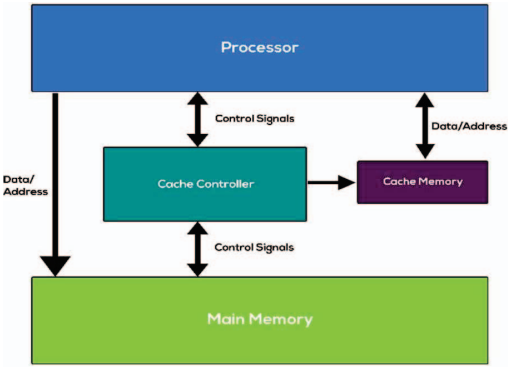


Figure 1. Memory hierarchy

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:34:57 UTC from IEEE Xplore. Restrictions apply.

taken place, and then a novel entry is assigned in the Cache that duplicates required data via primary memory. These operations are of two types: Read and Write. Hence, four types of such operations occur, i.e., 1. Read Hit operation, 2. Read Miss operation, 3. Write Hit operation, and, 4. Write Miss operation [1].

A. Architecture

The single banked cache memory of 256 bytes has been designed. It is inferred using flip-flops. The maximum depth of this cache memory comes out to be 16 entries, and the cache line/block size is sixteen bytes. The processor can read or write only one word at a time. Cache accepts 16 bytes block from the main memory [2]. It consists of two data lines - first from CPU and second from the primary, aka main memory. The line originating from the CPU comprises 8- bit data, whereas the line coming from the primary memory has the size 16 bytes. In order to send the data to the processor, there is an output line of 8 bits. [16]-[18]

B. Read operation

The read operation of the memory sends the required word or data to the address lines, and thereafter, the read control signal is activated. This operation is further of two types: Read Hit and Read Miss. When whatever information we need to be computed is present in the buffer, aka cache memory, it results in a read hit, and if that information which we require for a Read operation is not present, we say that a Read Miss has occurred. If read miss occurs, the information needs to be brought into the Cache by the processor, which causes extra delays [2].

C. Write operation

The write operation sends data and address at which it is to be written to the address lines, and after that, it activates the write control signal. This operation is further of two types: Write Hit and Write miss. When whatever information we need to be computed is present in the cache memory, it results in a Write Hit, and if that information which we require for a write operation is not present, we say that a write miss has occurred. If a write miss occurs, the information needs to be brought into the Cache by the processor, which causes extra delays [2].

III. CACHE CONTROLLER

It is a tool that is utilized to direct operations which concern cache memory, main memory, and the processor. It works as a liaison between these three units of the computer system. Anytime the processor needs to access any data in a particular location in the primary memory; first, it is checked in the buffer memory. This task is performed by the controller, for it checks for that site in the buffer first. In case the requested address is located in the Cache, the data/information is directed to the CPU. If the address is not found, it is the task of the controller to collect it from the primary memory and give it to the processing unit so that it can execute the current instruction. Cache memory is updated with the data before forwarding it to the processor [3]. Figure 3 displays the RTL schematic of a cache controller.

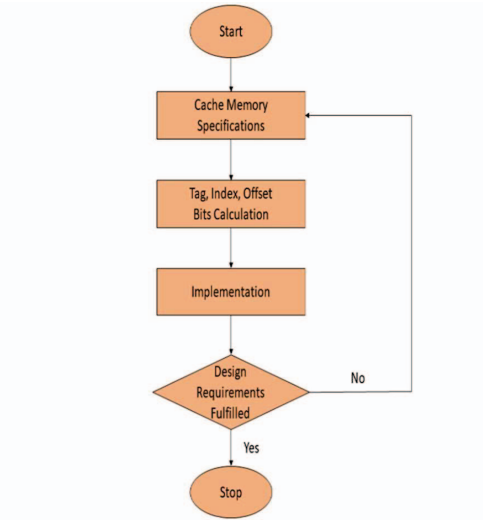


Figure 2. Cache Design

The architecture of the controller consists of the various input lines such as clock, reset, read, write, etc., along with output lines such as refill, update, stall, etc. *A reset* signal is an active low signal to reset all the states and signals involved in the processes. *Flush* is an active high signal which invalidates all the cache lines at once. *Read* and *write* are active high flag signals given by the processor to read from and write to Cache, respectively. *Index* and *tag* inputs are configurable lines that contain the index header and tag header, respectively, of the requested address. The *ready* signal is an active high signal coming from the main memory to acknowledge that the data requested is ready to be fetched. The *stall* is an outgoing signal to stall the processor in the events of read/write misses and write hits [3].

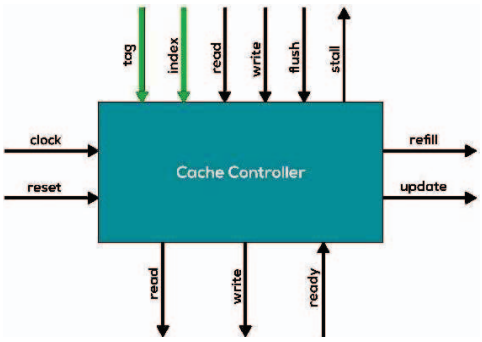


Figure 3. Schematic of a Cache controller

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:34:57 UTC from IEEE Xplore. Restrictions apply.

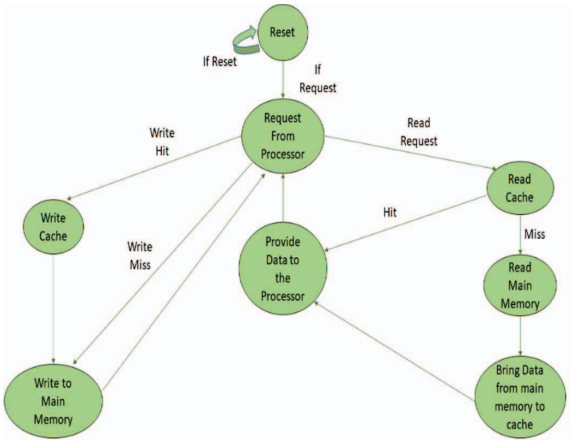


Figure 4. State diagram of the cache controller

The signal *refill* goes to the Cache, which tells it to refill the cache line using the data block read from the primary memory. Similarly, the *update* line tells the buffer memory to update corresponding work on the cache line on a write hit. *Read* (from main memory) is an active high signal given to main memory or primary memory to pick up the requested data block when read miss occurs. *Write* (to main memory) in a similar way, tells the main memory to write the word in it on a write hit/miss [3].

The four primary functions, which are: fetching the location (from the processing unit), reading from the cache memory and primary memory, writing to Cache as well as main memory, then delivering the desired block of data to the processing unit performed by the controller are described pictorially in Fig 4.

IV. MAPPING AND ITS TYPES

A technique or a process that focuses on how the contents or the matter from the main, that is, primary memory, is moved to the cache memory that rests between the CPU (processor) and the primary memory bank, is known as Mapping. In other words, it can also be explained as to how data is duplicated or mapped (from the primary memory to the buffer) when a cache miss occurs. That miss can be any of the two: read miss or write miss. Cache mapping is of three kinds: 1. Direct Mapping technique. 2. Fully Associative Mapping technique. 3. K-Way Set Associative Mapping technique [4].

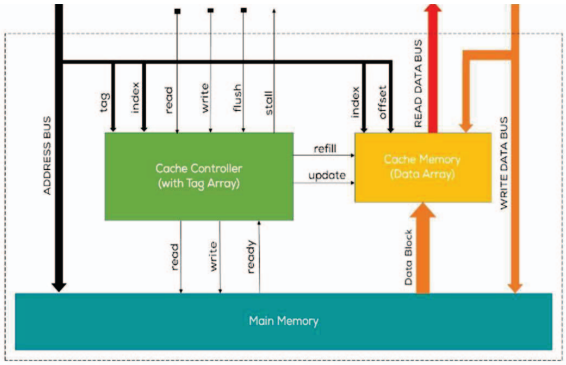


Figure 5. RTL of the implemented top module

To understand these techniques, one should know that the primary memory is split up in uniformly sized divisions called blocks, aka frames. Cache memory also has uniformly sized divisions called lines, and the size of these lines is similar to the blocks of main memory. While cache mapping is performed, the data in the block of the main memory is usually duplicated in the buffer memory rather than fetching the block from the primary memory.

A. Direct Mapping

It is a method of Mapping wherein a block in the primary memory can be copied into only one line that the Cache (buffer) memory contains. This line is a particular line that is calculated by taking the modulo of the address of that particular block which is to be duplicated, and the entire cache lines. Since a particular block can be mapped to a certain cache line, the algorithm of replacement is not required in this case.If the data is there in a particular line in which Mapping is to be done, it automatically gets replaced with the new contents,and the old data would be trashed. This is the most straightforward known technique for mapping [5].

B. Fully Associative Mapping

It is a method of Mapping in which the data from the block could be copied to either line of the Cache, provided it's readily available at that time. If none of the lines are free, then the subsisting data in any one of the lines need to be replaced. Hence, there is a need for an algorithm that can replace that data and make way for new block data to be copied. In such a case, algorithms such as Least Recently Used aka LRU; First Come, First Served aka FCFS, etc., are used. This technique is more flexible than that of direct Mapping.

C. K-way set-associative Mapping

There are some limitations of the direct mapping technique. They are removed in its enhanced version that is known as Set Associative Mapping. It can be called a combination of the two types of mapping techniques described above. In this technique, all the lines of the buffer memory are divided into various sets that depend on K we choose. Now, in this case, any specific block can be duplicated or mapped into a specific set. It cannot go into any other set. But, in that particular set, it can map into any line of that set that is readily accessible. However, if no line is available, then data in any one of those

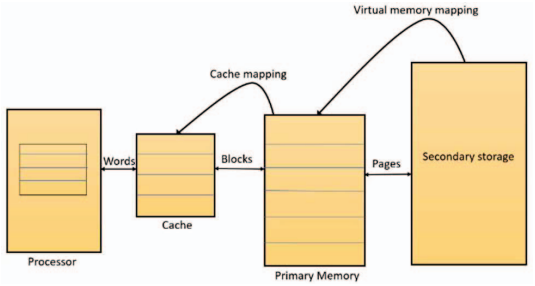


Figure 6. Mapping process

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:34:57 UTC from IEEE Xplore. Restrictions apply.

lines of that set ought to be changed, and therefore, a cache replacement algorithm is required in this case too. From this, it is evident that it is an amalgamation of the previously discussed mapping methods. If K here would be 1, it becomes direct Mapping, and if K were equal to the number of cache lines, it would become fully associative Mapping. Figure 6 describes the process of mapping [6]. This system can be used in various applications [10]-[15].

V. DIRECT MAPPING

The top module of the processor consists of various parts, namely 4 way set associative cache controller, cache memory data array, main memory system consisting of four memory banks. The read/write requests sent by the processor have been incorporated into the test bench. Each of these parts has been coded separately. CPI of the processor is assumed to be 4 plus memory access cycles. Figure 8 lists the test vectors for read/write requests. Simulation results for the same have been shown in Figure 7.

The cache memory data array is a synchronous memory to store cache lines. It is a 256 bytes single-banked Cache that is inferred using flip flops. This data array has a maximum depth of 16 entries and a block/cache line size of 16 bytes which is equivalent to four 32-bit words. It read/writes one word at a time from the processor. As per the R/W protocol, initially, the latency of two clock cycles occurs. In the first cycle, the address of the location is placed in the address register, and rd/wr signals are asserted by the processor simultaneously. In the next cycle, data is made available/written for read/write operation. At the occurrence of the 3rd cycle, the data is read by the processor. The main memory is a synchronous read/write multi-banked interleaved system comprising of 4 banks each of size 1kB. Hence, a total size of 4kB has been allotted to the main memory. Each bank infers a block RAM which accepts 32-bit data at 256 addressable locations. It writes only one word to one location at a time. As per the protocol, read and write operations occupy a total of 4 clock cycles in the primary memory system [4].

A. Read hit

In the event when the desired location is found inside the cache memory, the data is directly given at the output. In

```
test_data(0) <= x"12345678";
test_addr(0) <= x"00000201";

test_data(1) <= x"abcdef01";
test_addr(1) <= x"00000201";

test_addr(2) <= x"00000203";

test_addr(3) <= x"00000201";

test_data(4) <= x"7777ffff";
test_addr(4) <= x"00000202";

test_data(5) <= x"64646464";
test_addr(5) <= x"00000142";

test_data(6) <= x"12121212";
test_addr(6) <= x"00000201";
```

Figure 7. Test vectors for direct-mapped

Figure 7, at *test_address(3)*, the *rd* signal becomes 1, and a *hit* occurs, and the address in *test_address(3),* i.e., *'x201'* as per Figure 8, gets copied into the address register *addr[]*. No stall occurs for a read hit. Net access time is 3 clock cycles.

B. Read Miss

In Figure 8, when the test vector is *test_address(2) = 'x203'*, the signal *rd* becomes 1, and a *miss* occurs as per Figure 7. When this occurs, the data has to be fetched first from the memory banks of the primary memory and brought to the cache memory. It stalls the processor for 4 clock cycles. After that, it will be sent to the processor for further executions. Net access time is 6 clock cycles.

C. Write hit

In the event of a write hit, the address to be written to is found in the Cache, and the data is written at that location. As per Figure 7 and 8, when *test_address(4) = 'x202'* and data to be written to cache is *test_data(4) = 'x7777ffff'*, a write *hit* occurs. It stalls the processor for 2 cycles, and the total access time is 4 clock cycles because of the write-through scheme.

D. Write Miss

When *test_address(0) = 'x201'* and data to be written is *'x12345678'*, a write *miss* has occurred. After that, the data is loaded to the main memory directly. It has a net access time of 4 clock cycles, including 2 stall cycles. This can be verified in Figure 7.

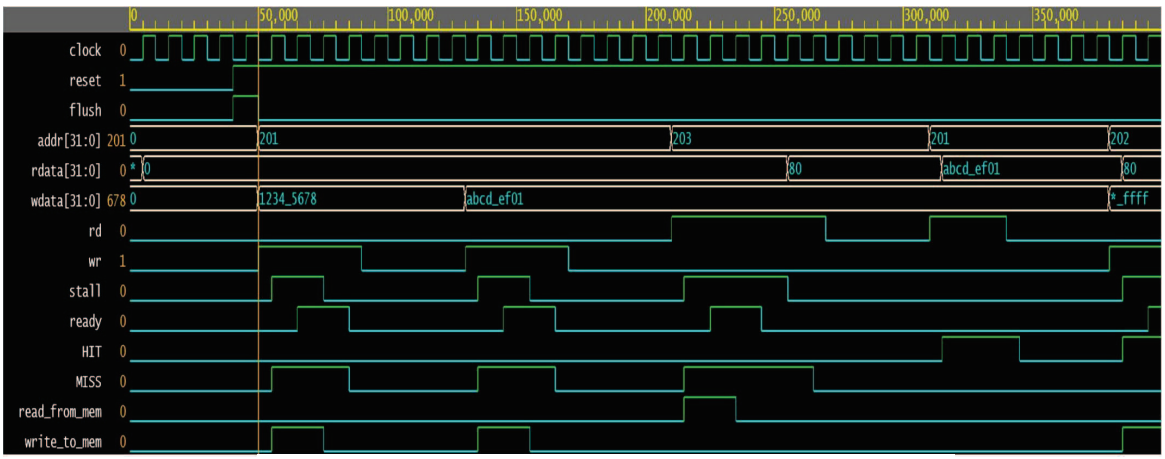


Figure 8. Simulation Results for Direct Mapped Cache Controller

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:34:57 UTC from IEEE Xplore. Restrictions apply.

VI. 4-WAY SET ASSOCIATIVE MAPPING

In order to demonstrate the Four-way set associative mapping technique, the processor is integrated with a Four-way set-associative mapped controller, cache memory data array, main memory system consisting of four memory banks. The read/write requests sent by the processor have been incorporated into the test bench. CPI of the processor is assumed to be 4 plus memory access cycles. Simulation results for the same have been shown in Figure 9.

The specifications of the cache memory data array and main memory system (including the memory banks) have been kept the same as that of the direct-mapped controller. The data array read/writes one word at a time from the processor. The main memory is a synchronous read/write multi-banked interleaved system comprising of 4 banks each of size 1kB. Hence, a total size of 4kB has been allotted to the main memory. It writes only one word to one location at a time. As per the protocol, read and write operations occupy a total of 4 clock cycles in the main memory system [5].

The 4 - Way Set Associative Mapped Cache controller is a single banked write through Cache Controller which addresses 16 entries and has a 256 bytes Associative Mapped Cache. It is segregated into four groups/sets, and each set has 4 lines inside the Cache Memory. A replacement algorithm will be required because it's a 4-Way Set Associative Mapped controller. Tree-PLRU replacement algorithm has been implied. [21]-[23]

A. Tree PLRU algorithm

Tree PLRU is a pseudo least recently used algorithm used for replacement in mapping techniques. In this type of algorithm, a binary tree is used instead of true Least Recently Used (LRU). Each node of the binary tree records which half is older or newer. On this basis, nodes are updates at each reference, and then older pointers are followed to find Least Recently Used victim. Basically, this algorithm is an improved version of the Least Recently Used (LRU) Algorithm [5]. The write-through policy and write buffers have not been incorporated due to complexity reasons. Tag array has been included. [19]-[20]

```
test_addr(0) <= x"00000201";
test_addr(1) <= x"00000203";
test_data(2) <= x"a0a0a0a0";
test_addr(2) <= x"00000201";
test_addr(3) <= x"00000201";
test_addr(4) <= x"00000301";
test_addr(5) <= x"00000001";
test_addr(6) <= x"00000101";
test_addr(7) <= x"00000211";
test_addr(8) <= x"00000221";
test_addr(9) <= x"00000101";
test_addr(10)<= x"00000271";
```

Figure 9. Test vectors for 4-way set associative mapped

The tag memory is implemented using flip-flops. The net access time for read hit is 3 cycles, and that for write hit is 5 cycles (due to the write-through policy). The net access time for read miss is 7 cycles inclusive of 4 stall cycles as a penalty. The write miss has a net access time of 5 cycles.

A. Read hit

When the cache memory has the desired or required location, the data is immediately given at the output. In Figure 9, at *test_address(1)*, *rd* signal becomes 1, and a *hit* occurs, and the address in *test_address(1),* i.e., *'x203'* as per Figure 10, gets copied into the address register *addr[]*. No stall occurs for a read hit.

B. Read Miss

In Figure 10, when the test vector is *test_address(0) = 'x201'*, a *miss* occurs as per Figure 9. When this occurs, the data has to be fetched first from the memory banks of the primary memory and brought to the cache memory. This stalls the processor for 4 clock cycles. After that, it is sent to the processor for further executions.

C. Write Mi ss

In the event of a write hit, the address to be written to is found in the Cache, and the data is written at that location. As per Figure 9 and 10, when *test_address(2) = 'x201'* and data to be written to cache is *test_data(2) = 'xa0a0a0a0'*, a write *hit* occurs. It stalls the processor for 2 cycles.

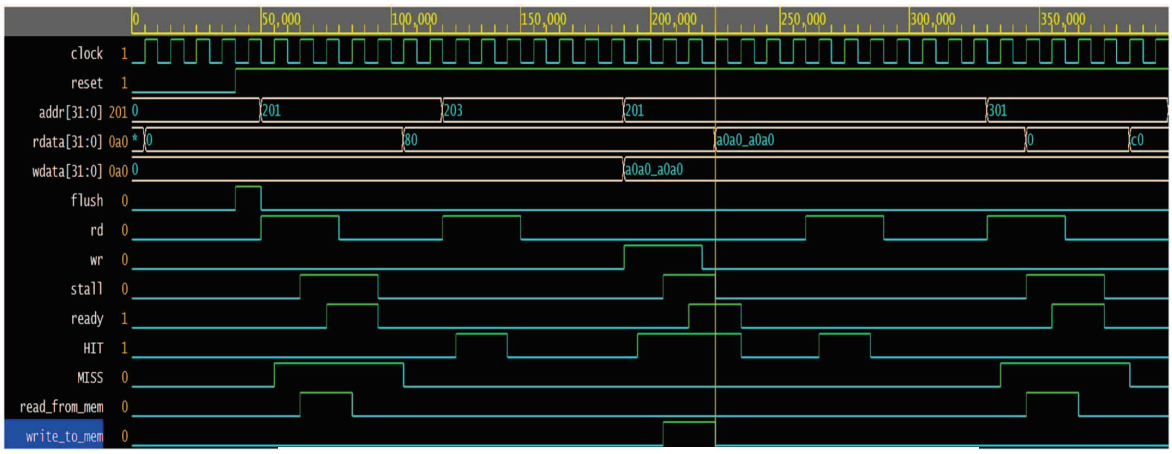


Figure 10. Simulation Results for 4-Way Associative Mapped Cache

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:34:57 UTC from IEEE Xplore. Restrictions apply.

TABLE I STALL AND LATENCY FOR HIT AND MISS OPERATIONS

Action	Direct-mapped		4-way set associative mapped	
	Latency (clock cycles)	Stall (clock cycles)	Latency (clock cycles)	Stall (clock cycles)
Read Hit	2	0	3	0
Read Miss	4	4	7	4
Write Hit	6	2	5	2
Write Miss	4	2	5	2

VII. CONCLUSION

A non-pipelined processor which generates memory read/write requests to a direct-mapped, and a Four-way set-associative mapped controller (Cache) frequently is designed and implemented using EDA Playground (online simulator) in VHDL. These techniques can further be improvised to optimize cache memory for missing data [7] and providing clock synced transient encryption [8]. This technique is used in Managing authorization tokens for third party vendors, integrating third party vendors, synchronization of multi-vendor platform etc. [24]-[30] The codes and the outputs have been verified using some test vectors defined inthe test bench of the cache memory. A memory system has been designed comprising of 4 banks each of size 1 kilobyte from which the inputs are fetched by the controller. The 4- way associative mapped Cache has a non-pipelined processorand uses a Tree-PLRU cache replacement algorithm. The sizeof synchronous cache memory designed is 256 bytes, and the default size of the block is 128 bits (16 bytes). As per the implemented controllers, it has been experimentally found that Set Associative mapping is an improvised version of Direct Mapping and has better efficiency than direct Mapping. In a direct-mapped cache controller, the net access time for read hit is 2 cycles; for read miss, it is 6 cycles, and that for write hit or miss is 4 cycles. In the 4 Way Set Associative Cache controller, the net access time for read, a hit is 3 cycles, for read miss, it is 7 clock cycles, whereas, for write hit/miss, it is 5 clock cycles.

REFERENCES

[1] Watile, Yogesh S. and A. S. Khobragade. "Design of Cache Memory with Cache Controller Using VHDL." *International Journal of Innovative Research in Science, Engineering and Technology* 2 (2013): 2914-2919.

[2] Lalwani, K. and M. Chawla. "FPGA Implementation of High Throughput Cache Controller." (2017).

[3] Hassan S.L.M, Johari M.N, Saparon A, Halim I.S.A, and Ab Rahim A.A, "Multi-sized Output Cache Controllers," in Proc. Int. Conf. Technology, Informatics, Management, Engineering & Environment (TIME-E 2013), Bandung, Indonesia, pp.186-191, June 2013.

[4] Manu Gupta, Jyotsna Singh, 2017, A Comparative Study of Cache Optimization Techniques and Cache Mapping Techniques, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 06, Issue 05 (May 2017), http://dx.doi.org/10.17577/IJERTV6IS050569

[5] Shree, Pavana. "Design and Implementation of Direct Mapped Cache Memory with Same Tag Bit Information." (2015).

[6] Chauan, Praveena & Singh, Gagandeep & Singh, Gurmohan. (2015). Cache Controller for 4-way set-associative Cache Memory. International Journal of Computer Applications. 129. 1-8. 10.5120/ijca2015906787.

[7] P Savla, GS SANDLE, G Vitchev, P Singh, SM Cohen, US Patent 10,909,118

[8] P Kannan, P Singh, US Patent App. 16/530,773.

[9] Singh, P., Elangovan, K., & Kudriavtsev, D. (2020). *U.S. Patent No. 10,616,230*. Washington, DC: U.S. Patent and Trademark Office.

[10] Singh, P. (2021). *U.S. Patent No. 10,971,159*. Washington, DC: U.S. Patent and Trademark Office.

[11] Elangovan, K., Singh, P., & Kudriavtsev, D. (2020). *U.S. Patent No. 10,616,352*. Washington, DC: U.S. Patent and Trademark Office.

[12] Kudriavtsev, D., Nachnani, P., Kashyn, D., Chen, B., Kamuju, S. V., Vaidhyanathan, H., ... & Singh, P. (2019). *U.S. Patent Application No. 16/006,775*.

[13] Kannan, P., & Singh, P. (2021). U.S. Patent Application No. 16/530,773.

[14] Singh, P. (2020). U.S. Patent Application No. 16/058,815.

[15] Yadav, V., Tomar, P., Singh, P., & Kaur, G. (2020). Improvement in XML Keyword Search and Ranking for Data Analytics. In *Smart Systems and IoT: Innovations in Computing* (pp. 339-349). Springer, Singapore.

[16] X. Pan, A. Bacha and R. Teodorescu, "Respin: Rethinking Near-Threshold Multiprocessor Design with Non-volatile Memory," 2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS), 2017, pp. 265-275, doi: 10.1109/IPDPS.2017.109.

[17] M. K. N, Kanagasabapati and S. S. Yellampalli, "Design of Reconfigurable Cache Memory Using Verilog HDL," 2018 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), 2018, pp. 1171-1176, doi: 10.1109/ICEECCOT43722.2018.9001646.

[18] K. Singh and S. Khanna, "Split Memory Based Memory Architecture with Single-ended High Speed Sensing Circuit to Improve Cache Memory Performance," 2020 6th International Conference on Signal Processing and Communication (ICSC), 2020, pp. 188-193, doi: 10.1109/ICSC48311.2020.9182771.

[19] Akanksha Jain; Calvin Lin, Cache Replacement Policies , Morgan & Claypool, 2019.

[20] Georgios Paschos; George Iosifidis; Giuseppe Caire, Cache Optimization Models and Algorithms , now, 2020.

[21] H. Jin, D. Chen, H. Liu, X. Liao, R. Guo and Y. Zhang, "Miss Penalty Aware Cache Replacement for Hybrid Memory Systems," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 39, no. 12, pp. 4669-4682, Dec. 2020, doi: 10.1109/TCAD.2020.2966482.

[22] A. V. Garashchenko, L. G. Gagarina, K. Z. Ye, E. Dorogova and M. Kochneva, "Development of an Approach to Automatic Test Generation Based on the Graph Model of a Cache Hierarchy," 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), 2020, pp. 1940-1944, doi: 10.1109/EIConRus49466.2020.9039334.

[23] J. Tada and R. Higashi, "A Cache Replacement Policy with Considering Fluctuation Patterns of Total Priority Value," 2019 Seventh International Symposium on Computing and Networking Workshops (CANDARW), 2019, pp. 309-312, doi: 10.1109/CANDARW.2019.00061.

[24] Singh, Prabhjot, Kavitha Elangovan, and Dmytro Kudriavtsev. "Managing authorization tokens for calling third-party vendors." U.S. Patent No. 10,616,230. 7 Apr. 2020.

[25] Singh, Prabhjot. "Cross account access for a virtual personal assistant via voice printing." U.S. Patent No. 10,971,159. 6 Apr. 2021.

[26] Elangovan, Kavitha, Prabhjot Singh, and Dmytro Kudriavtsev. "Integrating third-party vendors' APIs." U.S. Patent No. 10,616,352. 7 Apr. 2020

[27] Kudriavtsev, Dmytro, et al. "Multi-vendor synchronization platform supporting multiple formats." U.S. Patent Application No. 16/006,775.

[28] Kannan, Prashanth, and Prabhjot Singh. "Clock-synced transient encryption." U.S. Patent Application No. 16/530,773.

[29] Yadav, Vasudev, et al. "Improvement in XML Keyword Search and Ranking for Data Analytics." Smart Systems and IoT: Innovations in Computing. Springer, Singapore, 2020. 339-349.

[30] Singh, Prabhjot. "Managing access credentials for a service provider." U.S. Patent Application No. 16/058,815.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:34:57 UTC from IEEE Xplore. Restrictions apply.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/355369181

Intelligent Testing in Software Industry

Conference Paper · July 2021

DOI: 10.1109/ICCCNT51525.2021.9580012

CITATIONS
0

READS
45

2 authors:



Anushka Lal
Delhi Technological University
2 PUBLICATIONS **5** CITATIONS
SEE PROFILE



G. Kumar
Delhi Technological University
45 PUBLICATIONS **232** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Circular Economy Challenges View project

System availability modeling and analysis View project

All content following this page was uploaded by Anushka Lal on 18 October 2021.

The user has requested enhancement of the downloaded file.

Intelligent Testing in Software Industry

Anushka Lal
Department of Computer Science and Engineering
Delhi Technological University
New Delhi, India
anushkalal_2k18co083@dtu.ac.in

Girish Kumar
Department of Mechnanical Engineering
Delhi Technological University
New Delhi, India
girishkumar@dce.ac.in

***Abstract*— As the world is progressing towards digitization, software has become a critical component that is affecting the world's prosperity and economic competitiveness. This dependence on software and its corresponding impact in various sectors make it imperative for the need to accept and maintain the high quality of software. Testability is one of those software quality factors that has to be plugged in from the early stage of software development to upgrade the quality of software. Therefore, this study aims to investigate the approach of intelligent testing for improving software testability. Through an extensive literature review, numerous factors are identified that affect testability. Controllability and observability are two important factors. However, software complexity and understandability also play a significant role in this regard. Identifying the impact of testability on deliverability, an emphasis is given to the use of Artificial Intelligence (AI) based automated testing to greatly improve the effectiveness, coverage, and execution time of any software. The role of AI lies in its ability to proactively predict and solve any issues even before they have occurred. An illustration is provided on utilization of some openly available website testing AI tools to test the distinctive features of a university website. These tools have successfully assisted in testing and optimizing the sensitive areas of the website in least time possible. This study can be used by researchers to further explore the scope of intelligent testing and work towards the development of more robust AI testing tools to enhance software testability.**

Keywords—Software testing, Artificial Intelligence, website testing, Selenium IDE, Testim, Pingdom, Nibbler, Wave Accessibility tool

I. INTRODUCTION

Many organizations today, invest huge amounts in purchasing software in order to ease their business and switch to a cost effective system. Their roles expand to several fields like health, education, business, finance and security. In such cases, software failures can cause great inconveniences to users. Sometimes it may be leakage of sensitive information or in case of the software operating in medical sectors, it may even pose a threat to our lives. And this brings into attention the importance of quality assurance to prevent software failures [1].

Software quality is referred as its degree of customer satisfaction and to what extent it has met all the specified requirements [2]. Developers have emphasized on software development life cycle to help deliver quality systems that meet these expectations. They broadly focus on detection and correction of errors from an early stage of software development. This is when software testability comes into the picture. It is important to highlight the difference between testability and testing. While testing deals with assessing the software for its quality and results, testability refers to the degree to which a software can be tested for faults and errors [3]. This is a quality factor in sense that not all software applications are testable. It depends on the application architecture on how easily or how difficult it is to conduct tests and find bugs in it. The testability is defined as "the degree to which a system or component facilitates the establishment of test criteria and the performance of tests to determine whether those criteria have been met" (IEEE standard 610.12-1990). Also as per ISO the testability is expressed as "attributes of software that bear on the effort needed to validate the software product" (ISO standard 9126-1: 2001).

Testing is considered to be the critical stage of any software development process. Before the product is released in the market, it is important to check for any errors or faults that may cause the product to fail. Ideally, the software should be tested for all types of cases and faults i.e. all the functions and test case values should be tested. However with growing complexity of software, it is nearly impossible to achieve this goal either due to time constraints or technological constraints, hence the level of testability for every application differs. Testability is not just about testing but also checking how likely it is that these tests are carried out accurately and can detect faults in the system.

There are two different types of testing practiced in the industry today: manual and automated testing [4]. Manual testing though doesn't require a lot of knowledge about the testing tool, it is not at all feasible in current times. With the ever increasing complexity of software programs, manually testing for all possible test cases is an expensive and time-consuming process. Automating this process with the help of AI will significantly improve the software testability with larger coverage in the most cost-effective and timely manner.

The main objective of this paper is to evaluate software quality on the basis of testability and suggest the ways to enhancing it. The focus of this study is to utilize AI integrated tools for faster and successful execution of software testing while ensuring highest possible software standards. The paper is organized into 6 sections. Section II discusses latest literature in the field of software testability. Section III explains the software quality factors and factors affecting software testability. Section IV directs our discussion towards role of AI based automated testing in improving the efficiency of software testability. We have discussed the features and findings of our experiment on easily accessible website testing AI tools in Section V. Finally, a conclusion and future direction of the work is provided in Section VI.

TABLE I. SOFTWARE QUALITY FACTORS

Offered by	Quality Factors	Description
McCall's Product operation	Correctness	It refers to output specifications of a system. They are considered to be multidimensional for example accuracy of output, availability of information, up-to-datedness of information and documentation and coding standards of the software model.
	Reliability	Reliability deals with how often a system suffers from failure. Here, we determine how many failures of the entire system or the part of the system can be entertained.
	Efficiency	Requirements that address the system function's performance with regard to conformance to other requirements.
	Integrity	Makes sure only authorized personnel can access the system.
	Usability	It talks about how easily users can learn to use a product to achieve their goals and their level of satisfaction in that process.
McCall's Product revision	Maintainability	It defines the efforts needed by software to identify the causes of failures in the modular structure of software, internal documentation or programming manuals and work on improving and repairing them.
	Flexibility	It refers to how easily it can adapt and cater to the changes in future
	Testability	Before the official release of the software, it is important to check whether the system is working as intended and if not what are the detected errors.
McCall's Product transition	Portability	Requirements that suggest how well the software can adapt to other working environments and perform on different hardware and operating platforms.
	Reusability	Requirements that define whether this system can be reused for achieving different project goals in future, in order to save time and resources.
	Interoperability	These requirements define if the system requires creation of interfaces with any other existing system or equipment firmware for operation purpose.
Evans and Marciniak	Verifiability	It defines how easily a system can be checked for bugs and errors.
	Expandability	It refers to future efforts needed to cater larger populations, improve the services, or add new applications and functions to improve usability.
Deutsch and Willis	Safety	It refers to elimination of conditions arising due to errors in process control that can bring down the application by setting alarms or any such warning systems.
	Manageability	It is primarily meant for monitoring of performance or system configuration
	Survivability	It measures the ability of software to continue to exist.

II. LITERATURE SURVEY

Over the years several techniques and metrics for software testability have been comprehensively studied and proposed by researchers and practitioners around the world that helped in improving software testability further. Alanen and Ungar [5] have discussed the software design for testability practices as a way of promoting software testability in an early phase of software development lifecycle. Fault tolerance and fault handling is another topic of interest of their study. They have suggested methods of recovery like checkpoints for restoration of fault encountered systems and fault injection for testing fault tolerance. Voas and Miller [6] in their paper, highlighted the importance of selection of test cases such that they can successfully detect faults in a system. They also highlighted the necessity to identify and isolate potential code that can hide faults during the testing phase. They have also briefly described an effective dynamic technique called software sensitivity analysis for prediction of software testability. Pushpa and Harsha [7] have studied various testability metrics in object oriented systems to help researchers and practitioners decide the better metric for their design. They have also found out that most of these measures are fit for application only in the implementation phase.

Khan et al. [8] observed in their study that poor planning of requirements are common causes of software failure. Improving the quality of a software from an early requirement analysis phase greatly helps improve customer satisfaction and reach the goals in a most cost-effective manner. A review of more than 300 papers about software was carried out in [9] and they identified controllability and observability as the two key factors affecting software testability. They emphasized on improving these factors in order to enhance software testability. Divya and Mishra [10] explored the impact of test automation on the cost, time and quality of a software, with a main focus on automated regression testing. By carrying out experiments

they have concluded that automation is the long term effective solution for a precise and accurate software.

Despite so many contributions in this field, it is however interesting to note that no one method or technique has been accepted as a standard to be followed internationally. Most of these researchers have focused only enhancing factors like controllability and observability for improving testability. Though some have mentioned incorporating automated testing techniques, very few have actually explored the role of AI in this, which is exactly what we have tried to achieve in our study.

III. SOFTWARE QUALITY FACTORS

Many software quality factor models have been suggested over the years, the 11 factor model being the first one. It was suggested by James McCall in 1977. McCall's quality factors are classified under three categories: Product operation, Product revision and Product transition. Evans and Marciniak (1987) later added two new factors: Verifiability and Expandability. Three other important factors namely safety, manageability and survivability were again added by Deutsch and Willis. These are summarized in Table 1

A. Factors affecting Testability

While there are various factors affecting testability of software, there are two factors that have been studied to have maximum impact: Controllability and Observability [9]. Controllability has a direct influence on testability in a sense that higher is the ability to control a software easier it is to provide test inputs hence making tests more automated, specifiable and reproducible. On the other hand observability measures the software's ability to observe its behavior: how change in environment or hardware or software component affects the system. Observability helps test for distinct and predictable outputs. It helps examine the internal state of the system when an error has occurred.

In this study, we have also identified complexity and understandability as important factors that affect software testability. Less complex a software is, there is less probability of it having errors and thus, easier to test it. Complexity can be categorized under functional complexity, structural complexity and code complexity. Functional complexity is the complexity existing in the methods and model of the program. Structural complexity deals with how manageable the software is in terms of decomposition of modules and smaller units. Code complexity is about how easily the code is understood based on the adopted coding standards. The better testers understand the software, better they understand the software system. This factor is also not only related with improving the software design but also helps build the communication gap between all levels from designers to developers to testers.

IV. AI BASED AUTOMATED TESTING OF SOFTWARE

Automation testing as the word suggests performs tasks without any user intervention [11]. Automation testing has been adopted over manual testing for its faster performance and higher reliability [12]. Automation tools are utilized for automation of certain tasks but not all of them [13]. They are mostly suitable in carrying out repetitive tasks or tasks that are difficult to carry out manually, giving greater accuracy in least possible time. However, one of the requirements of automated testing requires testers with enough knowledge about the software and testing tools [14].

AI powered automation testing is what we are talking about here. The advancements in algorithms and its ability to collect massive data will extend the scope of automated testing of software. They help higher up the scales of testing. Giving importance to early testing of software in different phases of Software Development Life Cycle, every new modification requires to be tested on all hardware configurations and supported platforms. Manually running them each time is a very time consuming and expensive process. AI software testing helps overcome this difficulty. They even help run lengthy tests that may have been avoided under manual conditions. Without worrying about the ever growing complexity of software, these automated tools can look into all the minor details of the system executing multiple tests at the same time. They greatly enhance the depth of testing by investigating the internal workings of the program like memory states, data and other content in order to decide if the program is running as expected [15].

Visual testing is one other important feature of AI testing that can be beneficial to this industry. AI technology supports image and pattern recognition that make it possible to perform visual testing of all elements at pixel level. Automated tests perform the same steps several times in a very precise manner to help improve the accuracy. They can also help simulate numerous virtual users to perform tests in a controlled environment. A slight change in the code automatically triggers a test to be performed. This allows developers to focus on improving the features and quality of the software without worrying about its complexity and how long it will take to test its successful execution. Automation testing is largely known to increase the testing quality [16]. AI tools can not only rapidly carry out the testing, but also allow for larger test case coverage. Multiple testing tools can be used all at once allowing parallel testing in distinct test scenarios. They reduce the cost of failure as they can detect and fix errors at an early stage in the software development phase, thus reducing chances of delivering a faulty software product.

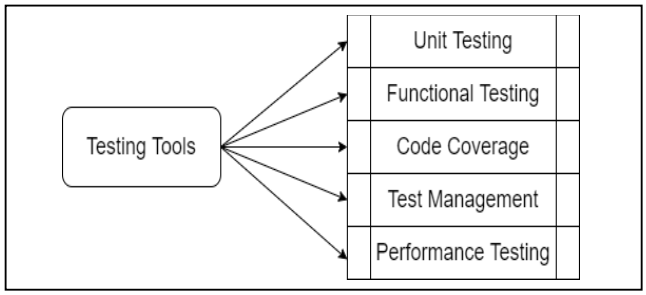
Despite these advantages there are some concerns regarding AI based automated testing that is preventing several developers from adopting it. Some have a very hard time deciding on the right tool for evaluation, some fear not being able to afford the tools along with its maintenance costs while others lack proficiency in writing down the automation test scripts [4].

V. APPLICATION OF AUTOMATED WEBSITE TESTING TOOLS

Automation testing is categorized into five separate categories in the software industry (Refer Fig. 1). Each of these tools aid to test different aspects of a software. Unit testing is the most common one that tests for the code structure and working of specific individual units in a software program. Functional testing tools help examine the obtained output from given inputs with the expected output for any function being tested. They help evaluate compliance of the system with

specified requirements [15]. Code coverage tools help identify the sections of the code that have been tested using automated tools, leaving behind those sections that are still vulnerable to faults and errors. Test management tools help automate the process of test planning, strategy designing and preparing test reports for efficient testing. Performance testing helps measure and investigate the responsiveness, user interface, graphics and workload balance of software applications.

Fig 1: Classification of Automated testing tools



A combination of all these tools has been considered for this work. This study aims to test a university website using various AI powered automated website testing tools and compare their features. The features and limitations of the tools used in this experiment have been summarized in Table II.

Starting with the Selenium IDE [17], its extension for Chrome was used for testing our website. The test cases were recorded as we interacted with different features of the website. All three kinds of Selenese command: actions, assertions and accessors were tested. The click and select options on the website were tested as part of actions command. For accessors we verified the title of the page and validated a text to check if it is identified as text or not. Some conditional logics were also tested for different inputs in a search box to check for its outcome. The test cases for basic login flow and navigation ran successfully, but there was a failed case for the click option on the logo of the university. However, this command was easily skipped to continue executing the remaining commands. The best part of this tool was the ease with which commands could modify the tests by inserting, altering, and deleting commands. Its other advantage is that it displays the documentation of the selected command making it quite user friendly and easy to use. Breakpoints were also easy to set for debugging purpose. Selenium IDE can be used by tester of any level for testing their web applications. This tool supports several kinds of testing types that involve acceptance testing, functional testing, performance testing, regression testing, test driven development (TDD) and behavior-driven development (BDD).

Testim.io [18] was next automated tool used in our experiment that required no deep level knowledge and understanding of programming language for execution. Testim.io is commonly referred to as an automated testing platform rather than a specific tool as it integrates several other features like running test locally or on cloud, capturing logs, screenshots and producing detailed reports of test runs through graphs, statistics and other information. Although it does have a feature to insert our own custom JavaScript code for testing, the predefined steps was used for test cases in our experiment. Having it installed as an extension on the Chrome browser, we started a recorded test by checking with the basic login flow to a student account. Next, amongst the various available options for validations and assertions, specific element visibility and text visibility was tested on our test site. Besides these, we played around with other options like creating breakpoints, adding hover and navigation actions, organizing test suites and running specific tests only. After completion of each step, a feedback was received with a green or red icon signaling a successful test and a failed test respectively. Testim.io is an AI tool best suited for testing small projects like landing pages or any simple applications. It has a very user friendly interface with simple settings and can be used by testers and developers of any level.

Since, our experiment only involved the testing of web applications, we also experimented with tools designed to individually test for specific aspects of a website like its speed, performance, coding standards and user interface specific elements of the website. These tools evaluated the website features with numeric values.

Nibbler [19] was used to test the overall performance of the website in range of 1-10 based on its evaluation of features like accessibility, experience, marketing and technology. All these measures help us identify how accessible the website is to different kinds of users, how satisfying it is, how well it is marketed and how well are the technical details designed. The test site received a score in range of 5-6 under all these categories, resulting in an overall performance score of 6 out of 10. Besides these scores, Nibbler also suggested steps to improve this score. For the test site, it highlighted the importance of defined headings for search engine optimization and suggested use of standard heading conventions like H1 tag for top level headings, H2 for sub headings, H3 for further sub headings and so forth.

Pingdom [20] specifically tested the loading speed of the website. It gave the test site a score of 77 with a grade C, signaling very low grade performance. The load time speed was 2.42 seconds which is greater than ideal speed of 1-2 seconds only. To help improve this performance, it requested the webpage to use cookie-free domains, expires headers, avoid the use of empty hyperlinks and compress website components with tools like ‘gzip’. The cookie free domains avoid unnecessary network traffic caused by cookies. Expires headers are responsible for caching website components to avoid repetitive and unnecessary HTTP requests on subsequent page visits. They also displayed requests individually made by images, scripts, HTML and CSS contents and domains linked to the website to specifically identify components making unnecessary requests and affecting the loading time of the website.

Wave accessibility tool [21] was used to test a completely different domain of website relating to its accessibility and usability by differently abled users. This tool highlighted large percentage of color contrast errors between texts and background color of the test site and suggested using higher contrast colors for better visibility.

TABLE II. COMPARISON OF EXPERIMENTED WEBSITE TESTING TOOLS

Tools	Features	Limitations
Selenium IDE [17]	• Record and run test cases automatically for web applications • Operates on the use of set of Selenese commands, which define the actions and operations to be performed • Allows easy debugging options by letting the test developers set breakpoints • Every element makes use of multiple-locators so the IDE ensures successful execution of test case • Test cases can be exported to various programming languages like Java, Ruby, C#, PHP, JavaScript, and Python • Execute parallel testing on different browsers and on Selenium WebDriver servers using the command-line test runner called the SIDE runner in Selenium IDE	• Database connections cannot be tested • Dynamic features of web-based applications cannot be handled • Does not support screenshot feature to capture test failures • Does not generate any overall result report
Testim.io [18]	• Supports both coded and codeless test cases • Preexisting cases for conditional logics, different types of validation, hover and navigation actions • Runs on both local and cloud platform • Option to organize, reuse and schedule test runs • The AI and parallel running capabilities help run several tests in very less time. • Its quick troubleshooting feature helps identify errors of failed test during execution • Runs bug trackers to report any existing bugs • Testim integrations are offered by vendors like Slack, Github, TestRail, JIRA Software and Trello.	• cross browser is not supported • image verification can be difficult at times
Nibbler [19]	• Provides a score out of 10 for different key features of a website like its accessibility, coding compliance with standard rules, social media handlers along with suggestions on how to improve or optimize that feature	• Does not test pages that include the anti-bot technologies to detect and protect from spammers • only tests around 5-10 pages
Pingdom [20]	• Allows testing of the performance from seven different locations worldwide • Evaluates the site's load time, page size and total number of HTTP requests it makes • Also provides a list of suggestions to further optimize the site's performance and speed	• Limited number of features available for free accounts
Wave Accessibility Tool [21]	• Investigates font size, colors, texts, use of imagery, alt tags and navigation around the site • Makes sure the website isn't discriminating against any differently abled users	• Only focused towards graphical user interface and does not test the actual performance of the website

However, based on all the tools used for this experiment, Testim.io proved to be the most efficient testing tool. Not only does it avoid the need of knowing any particular programming language, but also provides multiple other functionalities like maintenance, troubleshooting, reporting, reusability, and incorporating our own functionalities using JavaScript and HTML. It helps automate the tests in a simple, effective, stable, and collaborative manner.

Overall, most of these tools were able to generate a detailed report by scanning through all the sensitive areas of the website in just a few minutes which may have taken much longer in a manual process. Not only have these tools identified the weak points and errors in the website, but have also provided suggestions on how to further optimize the performance and enhance the features of the website. All of them had their distinct features. Selenium IDE and Testim.io allowed us to create our own test cases by simply interacting with various features of the website. While Nibbler, Pingdom, and Wave accessibility tools are domain specific and automatically scanned and generated a report of the test site.

VI. CONCLUSION

AI is today's key to help streamline software testing to make it more efficient and reliable. Software components need to be more independent, controllable, understandable and observable so that the ease of their testing is greatly increased and hence, the degree of testability increases. AI helps improve the quality of testing in a controllable manner and produce much more accurate results. Our experiment results clearly demonstrate how AI tools have assisted with testing and optimizing the sensitive testing areas in fraction of seconds which may not have been possible to test manually. AI powered automated testing tools should be widely used for different kinds of applications as they examine the correlations from previous results and recommend necessary steps based on risk to help enhance the application test coverage and resolve the defects. This study however, faces a limitation that it only experiments with the available web application testing tools. In future we aim to experiment with AI automated testing tools covering wide range of technical domains such as app development and API testing. Furthermore, researchers are recommended to research the scope of AI testing tools in other fields of development to help automate and reduce the amount of mundane and tedious tasks in development and testing.

REFERENCES

[1] "The Biggest Software Failures in Recent Years - DZone Agile." https://dzone.com/articles/the-biggest-software-failures-in-recent-years (accessed Sep. 20, 2020).J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] "Quality Assurance, Quality Control and Testing | AltexSoft." https://www.altexsoft.com/whitepapers/quality-assurance-quality-control-and-testing-the-basics-of-software-quality-management/ (accessed Dec. 29, 2020).

[3] M. M. Hassan, W. Afzal, M. Blom, B. Lindstrom, S. F. Andler, and S. Eldh, "Testability and Software Robustness: A Systematic Literature Review," *Proc. - 41st Euromicro Conf. Softw. Eng. Adv. Appl. SEAA 2015*, no. September, pp. 341–348, 2015, doi: 10.1109/SEAA.2015.47.

[4] M. A. Umar and C. Zhanfang, "A Study of Automated Software Testing : Automation Tools and Frameworks," *Int. J. Comput. Sci. Eng.*, vol. 8, no. 06, pp. 217–225, 2019.

[5] Alanen and L. Y. Ungar, "Comparing software design for testability to hardware DFT and BIST," in Proceedings of the 2011 IEEE International conference on automated test (AUTOTESTCON'11), 2011.

[6] J. M. Voas and K. W. Miller, "Software testability: The new verification," IEEE Software, vol. 12, no. 3, pp. 17–28, 1995..

[7] H. Singhani and P. R. Suri, "Testability Assessment of Object-Oriented Software Using Internal & External Factor Model and Analytic Hierarchy Process," *Int. J. Sci. Eng. Res.*, vol. 6, no. 6, pp. 268–277, 2015.

[8] Khan, R., Srivastava, A. K., Pandey, D., "Agile approach for Software Testing process", In System Modeling & Advancement in Research Trends (SMART), International Conference (pp. 3-6). IEEE, November, (2016)

[9] V. Garousi, M. Felderer, and F. N. Kiliçaslan, "A survey on software testability," *arXiv*, pp. 1–47, 2018.

[10] D. Kumar and K. K. Mishra, "The Impacts of Test Automation on Software's Cost, Quality and Time to Market," *Procedia Comput. Sci.*, vol. 79, pp. 8–15, 2016, doi: 10.1016/j.procs.2016.03.003.

[11] Andreas Leitner, Ilinca Ciupa, Bertrand Meyer, Mark Howard "Reconciling Manual and Automated Testing: the AutoTest Experience". ETH Zurich CH-8092 Zürich, AXA Rosenberg Investment Management LLC Orinda, California 94563.

[12] Jagdish Singh, Monika Sharma." A Comprehensive Review of Web-based Automation Testing Tools". International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3(10), October 2015

[13] 'Automated Software Testing - International Software Test Institute'. [Online]. Available:https://www.testinstitute.org/Automated_Software_Testing.php.

[14] C. Padmini, 'Beginners Guide To Software Testing', pp. 1–41, 2013

[15] "Why Automation Testing is Necessary." https://www.cprime.com/resources/blog/why-automation-testing-necessary/

[16] M. Polo, P. Reales, M. Piattini, and C. Ebert, 'Test automation', IEEE Softw., vol. 30, no. 1, pp. 84–89, 2013

[17] "Selenium IDE · Open source record and playback test automation for the web." https://www.selenium.dev/selenium-ide/ (accessed May 26, 2021).

[18] "Automated Functional Testing - Software Testing Tool - Testim.io." https://www.testim.io/ (accessed May 26, 2021).

[19] "Nibbler - Test any website." https://nibbler.silktide.com/

[20] "Website Performance and Availability Monitoring | Pingdom." https://www.pingdom.com/

[21] "WAVE Web Accessibility Evaluation Tool." https://wave.webaim.org/

[22] "5 Popular AI-powered tools for test automation." https://www.nextgenerationautomation.com/post/ai-powered-tools (accessed May 26, 2021).

View publication stats

2021 8th International Conference on Signal Processing and Integrated Networks (SPIN) | 978-1-6654-3564-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SPIN52536.2021.9565978

Interfacing of Digital and Resolver Type Systems for Transmission of Heading Angle

Vijendra Kumar Singh
Department of Electronics and Communication
Delhi Technological University, Delhi, India
Kvijendra.66@gmail.com

Dr. Yashna Sharma
Department of Electronics and Communication
Delhi Technological University, Delhi, India
yashnasharma@dtu.ac.in

***Abstract*—A marine vessel platform is fitted with sensors and equipments of diverse origin which are required to work in tandem in order to enhance its operation capability. However, due to incompatibilities in the data formats and protocols that are used, these equipments/ sensors cannot be integrated directly. As the systems available are a mixture of digital and analog types, this paper details the design and development of Embedded Interfacing Solution for interfacing of Digital and Resolver type systems for transmission of Heading Angle. An embedded interfacing board was first designed, using Cadence OrCAD proprietary software, and subsequently implemented using microcontroller 8051, Digital to Resolver converter modules, low power operational amplifiers and other required components. This electronic embedded interfacing board converts the digital input containing heading angle into Analog Resolver type of output. This was tested using simulated heading angle created with the help of COM Port Toolkit software application and the output of this embedded interfacing solution was measured on Angle Position Indicator (API) and with the help of Oscilloscope.**

Keywords—Serial asynchronous, UART, Resolver, RS422A, Heading Angle, Gyroscope

I. INTRODUCTION

It is well known that the old marine vessels are installed with various types of analog equipment, systems and sensors. As an example, the mechanical gyroscope fitted onboard marine vessels gives Heading Angle (HA) in Analog Resolver format to the various sink equipment, spread across the platform of the marine vessel. However, as a part of renovation and modernization, the old mechanical gyroscopes were replaced by the latest Digital Gyroscope based on Ring Lasers [1]. Therefore, the major challenge lies in the interfacing of these two type of equipments as the latest Digital Gyroscope is not compatible with the old mechanical Gyroscope. The format in which the various sink equipment desire the Heading Angle is not compatible with the format supported by the Digital Gyroscope. Therefore, an embedded interfacing solution has been designed and developed to interface the latest digital gyroscope with analog-resolver-type sink equipments [2]. It is to be noted that throughout this manuscript, the marine vessel's heading is described as the actual direction of marine vessel's central axis with respect to the true north in the clockwise direction.

II. BLOCK DIAGRAM OF PROPOSED EMBEDDED INTERFACING SOLUTION

The interfacing arrangement for the proposed embedded interfacing solution is shown in Fig.1. This shows the top level functionality of the embedded interfacing solution and the modules used in design [3]. The driver IC AM26LS32 [4] is used to convert the RS422A standard input into TTL format [5]. The microcontroller 8051 [6] is used to extract the Heading Angle information from the Serial Asynchronous input packet in TTL [7] and feed the Heading Angle to DSC11524 Modules [8]. These modules convert the Heading Angle information into Analog Resolver format and further this output is fed to the low power operational amplifiers.

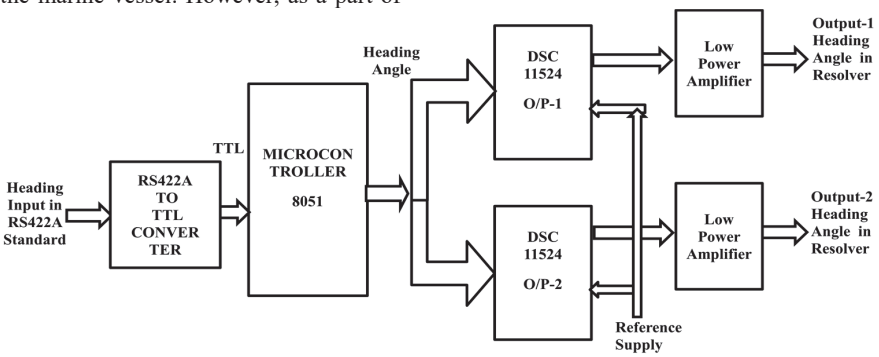


Fig.1. Block Diagram of Proposed Embedded Interfacing Solution.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:38:57 UTC from IEEE Xplore. Restrictions apply.

III. THE KEY FACTORS FOR THE DESIGN

The information of the heading angle was converted to the Analog Resolver types of format [9-10] by a precisely designed embedded interface solution. The key factors which led to this particular design [2] are mentioned below:-

i. The design to have *scalability* to cater for change in future requirements.
ii. The electronic embedded interface solution should not add any *latency* in the system.
iii. The electronic embedded interface solution should be *portable* and easy to maintain onboard marine vessels.
iv. The software algorithm should be *accurate* and should not introduce any error in the heading angle information during conversion.

IV. SCHEMATIC

The schematic diagram shown in Fig. 2 was designed using Cadence OrCAD Capture proprietary software tool suite [11]. OrCAD Capture is a powerful schematic drawing tool for implementing the design intent of electronic circuits. The major components used in designing of schematic diagram are the Microcontroller 8051 IC AT89C55, Digital to Resolver Converter Module DSC11524-303 [8], One IC for TTL to RS422A standard [12] conversion AM26LS32, four number of Low Power Operational amplifiers LM324 and a few other driver ICs.

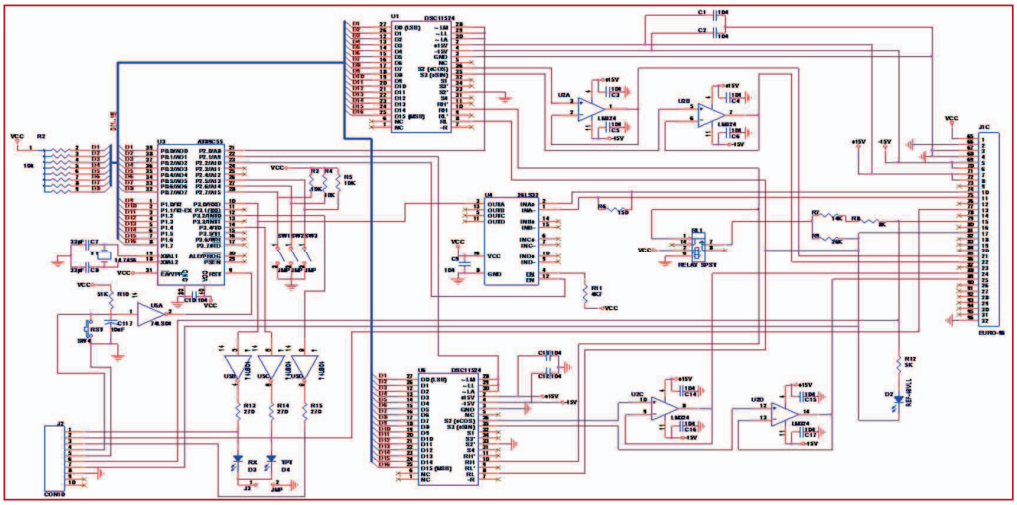


Fig.2. Schematic of Proposed Embedded Interfacing Solution.

V. PRINTED CIRCUIT BOARD (PCB)

The PCB of embedded interfacing solution is shown in Fig.3. This PCB was designed by using the OrCAD PCB Editor Cadence software [13]. Two layer PCB was used for designing the needed PCB. The basic steps utilized in PCB designing are: drawing the outline of the PCB, assigning the foot prints of every component and module, creating a netlist, automatic or manual routing, selection of top or bottom layer, selection of constraints, generating the artwork (Gerber file). A Gerber file keeps format of ASCII vector files that gives information of every physical board layer of PCB designing. Schematic board objects e.g. solder mask, vias, copper traces, silkscreen images and pads are shown by a flash or draw code, and defined by the series of coordinates of vector. These files are utilized by the manufacturer to generate the details of design in the physical form of the Printed Circuit Board.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:38:57 UTC from IEEE Xplore. Restrictions apply.

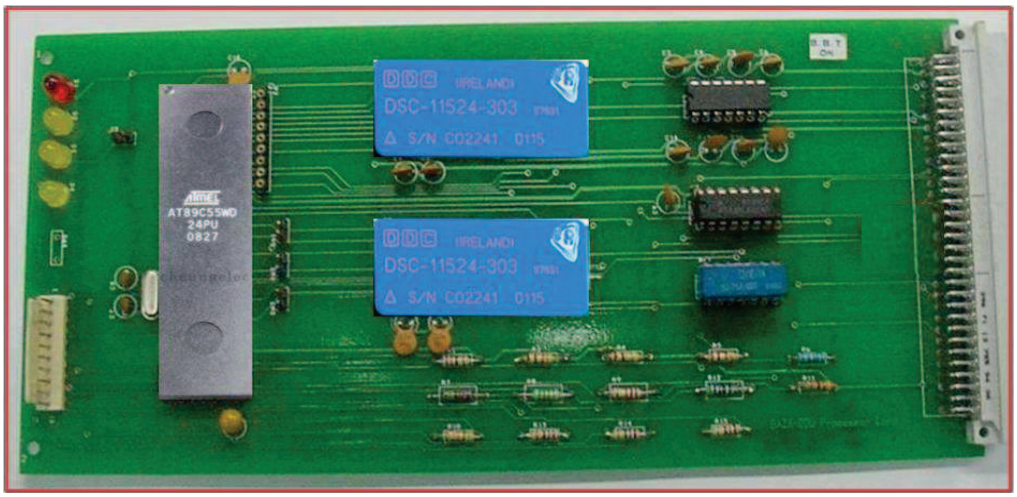


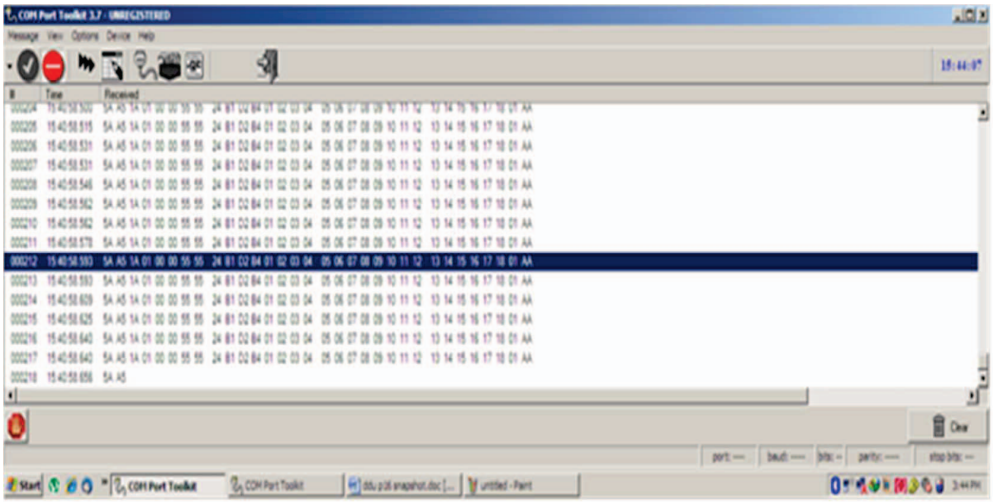
Fig.3. PCB of Proposed Embedded Interfacing Solution.

VI. ALGORITHM

Programming for 8051 microcontroller is written in C-language on Keil IDE (Integrated Development Environment). The microcontroller receives a serial packet in TTL format on its UART (Universal asynchronous receiver-transmitter) and on pin number 10. It extracts the heading angle from serial packet and keeps these values on port P0 of microcontroller. The digital to resolver converter module [8] is connected on port P0 and P1, which converts this heading angle in TTL into Resolver format and sends further to low power operational amplifier.

VII. HEADING ANGLE SIMULATION

The embedded interfacing solution has been tested with the help of simulated heading angle created by COM Port Toolkit [6-19]. This COM Port Toolkit is PC based application that mimics the output behavior of the latest digital Gyroscope (Ring Laser Types) [13]. This application software provides a user friendly interface to configure latest Gyroscope data. Whenever the simulation is needed on the latest Gyroscope data, the packet structure of latest Gyroscope is written in this application and the required serial parameters are set. The data is then sent forward, and the serial asynchronous packet with initially set serial parameters, can be received at any USB output port of PC. A commercial USB to RS422A converter has been utilized in this simulation [14] because the embedded card needs the input data in compliance with the RS422A standard. This RS422A standard data is fed into the input of the card and the output of the card is checked. Subsequently, the performance of the interface card was evaluated. A snapshot of the COM Port Toolkit with simulated values is shown in Fig.4. This kind of validation philosophy helps to capture error/ discrepancies, if present, at an early stage of the design and development of embedded interfacing solution, and thus saves a lot of time during the actual interfacing stage [13].


Fig.4. Screenshot of COM Port Toolkit with simulated Heading Angle

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:38:57 UTC from IEEE Xplore. Restrictions apply.

VIII. TESTING SETUP AND TESTING

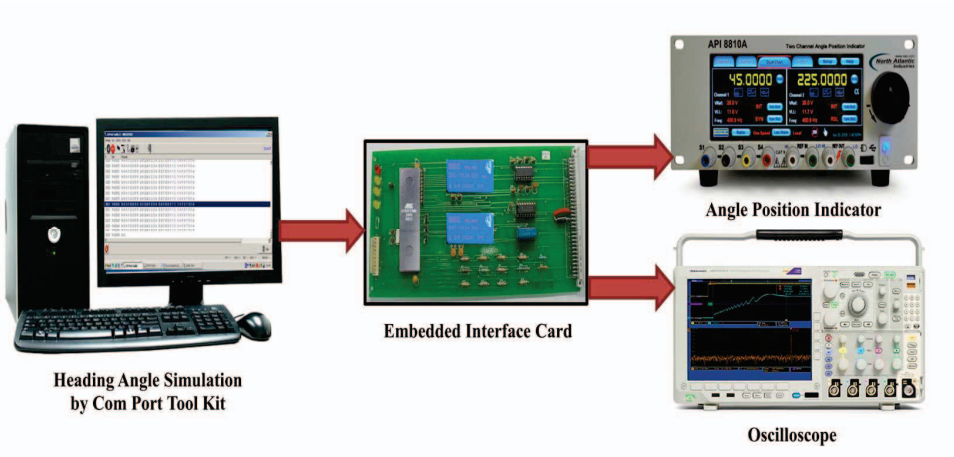


Fig.5. Testing Setup for Embedded Interfacing Solution.

The testing setup for embedded interfacing solution is mentioned at Fig.5. The simulated values of heading angle have been generated with the help of PC based COM Port Toolkit application [14] and fed to input of embedded card with the help of an USB to RS422A standard serial converter. The outputs of embedded card have been measured on Angle Position Indicator instrument.

IX. RESULTS

The testing setup was arranged as mentioned in Fig.5. The application software, COM Port Toolkit application was installed on Window operating system based PC. In this application, the value of heading angle was fed in required packet format with initially set serial communication parameters i.e. start bit, stop bit baud rate and refresh rate etc. The Range of Heading angle keeps within 0 to 360º. The simulated heading angle in serial asynchronous format was taken from USB port of PC and an USB to RS422A standard converter was utilized for conversion. This serial packet in RS422A standard was fed to input of embedded board and this board converts the heading angle in serial asynchronous to Resolver format. Analog Resolver has sinusoidal signal in four windings S1, S2, S3 and S4. This Resolver output of card fed to Angle Position Indicator Instrument (API). The API displays the heading angle values on its display. These heading values were matched with input heading angle values and found correct with minute deviation, reading are mentioned for different heading angle values in Table-2. The Line to line voltage (VLL) for different heading angles was measured with the help of oscilloscope and mentioned in Table 1.

A. Line to line Voltage (VLL) measurement for different Heading Angle with the help of Oscilloscope at the output of embedded interfacing board

Sink equipments have a motor which needs four winding Resolver input signal and when it gets the Resolver input signal then it starts moving. The amount of change in the angle depends upon the values of the RMS Line to Line Voltage (VLL). An oscilloscope is used for the measurement of output waveform and line to line voltages (VLL). Three channels are used for measurement of outputs with respect to the reference supply. The VLL voltages are mentioned in Table 1.

Table 1. VLL voltage w.r.t. different Heading Angles measured on Oscilloscope

S. No.	Simulated Input Heading Angle	Voltage (VLL)	
		S2,S4 Cos Winding	S1,S3 Sin Winding
i.	0	12.12	0.002
ii.	30	10.51	6.005
iii.	45	8.6	8.52
iv.	60	6.044	10.44
v.	120	6.040	10.43
vi.	90	12.15	0.002
vii.	180	12.15	0.002
viii.	270	0.003	12.06
ix.	360	12.12	0.002

B. Waveforms captured for different Heading Angle with the help of Oscilloscope at the output of embedded interfacing board

Three channels of oscilloscope were used as follows:-

i. Channel 1 (Yellow Colour): Used for Reference signal measurement

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:38:57 UTC from IEEE Xplore. Restrictions apply.

ii. Channel 2 (Blue Colour): Used for measurement of VLL across Sine windings (S1,S3)

iii. Channel 3 (Pink Colour): Used for measurement of VLL across Cosine windings (S2,S4) measurement

VLL Voltage Calculation formulas:-

i. VLL(S1,S3)=Vmax(VLL)* Sin(Heading angle)

ii. VLL(S2,S4)=Vmax (VLL)*Cos(Heading angle)

a) Case1

In this case, we feed Heading angle=60º with the help of simulator and measured the VLL voltage across S2, S4 Cosine and S1, S3 Sine Windings as shown in Fig.6(a).

b) Case2

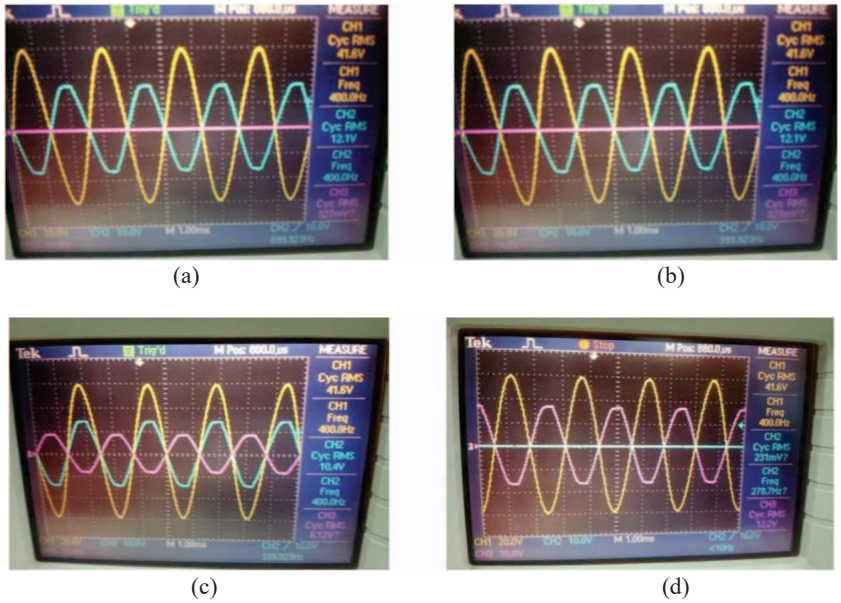
In this case, we feed Heading angle=90º with the help of simulator and measured the VLL voltage across S2, S4 Cosine and S1, S3 Sine Windings. Here no Cosine winding is displaying, because of Cos 90º=0. Therefore, the voltage across cosine winding is zero as shown in Fig.6(b).

c) Case3

In this case, we feed Heading angle=120º with the help of simulator and measured the VLL voltage across S2, S4 Cosine and S1, S3 Sine Windings as shown in Fig.6(c).

d) Case4

In this case, we feed Heading angle=180º with the help of simulator and measured the VLL voltage across S2, S4 Cosine and S1, S3 Sine Windings. Here no Cosine winding is displaying, because of Cos 90º=0. Here, the Sine winding is not displaying, because Sin 180º=0. Therefore, the voltage across sine winding is zero as shown in Fig.6(d).



(a) (b)

(c) (d)

Fig.6. (a) Waveform for 60º Heading Angle (b) Waveform at 90º Heading Angle (c) Waveform at 120º Heading Angle (d) Waveform at 180º Heading Angle

C. Heading Angle Measured on API at the Output of Embedded Interfacing Solution

Values of different-different heading angles measured at the output of Interface board with the help of API are mentioned in Table 2.

Table 2. Heading Angle measured on Angle Position Indicator

S. No.	Simulated Input	Output Heading
	Heading Angle	Angle on API
i.	0.0	0.0
ii.	1.0	0.94

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:38:57 UTC from IEEE Xplore. Restrictions apply.

iii.	2.0	1.89
iv.	3.9	3.71
v.	5.0	4.9
vi.	15	14.9
vii.	30	29.9
viii.	45	44.9
ix.	60	59.5
x.	75	74.9
xi.	90	89.8
xii.	105	104.9
xiii.	120	119.88
xiv.	135	134.88
xv.	150	149.88
xvi.	165	164.89
xvii.	180	180
xviii.	185	185.03
xix.	175	174.94
xx.	210	210.3
xxi.	225	225.3
xxii.	270	270.7
xxiii.	285	285.9
xxiv.	300	302.3
xxv.	315	316.8
xxvi.	330	331.3
xxvii.	345	345.8
xviii.	345.1	345.7
xxix.	344.9	345.59

X. CONCLUSION

An embedded interfacing solution has been designed, developed and tested with the help of simulated Heading Angle in serial asynchronous packet, created by COM Port Toolkit application software on Window operating system based PC. The heading angles in Resolver format at the output have been measured on API (Angle Position Indicator) and VLL voltages have been measured on Oscilloscope. The heading angle values have been found correct with some minute deviation as mentioned in Table-1 and 2. This embedded interfacing solution will allow seamless interfacing of equipment, sensors and weapon systems from diverse origin to work in tandem to enhance operation capability of a marine vessel platform. Here, the major module used for conversion of Digital to Resolver format has 16 bit resolution and ±4 minutes accuracy and it can operate from 0 to 70ºC wide temperature range [8]. The Embedded Solution also could be utilized to integrate various Digital and Resolver types of systems and allows distribution of important navigational heading angle parameters to various sink equipments, spread across the platform of the marine vessel.

REFERENCES

[1] Reddy, G.S. & Saraswat, V.K. Advanced navigation system for aircraft applications. Def. Sci. J., 2013, 63(2), 131-137. doi: 10.14429/dsj.63.4254.

[2] Sheetal Saroha; Rajeshwar Dass, IEEE 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN) - Noida, India (SPIN), Design & Implementation of Low Power Delta Sigma ADC. 193-197 doi:10.1109/SPIN48934.2020.9070852.

[3] Hoang, N.; Jenkins, M. & Karangelen, N. Data integration for military systems engineering. In Proceedings of the IEEE Symposium and Workshop on Engineering of Computer Based Systems, 1996. pp. 11-15. doi:10.1109/ecbs.1996.494504.

[4] Texas Instruments datasheet "onAM26LS32AC, AM26LS32AI, AM26LS32AM AM26LS33AC, AM26LS33AM" SLLS115F – October 1980 – Revised September 2016.

[5] Jie Liang and Ran Duan, "Design of the Embedded Serial and Parallel Communication Protocol Controller", IEEE, pp. 436-439, 2010.

[6] M. Muhammad Ali, Book "8051 Microcontroller and Embedded Systems using assembly and c", second edition.

[7] Serial Port Complete, 2nd Ed. by Jan Axelson.

[8] Datasheet of ©1995, 1999 DDC (Data Device Corporation) module, DSC-11524, 16-Bit Hybrid D/S and D/R converter pin programmable for Synchro or Resolver output.

[9] G. Ashok Kumar; R Ashish; K. Naveen; R. Ravi, IEEE 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN) - Noida, India (2020.2.27-2020.2.28), Design and Implementation of High-Speed Universal Asynchronous Receiver and Transmitter (UART), 295-300. doi:10.1109/SPIN48934.2020.9070856.

[10] Lei, D., Chang-shi, X., Yuan-qiao, W., & Le, W. (2015). Safe speed of ships passing access channel of port. 2015 International Conference on Transportation Information and Safety (ICTIS). doi:10.1109/ictis.2015.7232076.

[11] Okamoto, S. & Sycara, K. Augmenting ad hoc networks for data aggregation and dissemination. In Proceedings of the 28th IEEE Conference on Military Communications, 2009, pp. 2346-2352. doi:10.1109/milcom.2009.5379977.

[12] Yongjian Tang, Lenian He and Xiaolang Yan, "A novel data processing in high speed serial communication", IEEE, pp. 1228-1231, 2005.

[13] C, Kishor; K. Ashok, Defence Science Journal, Vol.64, No.5, September 2014, pp.445-450, DOI:10.14429/dsj.64.5035 "Designing Integration Unit to Integrate Navigational and Tactical Equipment onboard Naval Platforms".

[14] Popa, M., Popa, A. S., Cretu, V., & Micea, M. (2006). Monitoring Serial Communications in Microcontroller Based Embedded Systems. 2006 International Conference on Computer Engineering and Systems. doi:10.1109/icces.2006.320425.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:38:57 UTC from IEEE Xplore. Restrictions apply.

2021 Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 28-29, 2021

Internet of Things Based Energy Harvesting using Piezoelectric Discs

Deva Nand
Department of Electronics and Communication Engineering
Delhi Technological University
Delhi, India
devkamboj@dce.ac.in

Aakash Yadav
Department of Electronics and Communication Engineering
Delhi Technological University
Delhi, India
yaakash031@gmail.com

Pradeep Kumar
Department of Electronics and Communication Engineering
Delhi Technological University
Delhi, India
kspradeep06@gmail.com

Srinivas Kumar
Department of Electronics and Communication Engineering
Delhi Technological University
Delhi, India
srinivaskumar0999@gmail.com

2021 Asian Conference on Innovation in Technology (ASIANCON) | 978-1-7281-8402-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ASIANCON51346.2021.9545006

***Abstract*— Presently there has been an increased demand for energy as conventional sources of energy is going to reduce rapidly. It leads to create a high demand for energy sources and everybody knows that these sources are limited in the environment. So here require to think about other sources of energy and need to go with it while keeping in mind, that it should not affect human life. Various sources of energy conservation methods produce noise, air and water pollution, etc. The implementation of this also to be cost-effective. In this scope, piezoelectric discs become good for energy harvesting and storage for future applications. It takes advantage of the energy generated when people walking on it. This converts it into electric energy which used in an electronic gadget that requires power. This research paper is based on a new approach to solving the above problems by introducing here an Internet of things based energy harvesting system using piezoelectric discs. Which are capable to drive energy to electronic devices from stored energy in the battery. This setup is just not only harvested the energy although anyone can see the energy-saving related parameters voltage, current, battery level. It is also able to handle functions like mobile charging station application with the help of an Internet of thing based android application remotely in contrast with smart cities. Designing and implementing work is done for a real-time prototype system and tested for energy harvesting.**

Keywords*— *Energy harvesting, Piezoelectric, IoT, Power generation.

I. INTRODUCTION

Today everyone confronting an energy emergency expanding step by step traditional assets of energy are decreased. There is a need to go with the non-conventional source of energy while remembering that it can't influence human existence. Different ways of energy preservation produce commotion, air, water contamination, and so forth and furthermore, the execution of this ought to be cost-effective. So remembering this whole factor developed a piezoelectric plate-based energy harvesting framework where the expense of execution is extremely low. Quite possibly the most arising innovations these days is Foot Step Energy Harvesting utilizing a piezoelectric transducer. In this, a piezoelectric plate source conveys electric energy when mechanical pressure is applied by human movement. Pathways, shopping centers steps, parks can have piezoelectric circles inside them which generate energy when any individual step climbing or stroll on these pathways tracks the mechanical energy changes over into electrical energy and it is saved by battery. The delivered energy is likewise reliant on the number of footfalls and the heaviness of the people who are passing on these piezoelectric circles. So, as the more number of footfalls and the heaviness of the people are increased the power generation factor is increased. While later the energy is more noticeable [1]. The piezoelectric circle gives a decent voltage of generation while it gives little flow into microampere. For taking great energy need to do the series-parallel combination to upgrade the flow level of piezoelectric plates [2]. The same arrangement is shown in Fig. 1. So the resultant energy will likewise help by this arrangement. Energy is briefly put away in a capacitor and forever in a dc lithium-particle battery. The dc booster section is likewise being utilized for expanding further the limit of the framework for driving the loads which is a mobile charging station [3]. On the above said principle, deeply work on a prototype that is able to harvest the energy. Built a dashboard that is able to show all parameters while in operating state and also provide the functionality to automate the mobile charging station with the help of the Internet of things setup.

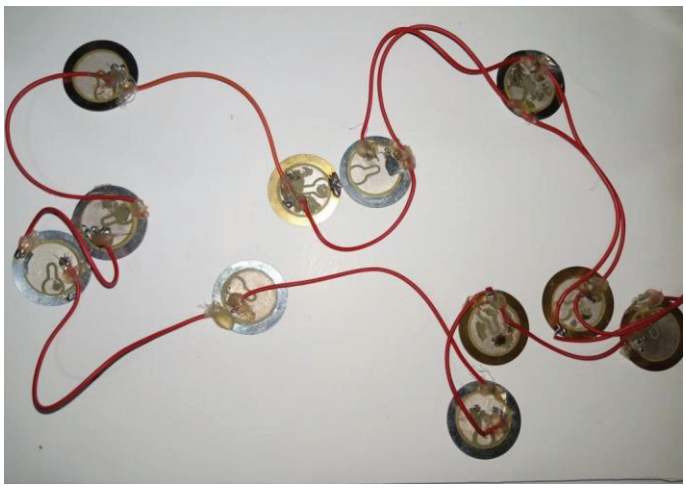


Fig. 1. Series and parallel combination of piezoelectric discs.

II. SYSTEM DESIGN

With the aim of validating the hardware of the proposed system. Basically, the system is divided into six sections.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:35:53 UTC from IEEE Xplore. Restrictions apply.

Energy generation is done with the help of piezoelectric discs connected in a series-parallel combination. The rectification section consists of a full-wave bridge rectifier that converts piezoelectric disc AC voltage into DC voltage which can be useful for many purposes. The energy boost section consists of an XL6009 DC-DC converter module which is used to boost the DC voltage and current level. The energy storing section consists of three 4.7 microfarad capacitors and a 3.7v li-ion battery. The electric energy generated when the mechanical pressure is exerted on the piezoelectric discs [4] charges the capacitor and then this charged capacitor is discharged through the battery. The load section consists of portable electronic devices such as a wifi router, Mobile phone which is operated from the energy source stored in the battery. The Internet of Things (IoT) section consists of NodeMCU [5], Voltage sensor and Arduino IDE. NodeMCU is the heart of this system and it binds the IoT and Arduino IDE acts as an interface to connect them. Voltage sensor work as a voltage divider. It is used to measure and calculate the DC voltage and current level using the analog read pin of NodeMCU IoT [6] is used to make this prototype remotely accessible energy saving related parameters like voltage, current, battery level and also handles functions like mobile charging station application very useful for smart cities [7].In this section, a brief introduction to every section of our designed system is discussed as shown in Fig. 2.



Fig. 2. Block diagram of proposed system.

A. Energy Generation Section

A piezoelectric disc is a device that works on the principle of piezoelectric effect which means whenever gives the mechanical pressure on a piezoelectric disc it delivers electrical energy. There are different types of piezoelectric materials. Some examples of piezoelectric materials are crystal quartz, Lead zirconate titanate, Polyvinylidene fluoride etc. In the proposed system the sixteen piezoelectric discs of 27mm in a series-parallel combination to maintain the balance of both current and voltage generation as shown in Fig. 1.

B. Rectification Section

A diode is a two-terminal semiconductor device that has anode and cathode terminals. Diodes lead current in just a single way that has low opposition one way and high obstruction toward another path. PN junction diode bridge as a full-wave rectifier is being used here the piezoelectric disc is produced alternating current and fed to the bridge diode network that converts AC to pulsating DC presented in Fig. 3(a).

C. Energy Boost Section

XL6009 DC-DC Step-up module is utilized to boost the DC voltage and current. It has one input pair where the input has to be given while a second pair worked as the boosted energy pair. The input range DC 3 to 32 volts while output 5 to 32 volts. It is a high exchanging recurrence, 400 kHz, the working temperature is - 40 ~ +85 as shown in Fig. 3(c).

D. Energy Storing Section

Capacitors are utilized to store instantaneous charges in electrical circuits consists of protecting dielectric material between two plates isolated by inverse and equivalent charges. Three 4.7 microfarad capacitors are utilized for the proposed framework by formula Q = C.V. The capacitor used as temporary storage while a lithium-ion battery work as permanent storing element of 3.7v as shown in Fig. 3(b) and Fig. 3(d).

It shows that the most extreme voltage of the cell is 3.7V. The current limit of this lithium Battery is 1300 mAh.

E. Load Section

This part comprises the utilities of gathered energy for utilizing this energy as charging electronic gadgets like cell phone, wifi router etc. The client utilizes assets like energy and utilities in an individual way. The mobile charging station comes under this segment as shown in Fig. 3(e).

F. IoT Section

Internet of Things (IoT) area comprises NodeMCU, Voltage sensor and Arduino IDE. NodeMCU is the core of this prototype and it ties the IoT and Arduino IDE as an interface to associate them. Voltage sensor used as a voltage divider. It is utilized to quantify and compute the DC voltage, current level utilizing the analog read pin of NodeMCU. It is used to providing the IoT dashboard and operation remotely and also to see energy-saving related parameters like the voltage, current, battery level as shown in Fig. 3(f) and Fig. 6.



Fig. 3. (a). Bridge rectification section.



Fig. 3. (b). Temporary energy storing section.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:35:53 UTC from IEEE Xplore. Restrictions apply.

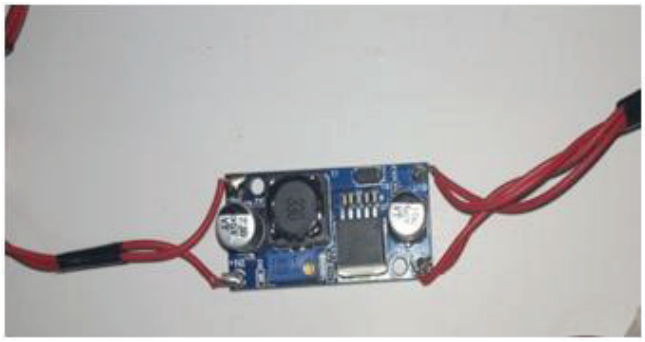


Fig. 3. (c). Energy boost section.



Fig. 3. (d). Permanent energy storing element.



Fig. 3. (e). Load mobile charging station.

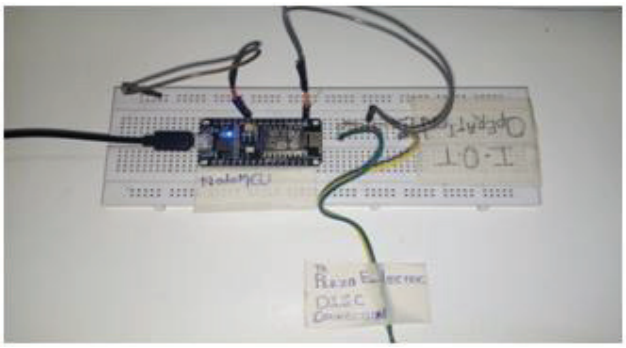


Fig. 3. (f). Internet of things section.

III. EXPERIMENTAL SETUP

In the system implementation, series and parallel combination of piezoelectric discs are employed to increase voltage and current capacity from each piezoelectric disc. It delivers electric energy when mechanical stress is applied i.e. when someone is pressed or extended after that bridge rectifier is to rectify it to pulsating DC. A capacitor is used for filtering purpose and to store an instantaneous electric charge 3.7V lithium-ion battery is going to harvest the energy from the piezoelectric disc and use this as a power source for other utilities. XL6009 converter module is used to boost the voltage and current obtained from the battery charged by the piezoelectric disc. It also provides power to NodeMCU. A voltage sensor is equivalent to a voltage divider circuit connected with an analog A0 pin of NodeMCU. To measure the voltage, current level and shown in the IoT dashboard application to control remotely. This stored energy in the battery is used as power stations to operate different loads and in various applications. Examples:-Micro USB-based electronic gadgets, i.e. Bluetooth earphones, Mobile phones etc.

The section design consists of all the related components according to the circuit diagram as shown in Fig. 4. The different components and arrangements such as piezoelectric discs in parallel and series combination are used, three capacitors whose value is 4.7 microfarad, 3.7 Volt battery, xl6009 boost converter, a resistor of value 10k, 100k and a NodeMCU ESP8266. For design a successful setup thereafter designed hardware prototype arrangement setup which provides current and voltage when real-time mechanical pressure applied on piezoelectric discs. In the hardware section, these IoT dashboard control by mobile application shows the different parameters in the form of current, voltage, power and also control the mobile charging station.

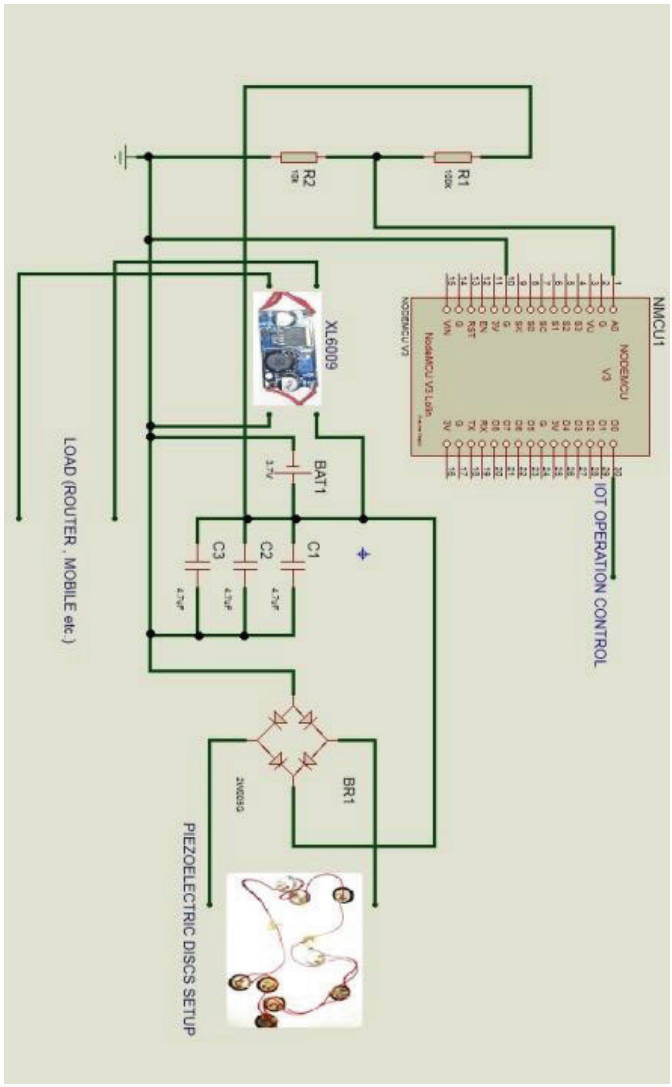


Fig. 4. System implementation.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:35:53 UTC from IEEE Xplore. Restrictions apply.

IV. HARDWARE SETUP

In this section, a hardware model is built on the basis of circuit diagram so all components connected according to the circuit on which worked in the experimental setup as shown in Fig. 5 and Fig. 6.

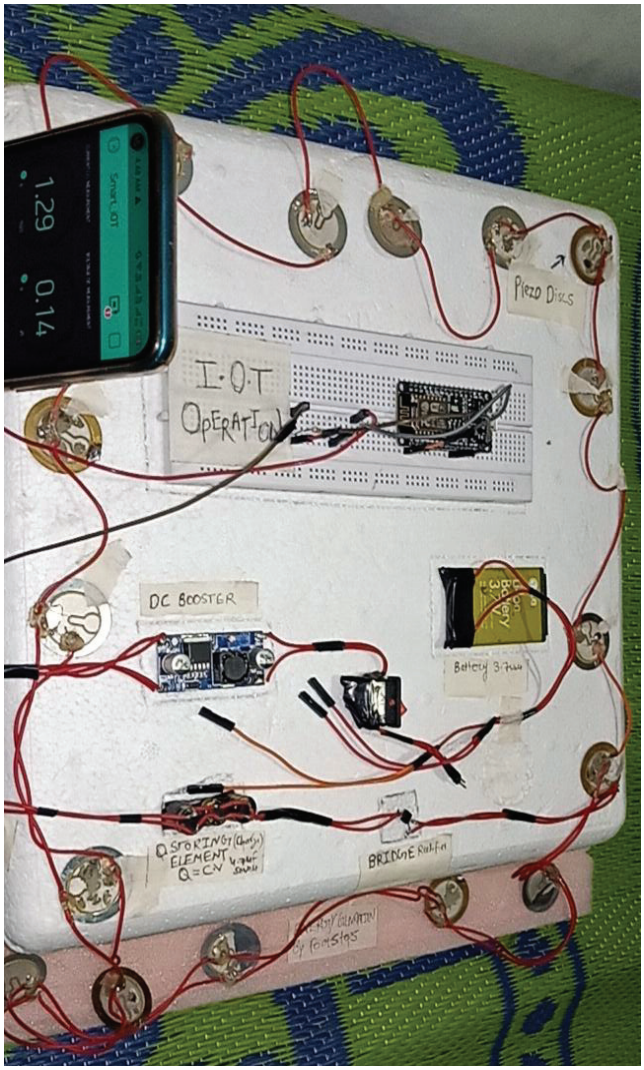


Fig. 5. Proposed prototype model.

An android mobile application [8] built with the help of blynk, the NodeMCU library, in this application anyone can see various parameters related to energy harvesting like the number of voltages and current power meanwhile also control the load section's mobile USB charging station via this mobile application as in Fig. 6. With the help of internet of things concept. The Arduino IDE used to program this NodeMCU controller and wrote a program to observe the parameters above discussed.

V. CALCULATION

A. Experimental Observation by Single Piezoelectric Disc

It is observed that various parameters result by pressing a single piezoelectric disc by hand and can be seen on blynk application as shown in Fig. 6 the parameters like voltage, current and power when pressing the disc with a different type of forces generate the different outputs based on how much pressure we exerted on the disc as can be shown in Table I.

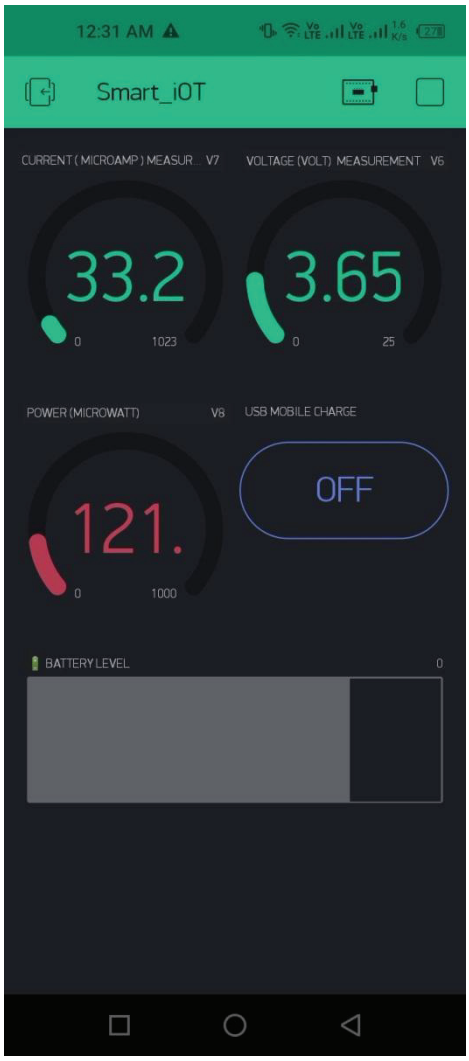


Fig. 6. Internet of things based dashboard application.

TABLE I. BLYNK APPLICATION OBSERVATION BY PRESSING SINGLE DISC

Observation count	Voltage *(V)*	Current *(μA)*	Power *(μW)*
1	1.03	9.36	9.640
2	1.35	12.2	16.50
3	1.66	14.5	23.10
4	2.02	18.3	37.10
5	2.23	20.3	45.40
6	2.41	21.9	52.90
7	2.66	24.1	64.30
8	3.19	29.0	92.70
9	3.65	33.2	121.0
10	3.76	34.1	128.0

Let's choose the one reading from Table 1. of voltage, current, and power when the pressure is applied using a hand on a piezoelectric disc. The chosen values 2.02 volt, 18.3 μA, and 37.10 μW respectively.

Total sixteen discs used in series-parallel combination of piezoelectric discs.

B. Voltage Generation by Piezoelectric Discs

We employed eight piezoelectric discs as series and others in parallel for extracting a good amount of voltages. Eight discs work as a series combination while another disc

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:35:53 UTC from IEEE Xplore. Restrictions apply.

equivalent to one disc. The total number of equivalent discs for voltage generation = 9

So the total generated voltage in the circuit will be = 9 * 2.02 V

=18.18 Volt

C. *Current Generation by Piezoelectric Discs*

Similarly, here eight discs work as a parallel combination while another disc equivalent to one disc.

The total number of equivalent discs for current generation = 9

The amount of generated current will be = 9 * 18.3 µA

= 164.7 µA

D. *Total Power Generated by the Series-Parallel Circuit of Piezoelectric Discs.*

Power (P) = Total voltage generated * Amount of current generated

= 18.18 Volt * 164.7 µA

= 2994.25 µW

E. *Energy Stored in Capacitor by Discs Combination*

Three capacitors are used in parallel combination to store the energy each of 4.7 µF having an equivalent capacitance of 14.1 µF.

Energy (E) = ½ CV^2

= ½ * (14.1) * $(18.18)^2$

= 2330.12 joule

F. *Charging Time Calculation of Battery*

The time required to charge the battery = Battery capacity / Charging current

Where, Battery capacity in mAh= 1300

Charging current in mA = 390

= 1300 mAh / 390 mA

= 3.33 Hour approximately.

VI. RESULT

In this research paper, a prototype of an energy harvesting power system is designed for keeping in mind the view of a smart city where these piezoelectric discs can employ anywhere like malls, parks, railway stations stairs etc. The energy can be generated easily with the help of piezoelectric discs. The harvested energy is used for mobile charging station where anyone can charge their electronic devices through harvested energy stored in battery via piezoelectric discs. Sixteen piezoelectric discs in the prototype are being used and a DC battery 3.7 Volt easily capable of charging the mobile phone and another electronic gadget too. Here included some remote function to disable enable the mobile charging station with the help of an IoT dashboard which shows in the android application. The application built with the help of the blynk library of NodeMCU to show various parameters like current, voltage, power, battery level and works with remote operation too.

Different test observations are recorded and different results as explained below. It was tested by pressing the piezoelectric discs by foot walks with a weight of 52 Kg.

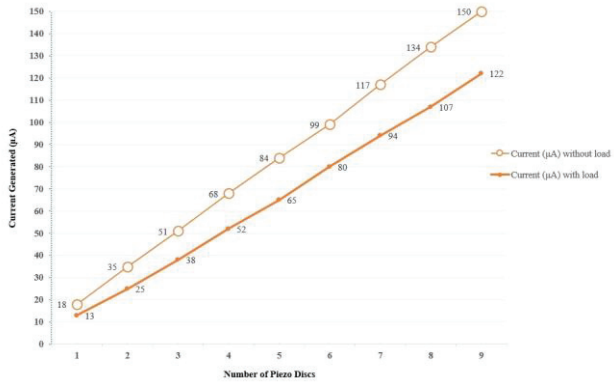


Fig. 7. (a). Current generation due to number of piezoelectric discs.

A. *Current Generation by Walking Test on Piezoelectric Discs*

Eight piezoelectric discs were placed in parallel and eight piezoelectric discs in series overall to act as one.

Every piezoelectric disc provides information on how much voltage is generated with load and without load when walked on it. The graph above Fig. 7(a) shows the Current generated (microampere) increases as more piezoelectric discs were pressed.

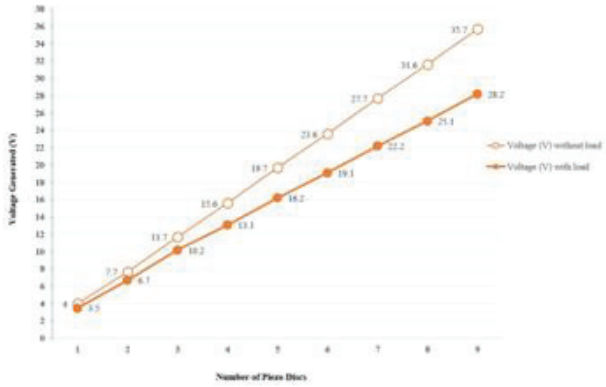


Fig. 7. (b). Voltage generation due to number of piezoelectric disc.

B. *Voltage Generation by Walking Test on Piezoelectric Discs*

Eight piezoelectric discs were placed in series and eight piezoelectric discs in parallel overall to act as one.

Similarly, the voltage generated after piezoelectric discs pressed. The graph above Fig. 7(b) shows the voltage generated (volt) increases as more piezoelectric discs were pressed.

VII. CONCLUSION

From all the different methods available to integrate into a real-time energy harvesting prototype from the simple human walk, piezoelectric discs seem to be one of the most promising elements. This research paper presents a prototype of piezoelectric energy harvesting with IoT enabled through android application for measuring different parameters like voltage, current, power, and battery level monitoring to access remotely. Around Billions individuals use smartphones, every mobile phone user wishes he save more battery and a method to charge easily whenever required from simply walking. Energy harvesting from human movements is an appealing methodology for getting perfect and feasible energy this

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:35:53 UTC from IEEE Xplore. Restrictions apply.

harvested energy in two ways: first for operating the IOT device NodeMCU which operates on +5 volts and second is the mobile charging station where anyone can charge the mobile. Here the DC buck converter is also playing an important role as it amplifies the harvested power from piezoelectric discs and the same is fed for operating both loads. This system commercially provides a vast energy-saving and is working best in crowded areas where people board and deboard in huge amounts like in parks, railway stations, malls, etc. finally observed that results obtained by this piezoelectric disc prototype totally depend on the pressure exerted on the discs, the number of piezoelectric discs used and the way of connecting piezoelectric discs, the number of footsteps, and the quality of piezoelectric discs.

ACKNOWLEDGMENT

We value the commitment of our guide mentor Deva Nand for assisting us with writing this research paper.

REFERENCES

[1] Johnny Nelson Savina Quispe and Alex Cartagena Gordillo, "Implementation of an Energy Harvesting System by Piezoelectric Elements Exploiting the Human Footsteps," IEEE URUCON, 2017

[2] Shrimoyee Poddar, Mohuya Dutta, Debashree Chowdhury, Abhinaba Dey, Debasis Maji, "Footstep Voltage Generator using Piezo-Electric Transducers," International Journal of Scientific & Engineering Research, Volume 8, Issue 3, March 2017

[3] Monika Jain, Usha Tiwari and Mohit Gupta, "Mobile Charger Via Walk," International Conference on Multimedia, Signal Processing and Communication Technologies, 2011

[4] Juliana Johari and Tuan Muhammad Afif Tuan Rashid, "Optimization of Piezoelectric Transducer Placement in Shoe Insole for Energy Harvesting," International Conference on Electrical, Electronics and System Engineering(ICEESE), 2017

[5] Teo Swee Ann, "Espressif Systems",

[6] https://www.espressif.com/en/products/modules/esp8266

[7] Joses Paul P, Samuel Desmond Tutu R, Kevin Richards W and Maria Jerome V, "Project Power Shoe: Piezoelectric Wireless Power Transfer - A Mobile Charging Technique", IEEE Global Humanitarian Technology Conference (GHTC), 2015

[8] Laurence Kemball-Cook, "Smart Cities Pavegen",

[9] https://pavegen.com/smart-cities/

[10] Steve August, "Blynk IoT platform", https://blynk.io/

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:35:53 UTC from IEEE Xplore. Restrictions apply.

2021 8th International Conference on Signal Processing and Integrated Networks (SPIN) | 978-1-6654-3564-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SPIN52536.2021.9566034

LCNT incorporated domino with footed diode inverter

NEETIKA YADAV[a,b]
a. Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, 110042, India

b. Department of Electronics and Communication Engineering, Amity School of Engineering and Technology (affiliated to GGSIPU), Sector 125, Noida 201303, India

yadav.neetika92@gmail.com

NEETA PANDEY[a]
a. Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, 110042, India
n66pandey@rediffmail.com

DEVA NAND[a]
a. Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, 110042, India
devkamboj07@gmail.com

***Abstract*— This contribution proposes a technique for leakage power reduction in footed domino circuits by incorporating Leakage Control NMOS Transistor (LCNT) technique along with a footed diode inverter. The analysis is done using footed domino gates for two, four and eight input OR gates to investigate the proposed techniques at 90nm,45nm and 32nm technology nodes in both precharge and evaluation phase using SymicaDE tool. Analysis of leakage power reveals that its value increases with technology scaling. The total leakage power saving is (23.7%-71.8%) in precharge phase and (22.6%-77.1%) in evaluation phase using the proposed approach. The efficiency of the proposed approach increases with scaling of technology i.e., the leakage power saving achieved by the proposed approach improves at lower technology node.**

Keywords— Dual mode logic, dynamic, footed diode transistor, GALEOR, static

I. INTRODUCTION

Nowadays, power consumption is a major concern in applications pertaining to portable electronics due to availability of limited battery resources [1-2]. Further, the device density on silicon is increasing with each technology generation which causes power density to rise [3]. In technology nodes below 90nm, there is proliferation in leakage power due to smaller feature size and threshold voltage reduction [4]. The leakage power is becoming comparable to dynamic power [4]. This has prompted the researchers to search for techniques that mitigate leakage power which is evident from the techniques available in [5-21]. These techniques are applied to static and dynamic circuits based on CMOS, FinFET etc. [5-21]. Domino logic is yet another logic family that acts as a high performance, low power and low area alternative to dynamic logic [22]. However, it also witnesses an increase in leakage power as technology node is scaled. To combat leakage power in domino logic, substantial research has already been done [18-21].

Leakage Control NMOS Transistor (LCNT) [13] is a self-controlled leakage reduction technique which minimizes the leakage current. It employs two NMOS transistors, which are referred to as Leakage Control Transistors (LCT), to introduce stacking effect. The gate terminal of these two transistors is tied together to form the output node. The voltage level at the output node governs the switching of two LCTs. These two leakage control transistors offer high resistance to the current flowing from power supply to ground, as a result of which the leakage current and eventually leakage power reduces. This work aims at incorporating LCNT technique with a footed diode for leakage power reduction in domino circuits at different technology node.

The organization of the paper is as follows: section 2 discusses footed domino logic and the associated leakage mechanism. In section 3, proposed LCNT incorporated domino logic with footed diode is explained. Section 4 states the simulation results for two, four and eight input OR gates based on proposed approach in pre-charge and evaluation phase and subsequently, the conclusion of the paper is placed in section 5.

II. FOOTED DOMINO LOGIC AND ITS LEAKAGE MECHANISM

Domino logic has two variants- Footed and Unfooted design. A footed domino configuration employs an additional footer transistor [22]. A standard footed OR-domino logic is shown in fig.1. It consists of a precharge transistor (TP_1) and a footer transistor (N_2) along with a keeper transistor (TP_2). The keeper transistor is used to preserve the logic level at the dynamic node (DN). The circuit is operated in two phases with the help of a clock (CLK) signal i.e., precharge phase and evaluation phase. When the CLK is LOW in precharge phase, the output becomes LOW as the dynamic node is charged to VDD through precharge transistor (TP_1). The applied inputs don't affect the output of the domino logic but the leakage current is governed by the inputs applied. For evaluation phase, a HIGH logic level is applied at the CLK signal. The precharge transistor (TP_1) is off and the footer transistor (N_2) is on. The output depends on the type of logic implemented in evaluation network. The switching of transistor, TP_2 depends on the voltage level at the output node. Charging and discharging of the dynamic node is dependent on the inputs applied. There is a leakage current which flows from supply to ground due to the presence of corresponding OFF transistors.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:34:02 UTC from IEEE Xplore. Restrictions apply.

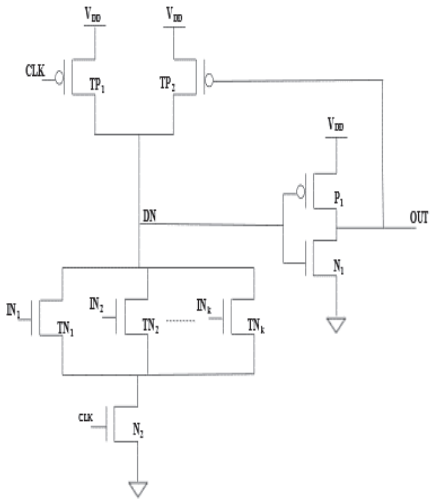


Fig.1.Footed domino OR-gate

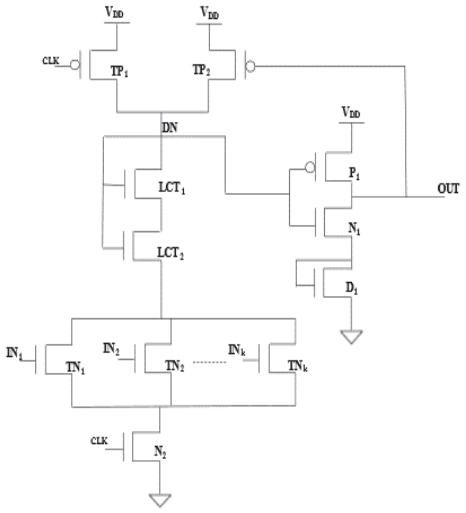


Fig. 2. LCNT incorporated footed domino OR gate with footed diode inverter

III. PROPOSED LCNT INCORPORATED DOMINO LOGIC WITH FOOTED DIODE

The proposed LCNT incorporated footed domino OR-gate with footed diode inverter is illustrated in fig.2, two LCT transistors- LCT_1 and LCT_2, are placed between the precharge and evaluation network of the domino design. The dynamic node is connected to the gate terminals of LCT_1 and LCT_2. The switching of LCT_1 and LCT_2 is governed by the voltage level at the dynamic node. A footed diode is an NMOS transistor with connected gate and drain terminals [13]. It is placed at the output inverter above the GND in the footed domino circuit. The footed diode introduces stacking effect which leads to a further decrease in leakage current.

In precharge phase (CLK=LOW), the transistor TP_1 is turned on and the dynamic node is charged to HIGH due to which the output becomes LOW. Both the LCT transistors LCT1 and LCT2 are turned off, which leads to an increase in the resistance to current flow from V_{DD} to ground. As a result, the leakage current decreases.

In evaluation phase (CLK=HIGH), when all the inputs are LOW, transistors TN_1-TN_k are turned off and there will be no discharge of dynamic node, as a result output would be LOW. Due to leakage in OFF transistors (TN_1-TN_k, TP_1), a path is created for the current to flow between V_{DD} and ground. This flow of current contributes to leakage in the circuit. Consider the case when all inputs are HIGH, there is discharging of dynamic node through ON transistors. As a result, the output becomes HIGH which makes transistor TP_2 off but due to leakage in transistor TP_1 and TP_2, a path is created between V_{DD} and ground which allows the leakage current to flow. In both the cases, LCT_1 and LCT_2 are turned OFF, which introduces stacking effect and reduces leakage current.

IV. SIMULATION RESULTS

The two, four and eight input OR gates are simulated in a footed domino and proposed LCNT incorporated domino with footed diode inverter configuration using 90nm, 45nm and 32nm PTM models for CMOS technology. SymicaDE tool is used for simulation at 1.2 V power supply and load capacitor, C=1fF. The (W/L) of all NMOS and PMOS are kept at (120nm/90nm), (120nm/45nm) and (120nm/32nm) for 90nm, 45nm and 32nm technology respectively. The leakage power is measured by applying the same inputs to both the circuits i.e., original and the proposed circuit. There may be switching at the output due to input change, therefore the power is measured after output settles down, thus ensuring that measured power is due to leakage only. Total leakage power evaluation is done for two input conditions of OR gates i.e., all inputs LOW and all inputs HIGH in precharge and evaluation phase of existing and proposed domino circuits. This section consists of three subsections-leakage power analysis in precharge phase, leakage power analysis in evaluation phase and effect of load capacitance.

A. LEAKAGE POWER ANALYSIS IN PRECHARGE PHASE

The leakage currents for domino and proposed design in precharge phase for OR2 gate at 32nm are shown in Fig.3. As observed from the Fig.3, the maximum leakage current in domino is 96.98 uA whereas for the proposed design, it has been reduced to 36.6 uA. The total leakage power of OR circuits - OR2, OR4 and OR8 at 90nm, 45nm and 32nm technology node at 27℃ in precharge phase is shown in fig. 4.

For precharge phase, leakage power saving is tabulated in table 1. As observed from table 1, the proposed technique is effective in reducing the leakage power. In precharge phase, the maximum power saving achieved using proposed approach is 28%,49.5% and 71.8% at 90nm, 45nm and 32nm respectively. The efficacy of the proposed approach in precharge phase improves as we move towards lower technology node as shown in fig. 5.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:34:02 UTC from IEEE Xplore. Restrictions apply.

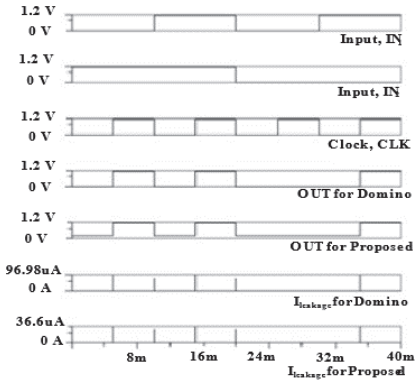
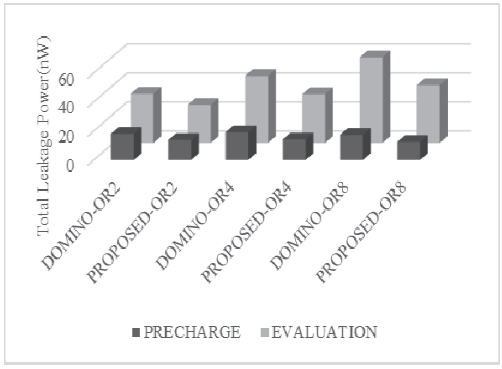
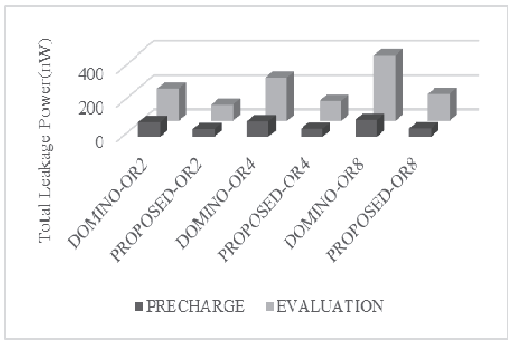


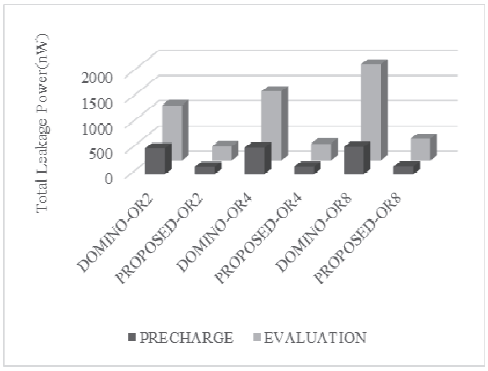
Fig.3. Timing diagram of domino and proposed approach for OR2 gate at 32nm



(a)



(b)



(c)

Fig.4. Total leakage power comparison for two, four and eight input OR gates in precharge and evaluation phase (a) 90 nm technology (b) 45 nm technology (c) 32 nm technology

Total Percentage Leakage Power Saving, Precharge phase			
	OR2	OR4	OR8
90nm	23.71	25.79	27.96
45nm	46.96	47.93	49.51
32nm	71.56	71.62	71.76

Table 1 Total percentage leakage power saving for two, four and eight input OR gates in precharge phase using proposed approach at 32nm, 45nm and 90 nm.

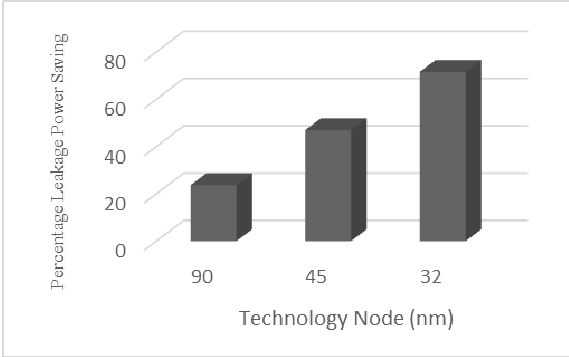


Fig.5. Percentage leakage power saving with technology node scaling for OR2 gate in precharge phase

B. LEAKAGE POWER ANALYSIS IN EVALUATION PHASE

The total leakage power of OR circuits- OR2, OR4 and OR8 at 90nm, 45nm and 32nm at 27℃ is shown in fig. 4. The leakage power saving achieved by the proposed approach in evaluation phase of operation is tabulated in table 2. As observed from table 2, the leakage power is significantly reduced by the proposed technique. In evaluation phase, the maximum power saving achieved using proposed approach is 32.5%, 59.6% and 77.1% at 90nm, 45nm and 32nm respectively. As illustrated in fig. 6, the efficacy of the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:34:02 UTC from IEEE Xplore. Restrictions apply.

proposed approach in evaluation phase improves as the technology node is scaled from 90nm to 32nm. The leakage currents for domino and proposed design in evaluation phase for OR2 gate at 32nm are shown in Fig.3.

Total Percentage Leakage Power Saving, Evaluation phase			
	OR2	OR4	OR8
90nm	22.58	27.23	32.49
45nm	51.75	55.17	59.58
32nm	74.07	75.26	77.07

Table 2 Total percentage leakage power saving for two, four and eight input OR gates in evaluation phase using proposed approach at 32nm, 45nm and 90 nm

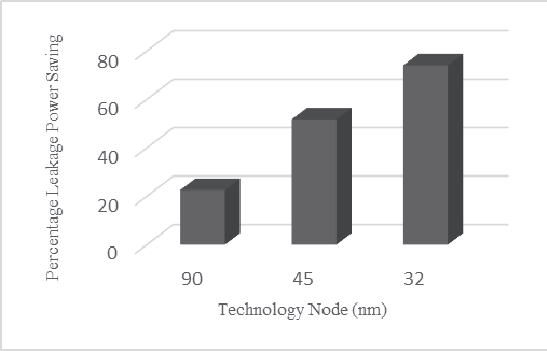


Fig.6. Percentage leakage power saving with technology node scaling for OR2 gate in evaluation phase

C. Effect of load capacitance

As the load capacitance value is changed for OR-2 gate from 1fF to 80fF at 32nm, there is no change in leakage power in both precharge and evaluation phase as enlisted in table 3. Leakage power, given by (1), is mainly due to subthreshold current [17].

$$P_{leakage} = I * V_{DD} \quad (1)$$

where V_{DD} is power supply

I is leakage current given by (2)

$$I = I_0\, e^{(V_{gs}-V_t)/(\alpha V_{th})} \quad (2)$$

where V_t is the device threshold voltage;
V_{th} is the thermal voltage (V_{th} = 25.9mV at room temperature);
I_0 is the current when $V_{gs} = V_t$

The parameter a is a constant depending on the device fabrication process, ranging from 1.0 to 2.5.

Leakage current is independent of load capacitance as observed from (2).

Leakage Power(nW), Precharge phase						
Capacitance (fF)	1	5	20	40	60	80
DOMINO-OR2	515.63	515.63	515.63	515.63	515.63	515.63
PROPOSED-OR2	146.66	146.66	146.66	146.66	146.66	146.66

Leakage Power(nW), Evaluation phase						
Capacitance (fF)	1	5	20	40	60	80
DOMINO-OR2	1097.7	1097.7	1097.7	1097.7	1097.7	1097.7
PROPOSED-OR2	284.64	284.64	284.64	284.64	284.64	284.64

Table 3 Leakage power for 2-input OR gate in precharge and evaluation phase for different load capacitance at 32nm

V. CONCLUSION

A novel LCNT based leakage power reduction technique for footed domino circuits is presented in this paper. The proposed design incorporates LCNT technique with a footed diode transistor, which is effective in reducing the leakage power in footed domino circuits. The simulation results show that leakage power is more dominant and the proposed approach shows better leakage power saving as we move towards lower technology node i.e., from 90nm to 32nm. In precharge phase of operation, the maximum power saving of is achieved for 71.6%,71.6% and 71.8% for two input (OR2), four input (OR4) and eight input (OR8) OR-gate. Corresponding values in evaluation phase are (74.0%, 75.3%,77.1%) for (OR2, OR4, OR8). So, the proposed technique provides substantial leakage reduction in footed domino circuits.

References

[1] G. K. Yeap, Practical low power digital VLSI design. Springer Science & Business Media, 2012.

[2] T. Sakurai, Perspectives of low-power VLSI's, IEICE transactions on electronics, vol. 87, no. 4, pp. 429-436, 2004.

[3] A. Pal, Low-Power VLSI circuits and systems. Springer, 2014.

[4] S. G. Narendra and A. P. Chandrakasan, Leakage in nanometer CMOS technologies. Springer Science & Business Media, 2006.

[5] J. Kao and A. Chandrakasan, "MTCMOS sequential circuits," in Proceedings of the 27th European Solid-State Circuits Conference, 2001: IEEE, pp. 317-320.

[6] F. Assaderaghi, D. Sinitsky, S. A. Parke, J. Bokor, P. K. Ko, and C. Hu, "Dynamic threshold-voltage MOSFET (DTMOS) for ultra-low voltage VLSI," IEEE Transactions on Electron Devices, vol. 44, no. 3, pp. 414-422, 1997.

[7] T. Kuroda, T. Fujita, F. Hatori, and T. Sakurai, "Variable threshold-voltage CMOS technology," IEICE Transactions on Electronics, vol. 83, no. 11, pp. 1705-1715, 2000.

[8] Y.-T. Ho and T.-T. Hwang, "Low power design using dual threshold voltage," in ASP-DAC 2004: Asia and

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:34:02 UTC from IEEE Xplore. Restrictions apply.

South Pacific Design Automation Conference 2004 (IEEE Cat. No. 04EX753), 2004: IEEE, pp. 205-208.

[9] N. Hanchate and N. Ranganathan, "LECTOR: a technique for leakage reduction in CMOS circuits," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 12, no. 2, pp. 196-205, 2004.

[10] P. Bajpai, N. Pandey, K. Gupta, and J. Panda, "LECTOR incorporated differential cascode voltage swing logic (L-DCVSL)," Analog Integrated Circuits and Signal Processing, vol. 100, no. 1, pp. 221-234, 2019.

[11] S. Katrue and D. Kudithipudi, "GALEOR: Leakage reduction for CMOS circuits," in 2008 15th IEEE International Conference on Electronics, Circuits and Systems, 2008: IEEE, pp. 574-577.

[12] V. L. Rani, M. M. Latha, and A. S. Ramesh, "Galeorstack-a novel leakage reduction technique for low power VLSI design," International journal of computer applications, vol. 48, no. 8, 2012.

[13] R. Lorenzo and S. Chaudhury, "LCNT-an approach to minimize leakage power in CMOS integrated circuits," Microsystem Technologies, vol. 23, no. 9, pp. 4245-4253, 2017.

[14] Bhardwaj, Aman, Vedang Chauhan, and Manoj Kumar. "Design of CMOS Based D Flip-Flop with Different Low Power Techniques." 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN). IEEE, 2019.

[15] J. W. Chun and C. R. Chen, "A novel leakage power reduction technique for CMOS circuit design," in 2010 International SoC Design Conference, 2010: IEEE, pp. 119-122.

[16] V. K. Sharma, M. Pattanaik, and B. Raj, "ONOFIC approach: low power high speed nanoscale VLSI circuits design," International Journal of Electronics, vol. 101, no. 1, pp. 61-73, 2014.

[17] N. Yadav, N. Pandey, and D. Nand, “Leakage reduction in dual mode logic through gated leakage transistors,” Microprocess. Microsyst., no. 104269, p. 104269, 2021.

[18] T. K. Gupta and K. Khare, "Lector with footed-diode inverter: a technique for leakage reduction in domino circuits," Circuits, Systems, and Signal Processing, vol. 32, no. 6, pp. 2707-2722, 2013.

[19] Z. Liu and V. Kursun, “PMOS-only sleep switch dual-threshold voltage domino logic in sub-65-nm CMOS technologies,” IEEE Trans. Very Large Scale Integr. VLSI Syst., vol. 15, no. 12, pp. 1311–1319, 2007.

[20] H. Mahmoodi-Meimand and K. Roy, "Diode-footed domino: A leakage-tolerant high fan-in dynamic circuit design style," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 51, no. 3, pp. 495-503, 2004.

[21] S. R. Ghimiray, P. Meher, and P. K. Dutta, “Ultralow power, noise immune stacked-double stage clocked-inverter domino technique for ultradeep submicron technology,” Int. J. Circuit Theory Appl., vol. 46, no. 11, pp. 1953–1967, 2018.

[22] SM Kang, Y. Leblebici, CMOS Digital Integrated Circuits, Tata McGraw-Hill Education, 2003.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:34:02 UTC from IEEE Xplore. Restrictions apply.

Soft Computing (2021) 25:15067–15081
https://doi.org/10.1007/s00500-021-06363-3

OPTIMIZATION

Matrix games with proportional linguistic payoffs

Parul Chauhan[1] · **Anjana Gupta**[1]

Accepted: 26 September 2021 / Published online: 22 October 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
This paper aims to present a novel approach for computing two-player constant-sum matrix games laid on the notion of a "symbolic proportional" linguistic term set. It is not always possible to lay assessments based on uniformly and symmetrically distributed linguistic term sets; hence, the defined concept motivates the decision-makers to represent their opinions using 2-tuples composed of two proportional linguistic terms. The proportional 2-tuple linguistic representation of payoffs concerns linguistic labels, which do not certainly have to be symmetrically distributed or do not have the conventional prerequisite of having uniform distance among them. This representation confers an opportunity to describe the linguistic payoffs of a matrix game using members of a continuous linguistic scale domain. In our work, we have defined a two-player constant-sum proportional linguistic matrix game and proposed an approach of Proportional Linguistic Linear Programming (PLLP) to evaluate these game problems. The framed PLLP problem is then transformed into a crisp LPP that can be easily solved, decreasing the computation complexities involved in solving the linguistic decision-making problems. This perspective of proportional 2-tuples provides decision-makers an approach to represent their opinions not by just using one label, rather by proportional linguistic labels of the form $(\delta u_i, \gamma u_{i+1})$, where u_i and u_{i+1} are two successive linguistic terms, with $0 \leq \delta, \gamma \leq 1$ and $\delta + \gamma = 1$. Besides, some test examples are also presented to show the consistency of our designed approach. Further, the PLLP formulation is utilized to solve a Multi-Criteria Decision-Making problem based on actual-time linguistic data.

Keywords 2-tuple linguistic term set · Proportional linguistic term set · Linguistic linear programming problems · Proportional linguistic matrix games · MCDM problems

1 Introduction

Game Theory emerges in almost every aspect of human interaction. Therefore, it is considered a powerful tool for showcasing the interactions of two or more independent decision-makers having conflicting interests. The innovative work of Von Neumann pioneers the concept of game theory in the early twentieth century. In 1944, Neumann and Morgenstern (1944) presented preliminary concepts of game theory in their published work named "Theory of Games and Economic Behavior," which forms a foundation for the present-day theories. The ultimate goal of game theory is to find the optimal solution(s) for players. This optimization means that we look for such values which maximize or minimize the payoffs of respective players. Considering the importance of optimization in different research domains, several optimization methods have been presented so far. The optimization algorithms, which cover both types of maximization and minimization issues, have many applications in quantum differential evolution Cai et al. (2021); Deng et al. (2020a, b), decision-making, differential evolution algorithms Deng et al. (2021a, b); Song et al. (2021), etc. Game theory and optimization have a conceptual overlap. In other words, game theory is a kind of approximating optimization.

The game theory framework is designed to study the strategic interactions among two or more players, who form their strategies to win the game. The outcome of these interactions typically depends on the current player's behavior and other players' conduct in the game. In a more general scenario, the result often depends on each player's expected payoff before making decisions, which measures each player's satisfaction from making the decision. Thus, the players must carry out actions and take decisions that will give them maximum payoff. However, it is envisioned

✉ Anjana Gupta
anjanagupta@dce.ac.in

[1] Department of Applied Mathematics, Delhi Technological University, Bawana Road, Rohini, Delhi, India

that several practical situations are concerned with data or elements involving uncertainty or impreciseness. Consequently, using the classical game-theoretic approaches in modeling such problems may offer results with imprecision and vagueness. In these cases, problems modeled via fuzzy subsets introduced by Zadeh (1965) seem appropriate and proffer results with high accuracy. As a result, several authors have studied fuzzy matrix games Butnariu (1978); Zaichenko (2020); Vijay et al. (2005a, b) in the literature and have proposed methodologies highly flexible to model game problems comprising uncertainty within the problem. For a detailed explanation of the methodologies, one may refer to the papers Bector et al. (2004a, b); Bector and Chandra (2005); Campos (1989); Dengfeng (1999).

Since the concept of fuzzy sets is uniquely stipulated in terms of membership functions associated with the elements, these memberships contribute significantly to fuzzy game theory. Nevertheless, the societal developments give rise to the complexity within the problems; as a result, representing the elements of payoffs in terms of fuzzy numbers is a challenging task. Moreover, in real decision-making problems, it is appropriate to estimate some outlooks using natural language. This implies the need to incline toward the linguistic aspect of a problem. Thus, Zadeh (1974, 1975a, b) introduced the concept of a linguistic variable to overcome this constraint. Till then, several linguistic computational models have been implemented to solve the problems having linguistic information Herrera and Martínez (2000, 2001a); Jiang et al. (2008); Herrera and Martínez (2001b). Further, to facilitate the players' straightforward selection of utilities, this newer form of ambiguous matrix game in terms of linguistics is suggested by Arfi (2006a, b) on the grounds of fuzzy linguistic approach. Delgado et al. (1990) have also focused on this non-conventional side of games where the players express their opinions in the forms of linguistic labels. Moreover, Sahoo (2017) has also solved matrix games having linguistic payoffs. In addition to these contributions, Singh et al. (2018) amplify the dimension of matrix games under a 2-tuple linguistic structure. In that paper, the authors have implemented the 2-tuple linguistic model Herrera and Martínez (2000) to solve a two-player linguistic matrix game having a constant sum. They have discussed the evaluation of saddle points and further formulated a Linear Programming approach involving linguistic terms to attain the respective optimal strategies of the game players. This approach has also been adopted for solving various Multi-Criteria Decision-Making (MCDM) problems. Furthermore, Singh and Gupta (2018), and Malhotra et al. (2019) extended the notion of matrix games to a newer approach of the 2-tuple linguistic model consisting of interval-valued games to increase the uncertainty incorporated within the game problem.

Considering the above-mentioned 2-tuple linguistic game problems, it is envisaged that the linguistic term set utilized for addressing the information is a balanced linguistic term set (BLTS) because its terms are symmetrically and uniformly distributed. However, in a more logical viewpoint, some problems are not well suited to deal with BLTS, instead incorporating linguistic information that is non-uniformly distributed. A linguistic term set possessing a non-symmetrical distribution of terms is named an unbalanced linguistic term set (ULTS). According to the literature, there are two types of ULTS. First, in which the linguistic terms are asymmetrically distributed, and thus the distance between them is unequal (described in Fig. 1). Furthermore, second, in which the linguistic set possesses equal terms on either side of the middle term, but the spacing among terms is uneven (as described in Fig. 2).

Several methodologies have been developed in the literature to deal with these two types of ULTS, such as Abchir and Truck (2013); Bartczuk et al. (2012); Cai et al. (2016); Cai and Gong (2017); Herrera et al. (2008); Pei and Zheng (2016); Zou et al. (2012). Dong et al. (2009) have developed a 2-tuple linguistic model based on the notion of numerical scale. Further, Wang et al. (2015) have presented a normalized numerical scaling method for dealing with ULTS. Moreover, in order to extend the existing 2-tuple model so that it might address ULTS, Wang and Hao (2006) have established an advanced form of the linguistic representation model for unbalanced sets, namely the symbolic proportional model. This newly designed so-called symbolic proportional model is interestingly easy-going and is flexible in a computation stage for CWWs allowing decision-makers to evaluate their alternatives via proportional 2-tuples of the form $(\delta u_i, \gamma u_{i+1})$ where u_i, u_{i+1} are represented as two successive linguistic terms, such that $\delta, \gamma \in [0, 1]$, $\delta + \gamma = 1$.

In this work, the scope of 2-tuple linguistic matrix games is extended within the aforementioned unbalanced linguistic framework introduced by Wang and Hao (2006) to manage the uncertainty to a large extent. For this, an approach is developed to solve constant-sum linguistic matrix games involving proportional linguistic information. The main highlights and contributions of this paper are described as follows:

1. The methodology introduced in this paper is an entirely new approach as it is concerned with the matrix games consisting of the payoffs of the type, say $(0.6M_1, 0.4M_2)$ where M_1 and M_2 denote month 1 and month 2, respectively. The notation represents the case when the production of a commodity is reported as 60% in month 1 and 40% in month 2. This representation of payoffs is based on the concept of "symbolic proportion."



Fig. 1 One type of ULTS



Fig. 2 Another type of ULTS

2. Under proportional 2-tuple linguistic context of payoffs, not only the process of "Computing with Words" may be carried out without any loss of information in some sense, but the linguistic labels can also be dealt with, without having to be symmetrically distributed around a central label and the without having the necessity of being equidistant.
3. A two-player constant-sum proportional matrix game is developed to model game problems having unbalanced proportional payoffs. To solve such game problems, a Proportional Linguistic Linear Programming (PLLP) methodology is proposed to assess the proportional matrix game.
4. Finally, this PLLP formulation is used to solve a multi-criteria decision-making (MCDM) problem based on actual-time linguistic information to strengthen the applicability of this new version of matrix game problems.
5. The novelty of this paper lies in the fact that the unbalanced information among the game players has been taken care of; the defined methodology takes into account the complicated game problems where the payoffs belong to an LT set in which the linguistic terms are distributed unevenly. Matrix games having payoffs from a balanced LT set have been considered in earlier studies, but to the best of our knowledge, this defined study is new to address matrix games with payoffs from an unbalanced LT set.

This PLLP model happens to be a modified version of the Linguistic Linear Programming Problem as it involves proportional coefficients in constraints belonging to a proportional linguistic term set. Moreover, this new methodology not just presents an appropriate scope of computing matrix games with payoffs in the form of "words," but it also offers an opportunity so that the decision-makers in several decision models can assess the functioning of alternatives by using not just a single label but with proportional 2-tuples (more flexible). To the extent of our knowledge, the proposed methodology is considered to be the first study within the literature addressing unbalanced linguistic information in matrix games.

The paper is set out in the following manner. Section 2 presents a brief review of proportional linguistic variables and few required conventions of matrix games. Then, we introduce a two-player constant-sum matrix game based on the proportional linguistic model in Sect. 3. In Sect. 4, a Proportional Linguistic Linear Programming is being formulated in relevance to the proportional matrix game. Section 5 illustrates the application of proportional matrix games in the MCDM problem, evaluating the perspective of doctors in first-line treatment of constipation based on several alternatives been chosen on different criteria. In Sect. 6, we analyze the work done in the paper. Finally, Sect. 7 offers the conclusions.

2 Preliminaries

In this segment, we present fundamental notations, definitions, and properties of proportional linguistic variables. Further, this section reviews the notion of two-player matrix games.

2.1 Proportional linguistic variables (PLV)

The linguistic 2-tuple model presented by Herrera and Martínez (2000) comprises of linguistic terms belonging to a linguistic term (LT) set $U = \{u_i,\ i = 0,\ 1, \ldots,\ g\}$. Generally, we assume the index of granularity in information uncertainty to be an even positive integer.
In the literature, Wang and Hao (2006) introduced an advanced version of the 2-tuple linguistic model as given below. This model is called the "symbolic proportional model" and is proficient enough in handling linguistic information asymmetrically.

Definition 1 Wang and Hao (2006) Suppose $U = \{u_0,\ u_1,\ \ldots,\ u_g\}$ is an ordered set, $I = [0,\ 1]$ and

$$IU \equiv I \times U = \{(\delta,\ u_i) : \delta \in [0,\ 1]\ \ \&\ \ i = 0,\ 1, \ldots,\ g\}.$$

Consider the pair $(u_i\ ,\ u_{i+1})$ of two succeeding members of the set U, then any two elements $(\delta,\ u_i), (\gamma,\ u_{i+1}) \in IU$ are named as a symbolic proportion pair and $\delta,\ \gamma$ are termed as pair of symbolic proportions of the pair $(u_i,\ u_{i+1})$ if $\delta + \gamma = 1$. A symbolic proportion pair $(\delta,\ u_i), (1 - \delta,\ u_{i+1})$ is represented as $(\delta u_i,\ (1 - \delta)u_{i+1})$, and we denote the set of all symbolic proportions as $\overline{U}$, that is

$$\overline{U} = \{(\delta u_i,\ (1 - \delta)u_{i+1}),\ \delta \in [0,\ 1],\ i = 0,\ 1, \ldots,\ g - 1\}$$

Definition 2 Wang and Hao (2006) **Negation of a Proportional Variable**

The negation of a PLV is described as:

$$Neg((\delta u_i,\ (1-\delta)u_{i+1})) = ((1-\delta)u_{g-i-1},\ \delta u_{g-i})$$

Definition 3 Wang and Hao (2006) **Position Index Function** Let $U = \{u_0,\ u_1, \ldots,\ u_g\}$ be an ordinal term set and $\overline{U}$ be the resulting ordinal proportional 2-tuple set.
The position index function $\pi : \overline{U} \rightarrow [0, g]$ of an ordinal 2-tuple is defined as

$$\pi((\delta u_i,\ (1-\delta)u_{i+1})) = i + (1-\delta)$$

where $i = 0,\ 1, \ldots,\ g-1$ and $\delta \in [0, 1]$.
π is a bijection under identification convention, and its inverse function $\pi^{-1} : [0, g] \rightarrow \overline{U}$ may be determined as

$$\pi^{-1}(z) = ((1-\gamma)u_i,\ \gamma u_{i+1})$$

where $i = E(z)$, E is the integer part function, $\gamma = z - i$ and $(1.u_g,\ 0.u_{g+1}) \equiv (0.u_{g-1},\ 1.u_g)$.

Definition 4 Wang and Hao (2006) **Comparison of Proportional 2-tuples.**
For any $(\delta u_i,\ (1-\delta)u_{i+1}), (\gamma u_j,\ (1-\gamma)u_{j+1}) \in \overline{U}$

$$(\delta u_i,\ (1-\delta)u_{i+1}) < (\gamma u_j,\ (1-\gamma)u_{j+1})$$
$$\Longleftrightarrow\ i + (1-\delta) < j + (1-\gamma)$$

- When $i < j$, then
  (a) $(\delta u_i,\ (1-\delta)u_{i+1}) = (\gamma u_j,\ (1-\gamma)u_{j+1})$ if $i = j-1$ and $\delta = 0\ \&\ \gamma = 1$.
  (b) Otherwise, $(\delta u_i,\ (1-\delta)u_{i+1}) < (\gamma u_j,\ (1-\gamma)u_{j+1})$.
- When $i = j$, then
  (a) If $\delta = \gamma$, then $(\delta u_i,\ (1-\delta)u_{i+1}) = (\gamma u_j,\ (1-\gamma)u_{j+1})$.
  (b) If $\delta < \gamma$, then $(\delta u_i,\ (1-\delta)u_{i+1}) > (\gamma u_j,\ (1-\gamma)u_{j+1})$.
  (c) If $\delta > \gamma$, then $(\delta u_i,\ (1-\delta)u_{i+1}) < (\gamma u_j,\ (1-\gamma)u_{j+1})$.

  It is worth noting that the above-mentioned comparison techniques can further be characterized by using position indices. Thus, for any $(\delta u_i,\ (1-\delta)u_{i+1}), (\gamma u_j,\ (1-\gamma)u_{j+1}) \in \overline{U}$, we have

$$(\delta\ u_i,\ (1-\delta)u_{i+1}) < (\gamma u_j,\ (1-\gamma)u_{j+1}) \Longleftrightarrow$$
$$\pi(\delta u_i,\ (1-\delta)u_{i+1}) < \pi(\gamma u_j,\ (1-\gamma)u_{j+1})$$

and

$$Neg(\delta u_i,\ (1-\delta)u_{i+1})$$
$$= \pi^{-1}(g - (\pi(\delta u_i,\ (1-\delta)u_{i+1})))$$

Definition 5 Wang and Hao (2006) Let U, $\overline{U}$ be as before and suppose $\Delta : [0,\ g] \rightarrow U \times [-0.5,\ 0.5)$ be the function defined by Herrera and Martinez in Herrera and Martínez (2000). Let $\overline{U}^T$ be the range of the Δ function. Define $h : \overline{U} \rightarrow \overline{U}^T$ by

$$h((\delta u_i, (1-\delta)u_{i+1})) = \begin{cases} (u_{i+1},\ -\delta), & \text{if } 0 \le \delta \le 1/2 \\ (u_i,\ 1-\delta), & \text{if } 1/2 < \delta \le 1 \end{cases}$$

The 2-tuples present on the right side of the above equation are traditional 2-tuples as defined by Herrera & Martinez Herrera and Martínez (2000). h is a bijective function that can be used to convert traditional 2-tuples into proportional 2-tuples.

2.2 A zero-sum matrix game

A two-player zero-sum game is the most elementary one, comprising of two players only. This type of game is assessed by assuming that both the players apply strategies that could be pure, mixed, or continuous (Parthasarathy and Raghavan (1971)). We recall a zero-sum game according to Barron (2011).

Definition 6 Barron (2011) A crisp two-player zero-sum matrix game is represented as a triplet (S^n, S^m, A), where

(i) S^n is the set of 'n' possible strategies of player I.
(ii) S^m is the set of 'm' possible strategies of player II.
(iii) A is the payoff matrix of order $n \times m$ of player I corresponding to player II, and $neg(A) = -A$ refers to the matrix consisting of payoffs of player II.

Definition 7 Barron (2011) A matrix game having $A = (a_{ij})_{n\times m}$ as the matrix of payoffs has the lower value

$$v^- = \max_{i=1,\ldots,n} \min_{j=1,\ldots,m} a_{ij}$$

and the upper value

$$v^+ = \min_{j=1,\ldots,m} \max_{i=1,\ldots,n} a_{ij}$$

The lower value v^- is that minimum amount that player I is secured to get, and the upper value v^+ is that assured maximum amount that player II might lose. We say that the game has a value if the upper value equals the lower value, and we express it as $v^+ = v^- = v$, and the row and column $(i^*,\ j^*)$ giving payoff $a_{i^*j^*} = v^- = v^+$ are called optimal, or more precisely a saddle point in case the pure strategies

are applied. Further, $a_{ij^*} \leq a_{i^*j^*} \leq a_{i^*j}, \forall\, i = 1,\, 2, \ldots,\, n$ and $j = 1,\, 2, \ldots,\, m$.

Definition 8 Barron (2011) A mixed strategy is a vector $\Phi = (\phi_1,\, \phi_2, \ldots,\, \phi_n)$ for player I and $\Psi = (\psi_1,\, \psi_2, \ldots,\, \psi_m)$ for player II, where

$$\phi_i \geq 0,\ \sum_{i=1}^{n} \phi_i = 1 \quad \& \quad \psi_j \geq 0,\ \sum_{j=1}^{m} \psi_j = 1.$$

ϕ_i represents the probability of using row 'i' by player I, and ψ_j, the probability of using column 'j' by player II. We represent the set of mixed strategies having 'k' components as

$$S^k = \{(z_1,\, z_2, \ldots,\, z_k) |\, z_r \geq 0, r = 1,\, 2, \ldots,\, k,\ \sum_{r=1}^{k} z_r = 1\}.$$

In this terminology, elements $\Phi \in S^n$ and $\Psi \in S^m$ represent the mixed strategies for player I and player II, respectively.

Now, in case of mixed strategies, the expected payoff of players is defined as follows:

Definition 9 Barron (2011) Given a choice of mixed strategy $\Phi = (\phi_1,\, \phi_2, \ldots,\, \phi_n)$ for player I and $\Psi = (\psi_1,\, \psi_2, \ldots,\, \psi_m)$ for player II, selected autonomously, the expected payoff of player I is

$$\begin{aligned} E(\Phi, \Psi) &= \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}\text{Prob(I uses } i \text{ and II uses } j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}\text{Prob(I uses } i)\ \text{Prob(II uses } j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} \phi_i a_{ij} \psi_j \\ &= \Phi A \Psi^T \end{aligned}$$

In both classical and fuzzy environments, numerous studies have been conducted for solving matrix games. However, this setup has been extended further to the linguistic games overcoming the limitation of providing precise membership values to the payoffs. The linguistic setup in games has provided a more flexible structure for the decision-makers to evaluate the game problems. This linguistic information may come from a balanced as well as an unbalanced LT set. The next section discusses the matrix games based on the information underlying an unbalanced (proportional approach) LT set.

3 Two-player constant-sum matrix games based on proportional linguistic information

In this section, we majorly focus on the non-conventional case of matrix games in which the players are capable of expressing knowledge in terms of unbalanced linguistic labels. First, we introduce a methodology to define and solve a two-player constant-sum proportional matrix game with terms in symbolic proportions. Furthermore, we also propose conditions that show the existence of duality in such types of matrix games.

Definition 10 Consider a two-player constant-sum proportional game $\widetilde{G}$ as a quadruplet $(S^n,\, S^m,\, \overline{U},\, \widetilde{A})$ where

$$S^n = \{\Phi = (\phi_1,\, \phi_2, \ldots,\, \phi_n) |\, \phi_i \geq 0,\ i = 1,\, 2, \ldots,\, n,\ \sum_{i=1}^{n} \phi_i = 1\}$$

and

$$S^m = \{\Psi = (\psi_1,\, \psi_2, \ldots,\, \psi_m) |\, \psi_j \geq 0,\ j = 1,\, 2, \ldots,\, m,\ \sum_{j=1}^{m} \psi_j = 1\}$$

represent the strategy sets of player I and player II, respectively. The members of S^n (S^m) of the form $\Phi = (0,\, 0, \ldots,\, 1, \ldots,\, 0)^T$ where 1 is at the ith position ($\Psi = (0,\, 0, \ldots,\, 1, \ldots,\, 0)^T$ where 1 is at the jth position) are called pure strategies of Player I (Player II). $\overline{U}$ is the set of symbolic proportions as defined in Definition 1, $\widetilde{A} = [\widetilde{a}_{ij}]_{n\times m}$ being the proportional payoff matrix of player I corresponding to player II given as:

$$\widetilde{A} = \begin{pmatrix} \tilde{a}_{11} & \tilde{a}_{12} & \ldots & \tilde{a}_{1m} \\ \tilde{a}_{21} & \tilde{a}_{22} & \ldots & \tilde{a}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{a}_{n1} & \tilde{a}_{n2} & \ldots & \tilde{a}_{nm} \end{pmatrix}$$

where $\widetilde{a}_{ij} \in \overline{U}$ represents payoff in the form of proportional 2-tuple. Basically if Player I chooses i^{th} pure strategy and Player II chooses j^{th} pure strategy, then $\widetilde{a}_{ij}$ is the amount paid by Player II to Player I. Moreover, $neg(\widetilde{A})$ comprehends the payoff matrix of player II which shows the amount paid by Player I to Player II. From the matrix $\widetilde{A}$, it can clearly be understood that if player I & II opt for pure strategies i and j, respectively, then the payoff of player I would be $\widetilde{a}_{ij}$ and that of player II would be $neg(\widetilde{a}_{ij})$, i.e., if $\widetilde{a}_{ij} = (\delta u_k,\, (1-\delta)u_{k+1})$ is the payoff of player I for some k, then payoff of player II would be $neg(\widetilde{a}_{ij}) = ((1-\delta)u_{g-k-1},\, \delta u_{g-k})$, where $g+1$ is the cardinality of the LT set U.

We next define the values of player I and player II and the value of the game in terms of symbolic proportions. The meaning of the game's solution is best recognized in terms of maxmin and minmax principles for Player I and Player II, respectively. As per these principles, both the players choose that strategy, resulting in the best of the worst outcomes. Player I (the maximizing player) opts to play a analogous strategy to the maximum of minimum benefits for his different approach. This is called the maxmin principle. In similar terms, Player II (the minimizing player) also likes to play safe, and in that case, he opts for that strategy which is analogous to the minimum of maximum losses for his different approach. This is called the minmax principle. Implementing these principles, we obtain the following values of players.

Definition 11 Let us consider $\widetilde{v}^{1-}$ and $\widetilde{v}^{2-}$ as the values representing the minimum gain (in terms of symbolic proportion) as received by player I and player II, respectively. Then, these values are defined in the following manner:

$$\widetilde{v}^{1-} = \max_{i=1,\ldots,n} \min_{j=1,\ldots,m} \widetilde{a}_{ij}$$

$$\widetilde{v}^{2-} = \max_{j=1,\ldots,m} \min_{i=1,\ldots,n} neg(\widetilde{a}_{ij})$$

where $\widetilde{a}_{ij} \in \overline{U}$.

Note: The lower and upper values of the game are defined as $\widetilde{v}^{-} = \widetilde{v}^{1-}$ and $\widetilde{v}^{+} = neg(\widetilde{v}^{2-})$, respectively. Also,

$$\widetilde{v}^{+} = neg(\widetilde{v}^{2-}) = neg(\max_{j=1,\ldots,m} \min_{i=1,\ldots,n} neg(\widetilde{a}_{ij})) = \min_{j=1,\ldots,m} \max_{i=1,\ldots,n} (\widetilde{a}_{ij})$$

The game has value $\widetilde{v}$ if $\widetilde{v} = \widetilde{v}^{+} = \widetilde{v}^{-}$.

Remark 1 Since it can be proved that π is a linear operator and is a monotonic function, we can easily conclude that if

$$\widetilde{a} = (\delta u_i,\ (1-\delta)u_{i+1})$$
$$\widetilde{b} = (\gamma u_j,\ (1-\gamma)u_{j+1})$$

are two proportional linguistic variables (PLV's), then

(I) $neg(neg(\tilde{a})) = \tilde{a}$
(II) $min(neg(\tilde{a}),\ neg(\tilde{b})) = neg(max(\tilde{a},\ \tilde{b}))$
(III) $max(neg(\tilde{a}),\ neg(\tilde{b})) = neg(min(\tilde{a},\ \tilde{b}))$

Remark 2 The maximum and minimum of two PLV's may be evaluated with the help of a comparison of proportional 2-tuples as described in Definition 4. That is, for every $j = 0,\ 1,\ \ldots,\ g-1$, we define maximum of proportional 2-tuples as

$$\max\left((\delta u_i, (1-\delta)u_{i+1})\right) = (\gamma u_j, (1-\gamma)u_{j+1}) \iff (\delta u_i, (1-\delta)u_i + 1) \leq (\gamma u_j, (1-\gamma)u_{j+1})\ \forall\ i = 0,\ 1,\ \ldots,\ g-1.$$

and similarly,

$$\min\left((\delta u_i, (1-\delta)u_{i+1})\right) = (\gamma u_j, (1-\gamma)u_{j+1}) \iff (\delta u_i, (1-\delta)u_i + 1) \geq (\gamma u_j, (1-\gamma)u_{j+1})\ \forall\ i = 0,\ 1,\ \ldots,\ g-1.$$

Next, for the proportional linguistic values of the game defined in Definition 11, the following relation holds for any linguistic proportional game.

Theorem 1 *Suppose $\widetilde{v}^{-}$ and $\widetilde{v}^{+}$ be the proportional lower and upper values of the game, respectively. Then, $\widetilde{v}^{-}$ and $\widetilde{v}^{+}$ are PLV's and $\widetilde{v}^{-} \leq \widetilde{v}^{+}$.*

Proof Using Remark 2, it can easily be deduced that the minimum of two PLV's is further a PLV, and so is the maximum of minimum. Hence, max-min of any PLV comes out to be proportional.

$$\implies \widetilde{v}^{-} = \widetilde{v}^{1-} = \max_{i=1,\ldots,n} \min_{j=1,\ldots,m} \widetilde{a}_{ij} \text{ is a PLV.}$$

Similar argument follows for $\widetilde{v}^{+}$.
Now, suppose $\widetilde{a}_{ij} \in \overline{U}$. Then, for any column j and a fixed row i, $\min_{j=1,\ldots,m} \widetilde{a}_{ij} \leq \widetilde{a}_{ij}$ (using Comparison defined in Definition 4), thus taking maximum of both the sides of rows $i = 1,\ \ldots,\ n$, we get

$$\widetilde{v}^{-} = \max_i \min_j \widetilde{a}_{ij} \leq \max_i \widetilde{a}_{ij}$$

This holds for any column $j = 1,\ 2,\ \ldots,\ m$. This means that $\widetilde{v}^{-} \leq \min_j \max_i \widetilde{a}_{ij} = \widetilde{v}^{+}$, and hence the proof. □

Definition 12 We call a particular row i^* and column j^* a saddle point in pure strategies of the game if

$$\widetilde{a}_{ij^*} \leq \widetilde{a}_{i^*j^*} \leq \widetilde{a}_{i^*j}$$

$\forall$ rows $i = 1,\ 2,\ \ldots,\ n$ and columns $j = 1,\ 2,\ \ldots,\ m$.

Lemma 1 *A proportional linguistic matrix game will have a saddle point in pure strategies if and only if*

$$\widetilde{v}^{-} = \max_i \min_j \widetilde{a}_{ij} = \min_j \max_i \widetilde{a}_{ij} = \widetilde{v}^{+}.$$

Proof Using Definition 12,

$$\widetilde{v}^{+} = \min_{j} \max_{i} \widetilde{a}_{ij} \leq \max_{i} \widetilde{a}_{ij^*} \leq \widetilde{a}_{i^*j^*}$$
$$\leq \min_{j} \widetilde{a}_{i^*j} \leq \max_{i} \min_{j} \widetilde{a}_{ij} = \widetilde{v}^{-}$$

But $\widetilde{v}^{-} \leq \widetilde{v}^{+}$ (Theorem 1), so we exhibit equality all round and $\widetilde{v} = \widetilde{v}^{-} = \widetilde{v}^{+} = \widetilde{a}_{i^*j^*}$.
Further, if $\widetilde{v}^{-} = \widetilde{v}^{+}$, then

$$\min_{j} \max_{i} \widetilde{a}_{ij} = \max_{i} \min_{j} \widetilde{a}_{ij}$$

Suppose i^* be such that $\widetilde{v}^{+} = \max_j \widetilde{a}_{i^*j}$ and j^* be such that $\widetilde{v}^{-} = \min_i \widetilde{a}_{ij^*}$. Then,

$$\widetilde{a}_{i^*j} \leq \widetilde{v}^{+} = \widetilde{v}^{-} \leq \widetilde{a}_{ij^*}$$

for any $i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m$.

Moreover, taking $j = j^*$ on left and $i = i^*$ on the right, we get $\widetilde{a}_{i^*j^*} = \widetilde{v}^{+} = \widetilde{v}^{-}$. This suffices (i^*, j^*) to be the saddle point. □

Now, we illustrate the above concept using an example.

Example 1 Suppose a builder tries to sell its two buildings I and II to a buyer. Three criteria namely Price, Architecture, and Gentry are selected for the sale. The payoffs are in terms of proportions chosen from the LT set

$$U = \{u_0 : \text{Very Low},\ u_1 : \text{Low},\ u_2 : \text{Average},$$
$$u_3 : \text{High},\ u_4 : \text{Very High}\}.$$

The payoffs correspond to what the builder gets based on the above criterion. Let the payoff matrix of building I be:

$$\widetilde{A} = \begin{array}{c|ccc} \text{Builder/Buyer} & \text{Price} & \text{Architecture} & \text{Gentry} \\ \hline \text{Price} & (0.2u_2,\ 0.8u_3) & (0.2u_2,\ 0.8u_3) & (0.2u_2,\ 0.8u_3) \\ \text{Architecture} & (0.4u_0,\ 0.6u_1) & (0.7u_1,\ 0.3u_2) & (0.2u_2,\ 0.8u_3) \\ \text{Gentry} & (0.4u_0,\ 0.6u_1) & (0.1u_3,\ 0.9u_4) & (0.1u_3,\ 0.9u_4) \end{array}$$

This means that if the buyer purchases building I by seeing its price and building II based on its architecture, respectively, then the builder is assured to receive a payoff which is 20% Average and 80% High of the market share.
Moreover,

$$neg(\widetilde{A}) = \begin{array}{c|ccc} \text{Builder/Buyer} & \text{Price} & \text{Architecture} & \text{Gentry} \\ \hline \text{Price} & (0.8u_1,\ 0.2u_2) & (0.8u_1,\ 0.2u_2) & (0.8u_1,\ 0.2u_2) \\ \text{Architecture} & (0.6u_3,\ 0.4u_4) & (0.3u_2,\ 0.7u_3) & (0.8u_1,\ 0.2u_2) \\ \text{Gentry} & (0.6u_3,\ 0.4u_4) & (0.9u_0,\ 0.1u_1) & (0.9u_0,\ 0.1u_1) \end{array}$$

is the payoff matrix for building II.

The payoffs are interpreted in a way that if the buyer purchases building I based on Price and building II based on Architecture, respectively, then the builder is assured to receive a payoff which is 80% Low and 20% Average of the market share.

The lower value for building I is

$$\begin{aligned} \widetilde{v}^{1-} &= \max_{i=1,2,3} \min_{j=1,2,3} (\tilde{a}_{ij}) \\ &= \max [(0.2u_2,\ 0.8u_3), \\ &\quad (0.4u_0,\ 0.6u_1), (0.4u_0,\ 0.6u_1)] \\ &= (0.2u_2,\ 0.8u_3) \end{aligned}$$

Hence, the builder is assured of achieving at least a market share which is 20% Average and 80% High by selling building I.

In the similar manner, the lower value of building II is

$$\begin{aligned} \widetilde{v}^{2-} &= \max_{i=1,2,3} \min_{j=1,2,3} \left(neg(\tilde{a}_{ij})\right) \\ &= \max [(0.8u_1,\ 0.2u_2), (0.8u_1,\ 0.2u_2), \\ &\quad (0.9u_0,\ 0.1u_1)] \\ &= (0.8u_1,\ 0.2u_2) \end{aligned}$$

Hence, the builder will surely receive a payoff of at least 80% Low and 20% Average by selling building II.
Also, the lower value of the game is

$$\widetilde{v}^{-} = \widetilde{v}^{1-} = (0.2u_2,\ 0.8u_3)$$

and the upper value is

$$\widetilde{v}^{+} = neg(\widetilde{v}^{2-}) = (0.2u_2,\ 0.8u_3)$$

$\Longrightarrow$ (1, 1) is the saddle point and $(0.2u_2,\ 0.8u_3)$ is the value of the game.

Next, it is worth noting that it is not always possible to apply pure strategy in every case, or a saddle point always exists. For example, if we interchange $\widetilde{a}_{21}$ with $(0.9u_3,\ 0.1u_4)$ in the above matrix $\widetilde{A}$, then $\widetilde{v}^{-} = (0.2u_2,\ 0.8u_3)$ and $\widetilde{v}^{+} = (0.1u_3,\ 0.9u_4)$. So we have to apply the case where mixed strategies are involved. Hence, if the players propose mixed strategies to play the game, then their payoffs can be calculated by finding what is "expected" by the player to receive.

In the next section, we step forward to propose a Proportional Linguistic Linear Programming (PLLP) approach for solving the Proportional Matrix Game wherein players undergo mixed strategies (as well as pure strategies) to play the game.

4 Proportional linguistic linear programming (PLLP) formulation of a proportional matrix game

In the following section, we formulate Proportional Linguistic Linear Programming (PLLP) to solve a proportional matrix game (involving games using mixed strategies) by making use of the expected payoff to evaluate the payoffs of the players.

Let $U = \{u_0, u_1, \ldots, u_g\}$ be the LT set and $\overline{U} = \{(\delta u_i, (1-\delta)u_{i+1}) : \delta \in [0,1] \text{ and } i = 0, 1, \ldots, g-1\}$ be the set of proportions, g being a positive integer. Consider the proportional linguistic matrix $\widetilde{A} = \left(\widetilde{a}_{ij}\right)_{n\times m}$, where each $\widetilde{a}_{ij} \in \overline{U}$ represents the payoff to player I received after opting the strategy 'i' while player II opts for strategy 'j'. Let $\Phi = (\phi_1, \phi_2, \ldots, \phi_n) \in S^n$ and $\Psi = (\psi_1, \psi_2, \ldots, \psi_m) \in S^m$ be the mixed strategies of player I & II, respectively. Then, in a proportional matrix game having uncertain and unbalanced payoff entries, the proportional expected payoff of players may be defined as

$$\widetilde{E}_{\widetilde{A}}(\Phi, \Psi) = \bigoplus_{i=1}^{n} \left(\phi_i \left(\bigoplus_{j=1}^{m} \widetilde{a}_{ij}\psi_j\right)\right). \tag{1}$$

Thus, the expected payoff of player I corresponding to the j^{th} strategy chosen by player II is given by $\widetilde{a}_{1j}\phi_1 \oplus \widetilde{a}_{2j}\phi_2 \oplus \ldots \oplus \widetilde{a}_{nj}\phi_n$. This expression is the weighted average of all $\widetilde{a}_{ij}$ which is defined as below:

Let $\widetilde{a}_{ij} = (\delta u_k, (1-\delta)u_{k+1}) \in \overline{U}$, $k = 0, 1, \ldots, g-1$ be a proportional 2-tuple variable and $w = (w_1, w_2, \ldots, w_n)$ be the associated weight vector satisfying $0 \le w_i \le 1$, $i = 1, \ldots, n$, $\sum_{i=1}^{n} w_i = 1$.

Then, the proportional weighted average operator may be defined in the following manner:

$$\begin{aligned} PWA[\widetilde{a}_{ij},\ i = 1, \ldots, n] &= \widetilde{a}_{1j}w_1 \oplus \widetilde{a}_{2j}w_2 \oplus \ldots \oplus \widetilde{a}_{nj}w_n \\ &= \pi^{-1}\left(\sum_{i=1}^{n} w_i\, \pi\left(\widetilde{a}_{ij}\right)\right), \\ &\quad j = 1, \ldots, m \end{aligned}$$

Consequently,

$$\pi\left(\bigoplus_{i=1}^{n} \tilde{a}_{ij}w_i\right) = \sum_{i=1}^{n} w_i\, \pi\left(\tilde{a}_{ij}\right) \tag{2}$$

Note: Several operators are used to aggregate PLV's, e.g., CCV operator as defined in Wang and Hao (2006) by Wang and Hao. But usage of the CCV operator draws our attention toward the defuzzification of variables which is not our objective because of the information loss involved in the method. So we use the π function for such purpose because here we deal with the position indices of the variables.

Hence, we need to evaluate the mixed strategy $\Phi \in S^n$ of player I such that the value of player I, i.e., $\widetilde{v}^{1-}$ is maximized. So the required PLLP for player I is:

$$\max \quad \widetilde{v}^{1-} \tag{P1}$$
$$\text{s.t.} \begin{cases} \bigoplus_{i=1}^{n} \widetilde{a}_{ij}\phi_i \ge \widetilde{v}^{1-} \\ \text{where} \\ \phi_i \ge 0,\ \sum_{i=1}^{n} \phi_i = 1. \\ \widetilde{a}_{ij} \in \overline{U}. \end{cases}$$

As π is a monotonically increasing function, we apply π to each constraint of P1 and then use Equation 2.

Thus, the strategies and minimum gain of player I can be obtained by solving the following optimization problem:

$$\max \quad \widetilde{v}^{1-}$$
$$\text{s.t.} \begin{cases} \sum_{i=1}^{n} \pi\left(\widetilde{a}_{ij}\right)\phi_i \ge \pi(\widetilde{v}^{1-}) \\ \text{where} \\ \phi_i \ge 0, \sum_{i=1}^{n} \phi_i = 1. \\ \widetilde{a}_{ij} \in \overline{U}. \end{cases}$$

Substituting $\pi\left(\widetilde{a}_{ij}\right) = a_{ij}$, and $\pi\left(\widetilde{v}^{1-}\right) = v$. Here, it is clear that $\pi\left(\widetilde{v}^{1-}\right) = v > 0$.

So, constraints of P1 become

$$\sum_{i=1}^{n} a_{ij}\phi_i \ge v$$
$$\text{s.t. } \phi_i \ge 0,\ \sum_{i=1}^{n} \phi_i = 1.$$

Set $X_i = \dfrac{\phi_i}{v};\ i = 1, 2, \ldots, n$ and $\dfrac{1}{v} = V$. This is summarized as a linear program:

$$\min \quad V = \sum_{i=1}^{n} X_i \tag{P2}$$
$$\text{s.t.} \begin{cases} \sum_{i=1}^{n} a_{ij}X_i \ge 1 \\ \text{where} \\ X_i \ge 0. \end{cases}$$

and then, we can solve P2 as a usual linear programming problem.

It must be noted that in all the above constraints and expressions, the variable i varies as $i = 1, 2, \ldots, n$ and the variable j varies as $j = 1, 2, \ldots, m$.

Similarly, the PLLP for player II would be

$$\max \quad \widetilde{v}^{2-} \tag{P3}$$
$$\text{s.t.} \begin{cases} \bigoplus_{j=1}^{m} neg\left(\widetilde{a}_{ij}\right)\psi_j \geq \widetilde{v}^{2-} \\ \text{where} \\ \psi_j \geq 0,\ \sum_{j=1}^{m}\psi_j = 1. \\ \widetilde{a}_{ij} \in \overline{U}. \end{cases}$$

We have to find the mixed strategy $\Psi \in S^m$ for player II. Let $neg(\widetilde{v}^{2-}) = \widetilde{v}^{+}$.

$\Longrightarrow$ From the properties of negation operator, the problem becomes

$$\min \quad \widetilde{v}^{+} \tag{P4}$$
$$\text{s.t.} \begin{cases} \bigoplus_{j=1}^{m} \widetilde{a}_{ij}\psi_j \leq \widetilde{v}^{+} \\ \text{where} \\ \psi_j \geq 0,\ \sum_{j=1}^{m}\psi_j = 1. \\ \widetilde{a}_{ij} \in \overline{U}. \end{cases}$$

Now, we apply π operator in P4 and use the expression in Equation 2.

Since earlier $\pi\left(\widetilde{v}^{1-}\right) = v$ and if v is the value of the game, then $\pi\left(\widetilde{v}^{+}\right) = v$.

So the problem for player II becomes

$$\min \quad \widetilde{v}^{+}$$
$$\text{s.t.} \begin{cases} \sum_{j=1}^{m}\pi\left(\widetilde{a}_{ij}\right)\psi_j \leq \pi(\widetilde{v}^{+}) \\ \text{where} \\ \psi_j \geq 0,\ \sum_{j=1}^{m}\psi_j = 1. \\ \widetilde{a}_{ij} \in \overline{U}. \end{cases}$$

Put $Y_j = \frac{\psi_j}{v};\ j = 1, 2, \ldots, m$ and $\pi\left(\widetilde{a}_{ij}\right) = a_{ij}$, the problem for player II reduces to the problem

$$\max \quad V = \sum_{j=1}^{m} Y_j \tag{P5}$$
$$\text{s.t.} \begin{cases} \sum_{j=1}^{m} a_{ij}Y_j \leq 1 \\ \text{where} \\ Y_j \geq 0. \end{cases}$$

So, by solving P2 and P5, we get the optimal solution $X = (X_1, X_2, \ldots, X_n)$ and $Y = (Y_1, Y_2, \ldots, Y_m)$ and the optimal value V. Hence, unwinding the formulation back to our original variables, the optimal strategies Φ and Ψ of player I and player II, respectively, could be achieved and the value v of the game follows. Also, the linguistic proportional values for player I and player II would be

$$\widetilde{v}^{-} = \widetilde{v}^{1-} = \pi^{-1}\left(\frac{1}{V}\right)$$

and,

$$\widetilde{v}^{2-} = neg\left(\pi^{-1}\left(\frac{1}{V}\right)\right)$$

respectively.

Again, the variables i & j vary as $i = 1, 2, \ldots, n$ and $j = 1, 2 \ldots, m$ in all of the above constraints.

Note: It is to be noted that applying the π function directly in the payoff matrix to convert payoffs into numbers must not be a good idea because the value of α is different for each cell. Instead, to do so, we apply the function π on both sides of the PLLP (π being a monotonic function, this is justifiable).

Remark 3 Note that the linear programming problems P2 and P5 obtained by using a transformation function π are particular crisp problems corresponding to given proportional linguistic 2-tuples and form primal-dual pair of each other. Since it is envisaged that the translation function π is an isomorphism, henceforth, there exists a similarity relation between the numerical and symbolic proportional linguistic spaces, i.e., $\overline{U} \cong [0,\ g]$ and both are structurally similar sets. As the duality result holds for the crisp LPP, therefore, it follows directly for PLLP as well.

Example 2 Consider a 3×3 proportional payoff matrix for a constant sum game whose elements belong to the set

$$U = \{u_0 : \text{Very Poor},\ u_1 : \text{Poor},\ u_2 : \text{Average},\ u_3 : \text{Good},\ u_4 : \text{Very Good}\}$$

given as

$$\widetilde{A} = \begin{bmatrix} (0.2u_0,\ 0.8u_1) & (0.3u_1,\ 0.7u_2) & (0.3u_1,\ 0.7u_2) \\ (0.3u_1,\ 0.7u_2) & (0.4u_2,\ 0.6u_3) & (0.9u_3,\ 0.1u_4) \\ (0.2u_0,\ 0.8u_1) & (0.2u_0,\ 0.8u_1) & (0.4u_2,\ 0.6u_3) \end{bmatrix}$$

Let $\Phi = (\phi_1, \phi_2, \phi_3), \phi_i \geq 0$ for $i = 1, 2, 3$ such that $\sum_{i=1}^{3}\phi_i = 1$, and $\Psi = (\psi_1, \psi_2, \psi_3), \psi_j \geq 0, j = 1, 2, 3$ such that $\sum_{j=1}^{3}\psi_j = 1$ be the mixed strategies for player I and player II, respectively. We formulate the PLLP for player I as in P1 in following manner:

$$\max \quad \widetilde{v}^{-}$$

$$\text{s.t.} \begin{cases} (0.2u_0, 0.8u_1)\phi_1 \oplus (0.3u_1, 0.7u_2)\phi_2 \oplus (0.2u_0, 0.8u_1)\phi_3 \geq \widetilde{v}^- \\ (0.3u_1, 0.7u_2)\phi_1 \oplus (0.4u_2, 0.6u_3)\phi_2 \oplus (0.2u_0, 0.8u_1)\phi_3 \geq \widetilde{v}^- \\ (0.3u_1, 0.7u_2)\phi_1 \oplus (0.9u_3, 0.1u_4)\phi_2 \oplus (0.4u_2, 0.6u_3)\phi_3 \geq \widetilde{v}^- \\ \phi_1 + \phi_2 + \phi_3 = 1 \\ \phi_1,\ \phi_2,\ \phi_3 \geq 0. \end{cases}$$

Applying π function in the constraints, we get

$$\begin{aligned} &0.8\phi_1 + 1.7\phi_2 + 0.8\phi_3 \geq \pi(\widetilde{v}^-) \\ &1.7\phi_1 + 2.6\phi_2 + 0.8\phi_3 \geq \pi(\widetilde{v}^-) \\ &1.7\phi_1 + 3.1\phi_2 + 2.6\phi_3 \geq \pi(\widetilde{v}^-) \\ &\phi_1 + \phi_2 + \phi_3 = 1 \\ &\phi_1,\ \phi_2,\ \phi_3 \geq 0. \end{aligned}$$

If $\pi(\widetilde{v}^-) = v$, substituting $X_i = \dfrac{\phi_i}{v}$ and $V = \dfrac{1}{v}$, the problem reduces in the similar manner as P2 like

$$\min \quad V = X_1 + X_2 + X_3$$
$$\text{s.t.} \begin{cases} 0.8X_1 + 1.7X_2 + 0.8X_3 \geq 1 \\ 1.7X_1 + 2.6X_2 + 0.8X_3 \geq 1 \\ 1.7X_1 + 3.1X_2 + 2.6X_3 \geq 1 \\ X_1,\ X_2,\ X_3 \geq 0. \end{cases}$$

The optimal solution of this problem is $(X_1,\ X_2,\ X_3) = (0,\ 0.59,\ 0)$ and $V = 0.59 \implies (\phi_1,\ \phi_2,\ \phi_3) = (0,\ 1,\ 0)$ and $v = 1.6949$. So, $\widetilde{v}^- = \pi^{-1}(v) = \pi^{-1}(1.6949) = (0.3051u_1, 0.6949u_2)$. Hence, the optimal strategy for player I is $\Phi = (0,\ 1,\ 0)$ and the value is $\widetilde{v}^- = (0.3051u_1,\ 0.6949u_2)$.
Similarly, the PLLP for player II as in P4 is

$$\min \quad \widetilde{v}^+$$
$$\text{s.t.} \begin{cases} (0.2u_0, 0.8u_1)\psi_1 \oplus (0.3u_1, 0.7u_2)\psi_2 \oplus (0.3u_1, 0.7u_2)\psi_3 \leq \widetilde{v}^+ \\ (0.3u_1, 0.7u_2)\psi_1 \oplus (0.4u_2, 0.6u_3)\psi_2 \oplus (0.9u_3, 0.1u_4)\psi_3 \leq \widetilde{v}^+ \\ (0.2u_0, 0.8u_1)\psi_1 \oplus (0.2u_0, 0.8u_1)\psi_2 \oplus (0.4u_2, 0.6u_3)\psi_3 \leq \widetilde{v}^+ \\ \psi_1 + \psi_2 + \psi_3 = 1 \\ \psi_1,\ \psi_2,\ \psi_3 \geq 0. \end{cases}$$

By substitution as given in P5 , the problem becomes

$$\max \quad V = Y_1 + Y_2 + Y_3$$
$$\text{s.t.} \begin{cases} 0.8Y_1 + 1.7Y_2 + 1.7Y_3 \leq 1 \\ 1.7Y_1 + 2.6Y_2 + 3.1Y_3 \leq 1 \\ 0.8Y_1 + 0.8Y_2 + 2.6Y_3 \leq 1 \\ Y_1,\ Y_2,\ Y_3 \geq 0. \end{cases}$$

The optimal solution is $(Y_1,\ Y_2,\ Y_3) = (0.59,\ 0,\ 0)$ and $V = 0.59 \implies (\psi_1,\ \psi_2,\ \psi_3) = (1,\ 0,\ 0)$ and $v = 1.6949$. So, $\widetilde{v}^+ = \pi^{-1}(v) = \pi^{-1}(1.6949) = (0.3051u_1,\ 0.6949u_2)$.
$\therefore$ the optimal strategy used by player II is $\Psi = (1,\ 0,\ 0)$ and optimal value is $\widetilde{v}^+ = (0.3051u_1,\ 0.6949u_2) = \widetilde{v}$.
It is noteworthy that $\widetilde{v}^- = \widetilde{v}^+ = (0.3051u_1,\ 0.6949u_2) = \widetilde{v}$ is the value of the game.

For a better understanding of the process of evaluating a proportional linguistic matrix game, one can refer to Fig. 3.

In the subsequent section, we apply the proposed proportional linguistic game-theoretic approach in solving a linguistic MCDM problem to validate the proposed methodology in a physical realm.

5 Proportional matrix game for MCDM problems

Decision-making is a necessary process to achieve the best possible solution for a problem as per the information provided by the decision-makers. For example, in an MCDM problem, there are different alternatives (A_i) which are to be chosen based on some criterion (C_j) using the previously defined linguistic terms set LT. As explained by Chen and Larbani (2006), the aim of the decision-maker (DM) is to pick the best alternative from the available data in the matrix with fuzzy payoffs. The decision maker (DM) and the Nature act as two players of the game called player I and player II, respectively, having strategies $\Phi = (\phi_1,\ \phi_2, \ldots,\ \phi_n)$ and $\Psi = (\psi_1,\ \psi_2, \ldots,\ \psi_m)$, where $\sum_{i=1}^{n} \phi_i = 1$, $\sum_{j=1}^{m} \psi_j = 1$ and $\phi_i,\ \psi_j \geq 0$. Hence, a two-player zero-sum game is obtained in which $\widetilde{a}_{ij}$ is treated as the payoff obtained by the DM after choosing the i^{th} strategy having probability ϕ_i and the Nature chooses the j^{th} strategy having probability ψ_j. This game problem is then solved using some models, and the strategies of both the players are evaluated along with the value of the game.
In the sequence, Singh et al. (2018) illustrated the applications of the linguistic matrix game approach in the MCDM problem. They aim to select the most appropriate alternative out of all based on some attributes. Analogously, we present the following example to solve an MCDM problem based on the proposed proportional linguistic matrix game approach.

Example 3 The pharmaceutical company Abbott is willing to understand the perspective of doctors in the "first-line treatment" of constipation. The study is conducted on different laxatives, namely

A_1: Milk of Magnesia (MoM)
A_2: Lactulose
A_3: Isphagula
A_4: PEG

The decision of doctors is based on the following attributes:

C_1: Safest Laxative

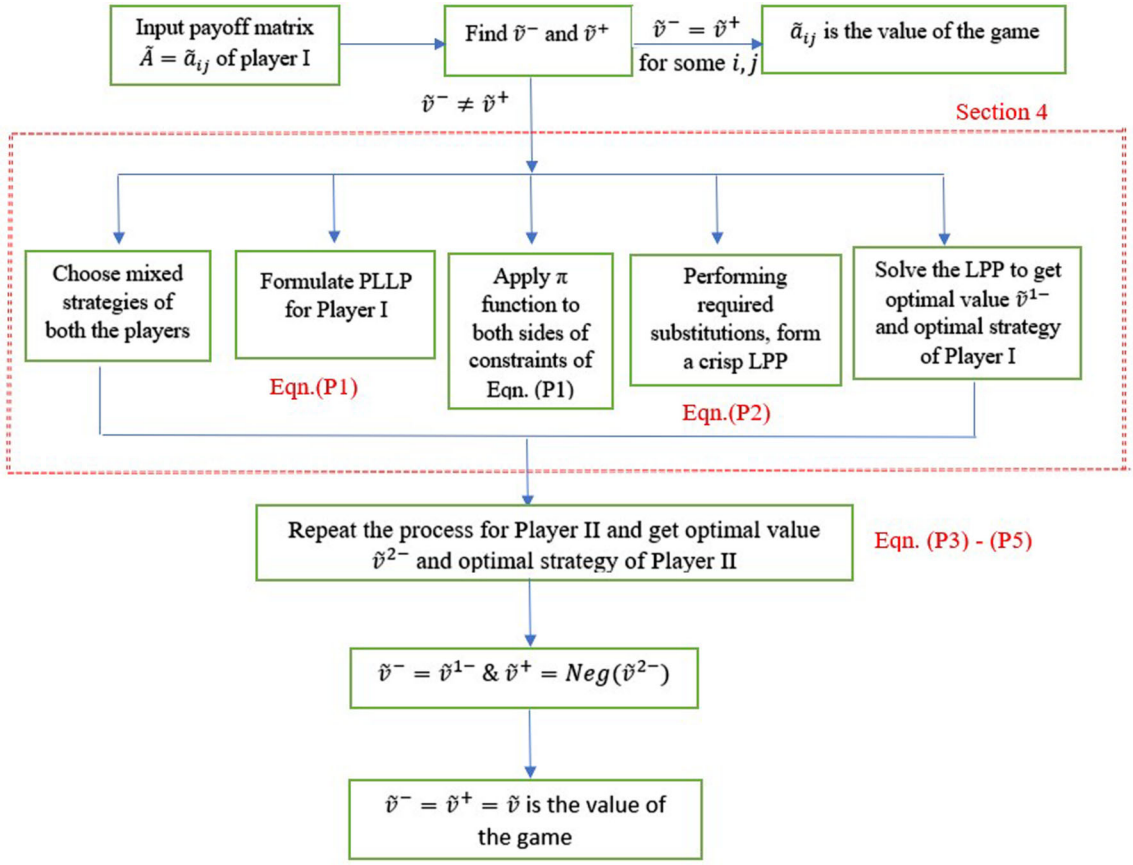


Fig. 3 The methodology to evaluate Proportional Linguistic Matrix Game problem

C_2: Fast Acting
C_3: Provides Overnight Relief
C_4: Provides Long-lasting Relief
C_5: Cost Effective
C_6: Is a mild laxative
C_7: Suitable for Long Term Usage
C_8: Backed by Strong Clinical Evidence
C_9: Non-habit forming

The doctors provide the response to each alternative based on given criteria using a predefined linguistic term set:
$U = \{u_0$: Not Effective, u_1 : Less Effective, u_2 : Average Effective, u_3 : Very Effective, u_4 : Most Effective$\}$
The following matrix having proportional linguistic payoffs is for the game problem between alternatives and criterion:

$$\tilde{A} = \begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \\ A_4 \end{array} \begin{array}{c} \begin{array}{ccccccccc} C_1 & C_2 & C_3 & C_4 & C_5 & C_6 & C_7 & C_8 & C_9 \end{array} \\ \begin{bmatrix} (u_0, 0) & (0.35u_0, 0.65u_1) & (0.71u_0, 0.29u_1) & (u_0, 0) & (0.76u_0, 0.24u_1) & (0.76u_0, 0.24u_1) & (u_0, 0) & (0.79u_0, 0.21u_1) & (u_0, 0) \\ (0.44u_1, 0.56u_2) & (0.61u_2, 0.39u_3) & (0.43u_1, 0.57u_2) & (0.35u_2, 0.65u_3) & (0.42u_1, 0.58u_2) & (0.77u_0, 0.23u_1) & (0.56u_1, 0.44u_2) & (0.66u_2, 0.34u_3) & (0.58u_1, 0.42u_2) \\ (0, u_4) & (u_0, 0) & (0.33u_0, 0.67u_1) & (0.56u_1, 0.44u_2) & (0.24u_2, 0.76u_3) & (0.22u_2, 0.78u_3) & (0.22u_2, 0.78u_3) & (0.54u_1, 0.46u_2) & (0.25u_2, 0.75u_3) \\ (u_0, 0) & (0.53u_1, 0.47u_2) & (0.37u_0, 0.63u_1) & (0.52u_1, 0.48u_2) & (0.8u_0, 0.2u_1) & (u_0, 0) & (u_0, 0) & (0.42u_1, 0.58u_2) & (u_0, 0) \end{bmatrix} \end{array}$$

The payoffs are described so that the payoff $(0.35u_0, 0.65u_1)$ depicts that 35% of the doctors consider the laxative as non-effective while 65% doctors find it effective, but some less.

In accordance with P2, the corresponding linear programming problem for alternatives would be

$$\min \quad V = X_1 + X_2 + X_3 + X_4$$

$$\text{s.t.} \begin{cases} 0X_1 + 1.56X_2 + 4X_3 + 0X_4 \geq 1 \\ 0.65X_1 + 2.39X_2 + 0X_3 + 1.47X_4 \geq 1 \\ 0.29X_1 + 1.57X_2 + 0.67X_3 + 0.63X_4 \geq 1 \\ 0X_1 + 2.65X_2 + 1.44X_3 + 1.48X_4 \geq 1 \\ 0.24X_1 + 1.58X_2 + 2.76X_3 + 0.2X_4 \geq 1 \\ 0.24X_1 + 0.23X_2 + 2.78X_3 + 0X_4 \geq 1 \\ 0X_1 + 1.44X_2 + 2.78X_3 + 0X_4 \geq 1 \\ 0.21X_1 + 2.34X_2 + 1.46X_3 + 1.58X_4 \geq 1 \\ 0X_1 + 1.42X_2 + 2.75X_3 + 0X_4 \geq 1 \\ X_1,\ X_2,\ X_3,\ X_4 \geq 0. \end{cases}$$

The optimal solution of above LPP is $(0, 0.5, 0.32, 0)$ & V = 0.82, and hence by implementing the required computations, the optimal strategy of player I comes out to be $(\phi_1, \phi_2, \phi_3, \phi_4) = (0, 0.6095, 0.39008, 0) \approx (0, 0.6, 0.4, 0)$. Moreover, $v = 1.219$ and hence $\widetilde{v}^- = (0.781u_1, 0.219u_2)$.

Similarly, for attributes, the LPP formulated using P5 is

$$\max \quad V = Y_1 + Y_2 + \ldots + Y_9$$

$$\text{s.t.} \begin{cases} 0Y_1 + 0.65Y_2 + 0.29Y_3 + 0Y_4 + 0.24Y_5 + 0.24Y_6 + 0Y_7 + 0.21Y_8 + 0Y_9 \leq 1 \\ 1.56Y_1 + 2.39Y_2 + 1.57Y_3 + 2.65Y_4 + 1.58Y_5 + 0.23Y_6 + 1.44Y_7 + 2.34Y_8 + 1.42Y_9 \leq 1 \\ 4Y_1 + 0Y_2 + 0.67Y_3 + 1.44Y_4 + 2.76Y_5 + 2.78Y_6 + 2.78Y_7 + 1.46Y_8 + 2.75Y_9 \leq 1 \\ 0Y_1 + 1.47Y_2 + 0.63Y_3 + 1.48Y_4 + 0.2Y_5 + 0Y_6 + 0Y_7 + 1.58Y_8 + 0Y_9 \leq 1 \\ Y_1,\ Y_2, \ldots,\ Y_9 \geq 0. \end{cases}$$

The solution obtained from this LPP is (0, 0, 0.61, 0, 0, 0.21, 0, 0, 0) & $V = 0.82$.

$\therefore$ The optimal strategy used by player II comes out to be $(\psi_1, \psi_2, \ldots, \psi_9) = (0, 0, 0.7439, 0, 0, 0.2559, 0, 0, 0)$. Also, $v = 1.219$. So, $\widetilde{v}^+ = (0.781u_1, 0.219u_2)$.

$\Longrightarrow$ $\widetilde{v}^- = \widetilde{v}^+ = \widetilde{v} = (0.781u_1, 0.219u_2)$ is the value of the game.

This means that based on overnight relief and the power of laxative, 78.1% doctors find the laxative "Lactulose" followed by "Isphagula" as effective (but less) and only 21.9% doctors find it not effective out of all other laxatives in the first line treatment of constipation.

6 Analysis of the work done

This section presents the analysis of the work done in the paper. Section 6.1 offers the analysis of the results obtained in the previous sections, and Sect. 6.2 compares the defined methodology with the existing methodologies. In Sect. 6.3, we discuss the computation complexity of the proposed methodology.

6.1 Result analysis

In this subsection, we discuss the results of the game's value and the players' strategies obtained in earlier sections.

Lemma 1 in Sect. 3 describes the case of having saddle point if the players in the game apply the pure strategies. Example 1 depicts the existence of saddle point in case pure strategies are played as $\widetilde{v}^-$ comes out to be equal to $\widetilde{v}^+$. Hence, the value of the game may be obtained as $\widetilde{v} = \widetilde{v}^- = \widetilde{v}^+$ and the row (column) for which value of the game is obtained is called the optimal pure strategy of player I (player II). The same may be evaluated for player II by applying this procedure on the negation of the game matrix corresponding to player I.

In case the pure strategies cannot be applied, the players use mixed strategies to evaluate the value of the game. This occurs when $\widetilde{v}^- \neq \widetilde{v}^+$, which is depicted in Example 2 of Sect. 4. For such a case, a PLLP is formulated for player I in equation P1 which consists of proportional linguistic payoffs. These proportional linguistic payoffs maybe then converted into numerical terms by doing required substitutions. By doing so, the original PLLP may be converted into a usual crisp LPP that can be solved directly. The solution of the LPP maybe then re-converted into linguistic form by reverting the substitutions. This gives the values of both the players and the value of the game, and the players' strategies

may be found. The same process is then repeated for player II whose PLLP is as given in equation P3. Equal values of both the players provide the value of the game.

6.2 Comparative analysis

This subsection compares the proposed study with the existing theories of matrix games having payoffs from an uncertain domain.

Many theories deal with matrix games having fuzzy payoffs, e.g., Bector et al. (2004a, b); Bector and Chandra (2005); Campos (1989); Dengfeng (1999); Vijay et al. (2005a); Zaichenko (2020). These theories discuss matrix games whose payoffs belong to a fuzzy set. Nevertheless, these theories have an associated limitation that the payoffs may take the form of linguistic variables as well. Hence, matrix games are important to study in the realm of linguistic term sets also. Researchers have worked in the field of linguistic matrix games and have found significant results Arfi (2006a, b); Sahoo (2017). However, in these methods, the linguistic payoffs are defuzzified to convert them into numeric LPPs. Defuzzification of payoffs causes the loss of information. The loss of information of payoffs is handled well by Singh et al. (2018); Singh and Gupta (2018) and Malhotra et al. (2019) where the payoffs belong to a balanced LT set, and these payoffs are converted into crisp values using a bijective function. But these theories are limited to a balanced LT set only. The theory defined in this paper is believed to be the first study considering matrix games having payoffs from an unbalanced LT set. The payoffs belong to a proportional LT set whose proportions belong to the interval [0, 1]. Moreover, the payoffs belonging to balanced LT sets may also be converted to their corresponding proportional ones by using the inverse of the h function defined in Definition 5. Note that this can only be done for a particular δ because δ is different for each cell. Hence, the matrix games with 2-tuple linguistic payoffs as defined by Singh et al. (2018) may also be solved by considering it as a particular case of the methodology defined in this paper.

6.3 Computation complexity

This subsection discusses the computation complexity involved in the defined methodology of unbalanced linguistic matrix games.

If we consider a game problem as a decision-making problem, then rather than picking random weights, the weights in this method are calculated within the game problem itself (considering that the mixed strategies play the role of weights). This makes the computation of this approach more appropriate than the existing decision-making methodologies. Moreover, the PLLP of Players I and II is first converted into usual crisp LPPs that are easy and fast to compute (manually or by using any optimization software), and the result is then re-translated into the linguistic one using bijective function. Thus, the defined methodology considers the complexity in solving game problems and computes them without making any "loss of information."

7 Concluding remarks

The payoffs based on symmetrical linguistic term sets have been successfully applied in many game problems to date. However, in a more realistic viewpoint, the payoffs of a matrix game may not always belong to an LT set where the terms are equidistant. Overcoming this limitation, we consider matrix games having payoffs belonging to an asymmetrical LT set in our paper. These payoffs are expressed in ordinal linguistic information via 2-tuples, comprising of two proportional linguistic terms with coefficients in [0, 1]. This representation of payoffs is more advantageous as it assesses the payoffs consisting of linguistic terms that belong to a continuous scale. To compute such game problems, first, the case of pure strategies is discussed where the existence of saddle points is shown. Further, when pure strategies cannot be applied, the application of mixed strategies is discussed. For this, a Proportional Linguistic Linear Programming (PLLP) approach is introduced to obtain the optimal strategies for players I and II. The value of such a game, where the conventional requirement of linguistic terms being equidistant may not be required, is also discussed. This approach is then illustrated with the help of some examples.

Lastly, we develop an application of our approach for solving the linguistic MCDM problem to demonstrate its validity in a practical sense. The suggested approach provides an integrated structure as it can be accepted to solve unbalanced linguistic matrix games and PLLP problems, and MCDM problems.

Although our approach is advantageous in solving matrix games where the data are asymmetrical or unbalanced, giving rise to unbalanced linguistic matrix games, it has a limitation associated with it. Concerning the proportional linguistic term set, the symbolic proportions are allocated only to two consecutive terms of a linguistic term set. So, in the future, we look forward to extending our methodology to solve unbalanced linguistic matrix games overcoming the current limitation and enhancing the applicability of linguistic matrix games in this ambiguous world. Further, we anticipate that this methodology may be extended to solve n-person games and can be applied to evaluate the results in cooperative game theory.

Author Contributions Parul Chauhan is a research scholar in the Department of Applied Mathematics, Delhi Technological University under the supervision of Anjana Gupta. Parul Chauhan has designed

the algorithm and prepared the manuscript in the present form with the guidance of Anjana Gupta.

Funding This research received no external funding.

Declarations

Conflict of Interest 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.

References

Abchir MA, Truck I (2013) Towards an extension of the 2-tuple linguistic model to deal with unbalanced linguistic term sets. arXiv:1304.5897v1 : 1-17

Alonso S, Pérez IJ, Herrera-Viedma E (2008) A consensus model for group decision making unbalanced fuzzy linguistic contexts. XIV Congreso Español sobre Tecnologías y Lógica fuzzy : 449-456

Arfi B (2006) Linguistic fuzzy-logic game theory. J Conf Resol 50(1):28–57

Arfi B (2006) Linguistic fuzzy-logic social game of cooperation. Rational Soc 18(4):471–537

Barron EN (2011) Game theory: an introduction. Wiley, New Jersey

Bartczuk L, Dziwiński P, Starczewski JT (2012) A new method for dealing with unbalanced linguistic term set. ICAISC 2012 Part I. LNCS 7267:207–212

Bauso D (2016) Game theory with engineering applications. Advances in design and control, philadelphia: SIAM

Bector CR, Chandra S (2005) Fuzzy mathematical programming and fuzzy matrix games. Springer, Berlin

Bector CR, Chandra S, Vijay V (2004) Duality in linear programming with fuzzy parameters and matrix games with fuzzy payoffs. Fuzzy sets and systems 146:253–269

Bector CR, Chandra S, Vijay V (2004) Matrix games with fuzzy goals and fuzzy linear programming duality. Fuzzy optimization and decision making 3:255–269

Benz A, Jåger G, Rooij RV (2006) Game theory and pragmatics. Palgrave Studies in Pragmatics, Language and Cognition, pp 1–82

Butnariu D (1978) Fuzzy games: a description of the concept. Fuzzy Set Syst 1:181–92

Cai M, Gong Z (2017) Group decision making using distances between unbalanced linguistic assessments. Appl Soft Comput 67:613–624

Cai M, Gong Z, Yu X (2016) A method for unbalanced linguistic term sets and its application in group decision making. Int J Fuzzy Syst 19(3):671–682

Cai X, Zhao H, Shang S, Zhou Y, Deng W, Chen H, Deng W (2021) An improved quantum-inspired cooperative co-evolution algorithm with muli-strategy and its application. Expert Syst Appl 171. https://doi.org/10.1016/j.eswa.2021.114629

Campos L (1989) Fuzzy linear programming models to solve fuzzy matrix games. Fuzzy Sets Syst 32:275–289

Charilas DE, Panagopoulos AD (2010) A survey on game theory applications in wireless networks. Comput Netw 54(18):3421–3430

Chen YW, Larbani M (2006) 2-person zero-sum game approach for fuzzy multiple attribute decision making problems. Fuzzy Sets Syst 157:34–51

Delgado M, Verdegay JL, Vila MA (1990) Playing matrix games defined by linguistic labels. Multiperson decision making models using fuzzy sets and possibility theory, theory and decision library (Series B: Mathematical and Statistical Methods) 18 Springer, Dordrecht

Dengfeng L (1999) A fuzzy multi-objective programming approach to solve fuzzy matrix games. J Fuzzy Math 7(4):907–912

Deng W, Shang S, Cai X, Zhao H et al (2021a) Quantum differential evolution with cooperative coevolution framework and hybrid mutation strategy for large scale optimization. Knowledge-Based Systems 224(107080):1–14. https://doi.org/10.1016/j.knosys.2021.107080

Deng W, Shang S, Cai X, Zhao H, Song Y, Xu J (2021) An improved differential evolution algorithm and its application in optimization problem. Soft Comput. https://doi.org/10.1007/s00500-020-05527-x

Deng W, Xu J, Gao XZ, Zhao H (2020) An enhanced MSIQDE algorithm with novel multiple strategies for global optimization problems. IEEE Trans Syst Man Cybernet Syst. https://doi.org/10.1109/TSMC.2020.3030792

Deng W, Xu J, Zhao H, Song Y (2020) A novel gate resource allocation method using improved PSO-based QEA. IEEE Trans Intell Transp Syst. https://doi.org/10.1109/TITS.2020.3025796

Dong YC, Xu YF, Yu S (2009) Computing the numerical scale of the linguistic term set for the 2-tuple fuzzy linguistic representation model. IEEE Trans. Fuzzy Syst. 17:1366–1378

Herrera F, Martínez L (2000) A 2-tuple fuzzy linguistic representation model for computing with words. IEEE Trans Fuzzy Syst 8(6):746–752

Herrera F, Martínez L (2001) The 2-tuple linguistic computational model, advantages of its linguistic description, accuracy and consistency. Int J Uncert Fuzz Knowl Based Syst 09:33–48

Herrera F, Martínez L (2001) A model based on linguistic 2-tuples for dealing with multigranular hierarchical linguistic contexts in multi-expert decision making. IEEE Trans. Syst. Man. Cybern. B 31:227–234

Herrera F, Viedma EH, Martínez L (2008) A fuzzy linguistic methodology to deal with unbalanced linguistic term sets. IEEE Trans Fuzzy Syst 16(2):354–370

Jaeger G, Applications of game theory in linguistics. Lang Linguist Comp 2(3) : 406–421

Jiang YP, Fan ZP, Ma J (2008) A method for group decision making with multigranularity linguistic assessment information. Inf. Sci. 178:1098–1109

Malhotra T, Gupta A, Singh A (2019) Methodology for interval-valued matrix games with 2-tuple fuzzy linguistic information. International Conference on Numerical Computations: Theory and Algorithms Springer, Cham : 154-168

Mazalov V (2014) Mathematical game theory and applications. Wiley, New Jersey

Neumann JV, Morgenstern O (1944) Theory of games and economic behavior. Princeton University Press, New York

Parthasarathy T, Raghavan TES (1971) Some topics in two-person games. American Elsevier Pub. Co., New York

Pei Z, Zheng L (2016) New unbalanced linguistic scale sets: the linguistic information representations and applications. Comput Indus Eng 105:377–390

Sahoo L (2017) Solving matrix games with linguistic payoffs. Int J Syst Assur Eng Manag 10(4):484–490

Singh H (1999) Introduction to game theory and its application in electric power markets. IEEE Comput Appl Power 12(4):18–20

Singh A, Gupta A (2018) Matrix games with interval-valued 2-tuple linguistic information. Games 9(3):62

Singh A, Gupta A, Mehra A (2018) Matrix games with 2-tuple linguistic information. Ann Oper Res 287:895–910

Song Y, Wu D, Mohamed AW, Zhou X, Zhang B, Deng W (2021) Enhanced success history adaptive de for parameter optimization of photovoltaic models. Complexity 6660115:1–22. https://doi.org/10.1155/2021/6660115

Straffin PD Jr (2010) Game theory and strategy. Math Assoc Am 36
Vijay V, Chandra S, Bector CR (2005) Matrix games with fuzzy goals and fuzzy payoffs. Omega 33(5):425–429
Vijay V, Mehra A, Chandra S, Bector CR (2005) Fuzzy matrix games via a fuzzy relation approach. Fuzzy Opt Dec Mak 6:299–314
Wang JH, Hao J (2006) a new version of 2-tuple fuzzy linguistic representation model for computing with words. IEEE Trans Fuzzy Syst 14(3):435–445
Wang B, Liang J, Qian Y (2015) A Normalized numerical scaling method for the unbalanced multi-granular linguistic sets. Int J Uncert Fuzz Knowl Based Syst 23(2):221–243
Zadeh LA (1974) The concept of a linguistic variable and its application to approximate reasoning. Learning Systems and Intelligent Robots : 1-10, Boston
Zadeh LA (1965) Fuzzy Sets. Inf Control 8:338–353
Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning- II. Inf Sci 8(4):301–357
Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning- III. Inf Sci 9(1):43–80
Zaichenko H (2020) Fuzzy cooperative games of two players under uncertainty conditions. IEEE 2nd International Conference on System Analysis & Intelligent Computing (SAIC) : 1-6. https://doi.org/10.1109/SAIC51296.2020.9239214
Zou L, Pei Z, Karimi HR, Shi P (2012) The Unbalanced Linguistic Aggregation Operator in Group Decision Making. IEEE International Symposium on Intelligent Control (ISIC) : 703-707

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

Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-021-06154-w

RESEARCH ARTICLE-CIVIL ENGINEERING

Mechanical and Microstructural Properties of Rubberized Concrete After Surface Modification of Waste Tire Rubber Crumb

Rahul Kumar[1] · **Nirendra Dev**[1]

Received: 15 June 2021 / Accepted: 29 August 2021
© King Fahd University of Petroleum & Minerals 2021

Abstract
Rubberized concrete is highly effective in areas where more flexibility is required. Compared to standard concrete, rubberized concrete has greater flexibility, impact resistance, and lower unit weight. In this study, sulfuric acid is used in the surface treatment process of rubber particles to strengthen the bonding property of rubber particles and cement paste. Six waste tire rubber crumb contents (0, 10, 15, 20, 25, and 30%) and particle sizes of 0.6–2.36 mm were used as partial replacement of fine aggregate in concrete. Microstructural analyses (EDX, TGA, XRD) are conducted on rubber crumbs to examine rubber's compatibility before using it in concrete. The fresh, hardened, microstructural properties (SEM and XRD) and stress vs. strain behavior of rubberized concrete are investigated. From the SEM image, it is analyzed that surface treatment improves the bonding properties of rubber particles and cement paste compared to untreated rubber particles. The mechanical properties of rubberized concrete reduce by increasing the rubber crumb in the mix. Still, the results are not much troubling up to 15% replacement level of rubber crumb, and the rubberized concrete has a highly deformable capacity than standard concrete.

Keywords Rubberized concrete · X-ray diffraction · Energy dispersive X-ray analysis · Scanning electron microscope · Thermal gravimetric analysis · Mechanical properties · Modulus of elasticity

1 Introduction

Over the years, the disposal of waste tires has been increasing the issue continuously for the environment. Waste rubber tires have little scope of being recycled and mostly form a landfill and degrade the environment [1–4]. There is an excellent potential for the rubber to be used in concrete, thus saving the area from becoming a landfill, which means eco-friendly. Rubber is readily available from used tires and is not as costly as other concrete constituents. Thus, if it is used, it will result in more economical concrete. The adhesive property of rubber particles and cement paste is the main issue, and another is the strength decrement in rubberized concrete. These are the gaps that this research is attempting to solve.

According to previous research, the acid (H_2SO_4 and HCl) resistance characteristics and mechanical properties are investigated for 90 and 180 days. The utilization of rubber shreds in concrete enhanced the property of resistance to acid attack but decreased the compressive strength of RuC (rubberized concrete). The mechanical and durability properties of RuC were significantly improved when silica fume is added [5]. The addition of rubber aggregate's proportion and size in concrete increases the microcrack width of RuC [6]. The use of steel fibers has more negative impacts than polypropylene fibers in RuC [7]. The curing methods of concrete are also analyzed, and it is observed that the oven-curing approach gives better strength results than the ambient curing approach for concrete specimens [8]. The contact angle and infrared spectroscopy analyses are done between the rubber and water droplet. The rubber crumbs' modification can improve the adhesion property between cement paste and rubber particles. The compressive strength of RuC after treatment with sodium hydroxide is slightly enhanced compared with ordinary concrete [9]. Ismail et al. [10] investigated fresh and hardened properties self-consolidating RuC using waste rubber crumb and powder rubber as a partial replacement for sand and supplementary

✉ Rahul Kumar
rahul_phd2k17@dtu.ac.in

Nirendra Dev
nirendradev@dce.ac.in

[1] Department of Civil Engineering, Delhi Technological University, Delhi 110042, India

Published online: 27 September 2021

cementing materials. The rubber crumb and powder rubber results are economical and environmentally friendly to alternative aggregates with higher ductility. The inclusion of powder rubber in concrete shows better performance of fresh and mechanical properties of RuC than the inclusion of rubber crumbs [10]. The five methods (cement coating, silica coating on rubber particles, and other three processes are soaking rubber crumb in NaOH, $KMnO_4$, and H_2SO_4 solutions) enhance the strength of RuC compared to RuC without surface treatment of rubber. The silica fume coating and 10% H_2SO_4 solution treatments give the best result for improving the strength of RuC [11].

Gupta et al. [12] uses six levels of rubber fiber (0, 5, 10, 15, 20, and 25%) for replacement of FA and three *w/c* ratios (0.35, 0.45, and 0.55) and three stages of silica fume (0, 5, and 10%) as partial replacement of cement. XRD (X-ray diffraction) and SEM (scanning electron microscope) do the microstructural analysis; however, the compressive strength and MOE (modulus of elasticity) decreased with increasing rubber fiber quantity. Chloride-ion resistance is not changing with the increase in rubber fiber in concrete, but chloride-ion penetration increases when cement is replaced by silica fume. The RuC can be used to design earthquake shock wave absorbers, machinery base foundations, roadway pavement, and airport runways because it has a highly deformable capacity [12]. The flexural strength of RuC decreases with increasing quantity of rubber ash but increases with the increasing content of rubber fiber. "The reduction in static and dynamic modulus of elasticity indicates higher flexibility, which can be viewed as a positive gain in rubberized concrete mixtures" [13]. For the treatment process, He et al. [14] investigated after oxidizing the rubber crumb with $KMnO_4$ solution, then sulfonated with $NaHSO_3$ solution, used the FT-IR for oxidation and sulfonation progression, and presented a large amount of hydroxyl and sulfonate to rubber crumbs. The bonding property of rubber and cement is enhanced by 41.1% after the surface treatment and concluded that the treatment of rubber crumb with $KMnO_4$ and $NaHSO_3$ improves the bonding property with the cement matrix [14]. According to Samar Raffoul et al. [15], 40 RuC mixes were investigated for replacement levels from 0 to 100% for both fine and coarse aggregate. After the surface treatment of rubber crumbs by silica fume, RuC was compared with control concrete and found that rubber with silica fume's surface treatment did not affect RuC's bonding properties. The compressive strength of RuC decreases with increasing quantity of rubber in concrete [15]. The bonding properties of rubber particles with cement can be enhanced by surface treatment of rubber crumbs by NaOH [16–20].

The study's main aim is to minimize the loss of strength of RuC and enhance the bonding between the rubber crumb and concrete matrix. Many preceding kinds of research in this area give nearly the same results: decreased compressive strength, increased flexibility, ductility, impact and toughness resistance, and enhanced and better sound insulation. However, these results are due to untreated rubber or other treatment methods [5, 12, 13, 16–20]. This article analyzes the mechanical and microstructural properties after the rubber crumb's surface treatment by 15% H_2SO_4 solution. Additionally, PPC (portland pozzolana cement) is used in this study because it has a more fineness-specific surface than OPC (ordinary portland cement). It fills the gaps between rubber particles and cement paste and makes an enhanced bond. Pozzolana, such as fly ash, is a siliceous material that has little or no binding properties. In the presence of water, it interacts with $Ca(OH)_2$ to produce compounds having cementitious characteristics. This approach is selected for this study to minimize the loss in strength.

2 Materials and Testing

2.1 Portland Pozzolana Cement

In this study, PPC JK Laxmi pro+, up to 30% replacement by fly ash-based, is used. The compressive strength of PPC is 44 MPa at 28 days [21]. The standard consistency of the cement is obtained at 33% [22]. Sp. Gr. of cement is found to be 2.97 [23]. PPC's initial setting time is 90 min, and the final setting time is 155 min [24]. The soundness of PPC is tested by the Le-Chatelier method, and the cement expansion is noted as 3 mm [25]. The XRD analysis is done to identify the crystalline phases of cement powder passing through an 850-µm sieve, according to ASTM (American Society for Testing and Materials) [26]. The results of XRD are analyzed by the software X'Pert HighScore Plus. The XRD data of PPC are shown in Fig. 1, and peaks are found

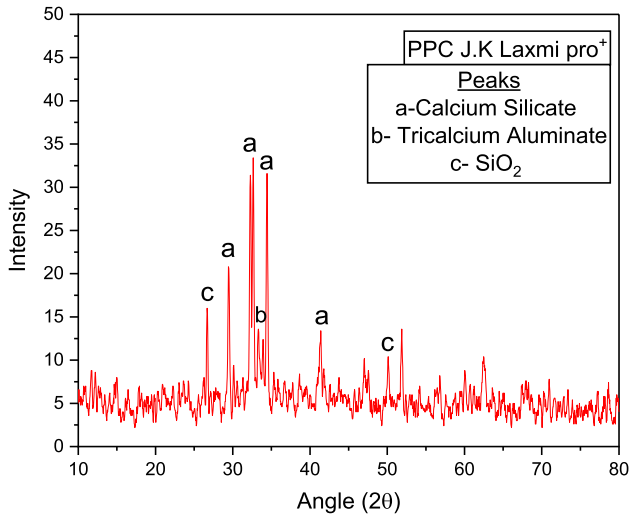


Fig. 1 XRD of J.K Laxmi pro + cement (PPC)

for calcium silicate, tricalcium aluminate, and silica. The chemical composition of PPC is seen from the XRF (X-ray fluorescence), which is presented in Table 1.

2.2 Aggregate

FA in the form of stone dust is procured from a local material shop in Delhi, which falls in zone II with a finesse modulus of 2.75 [27], Sp. Gr. is 2.62, and water absorption is 1.21% [28]. Crushed rock as CA (coarse aggregate) is also procured from a local material shop in Delhi, which has Sp. Gr. of 2.79 at 28 °C, water absorption of 0.2% [28], finesse modulus of 7.29 [27], impact value of 26.3, and Loss Angeles Abrasion Value of 36 [29].

2.3 Superplasticizer (Polycarboxylate Ether-Based)

Auramix 400 SP (superplasticizer) based on polycarboxylate ether is used and obtained from Fosroc Chemical (India). The Sp. Gr. (specific gravity) of SP is 1.09, and the pH value is 6. The process for determining the SP's optimum dosage and efficiency is obtained from concrete technology by Shetty [30]. SP's optimum dosage and efficiency are determined from the marsh cone apparatus with a nozzle aperture diameter of 8 mm. The optimum dosage of SP is found to be 1.4%. The efficiency of the Auramix 400 SP for JK Laxmi pro^{+} cement (PPC) allows the reduction of 22.73% of water.

2.4 H_2SO_4 Solution

H_2SO_4 is procured from RANKEM (Avantor) (UN1830) with 36.4 normality with 98% purity [31].

2.5 Waste Tire Rubber Crumb

Rubber is obtained from a local waste tire rubber factory with a particle size of 0.6–2.36 mm. Rubber is a lightweight material because its Sp. Gr. is 1.07. Rubber crumb is used as a partial replacement of FA. The particle size distribution curve of FA and rubber is shown in Fig. 2. Rubber crumb particles are gained from the waste rubber tire after grinding in a mechanical grinder. Rubber crumb is a waste material, and it is necessary to find compatibility with concrete; some physical and microstructural analyses are essential before use in concrete. For evaluating the chemical elements and mineralogical properties present in waste tire rubber, EDX (energy-dispersive X-ray) and XRD analyses have been done. A high amount of carbon was detected from the rubber

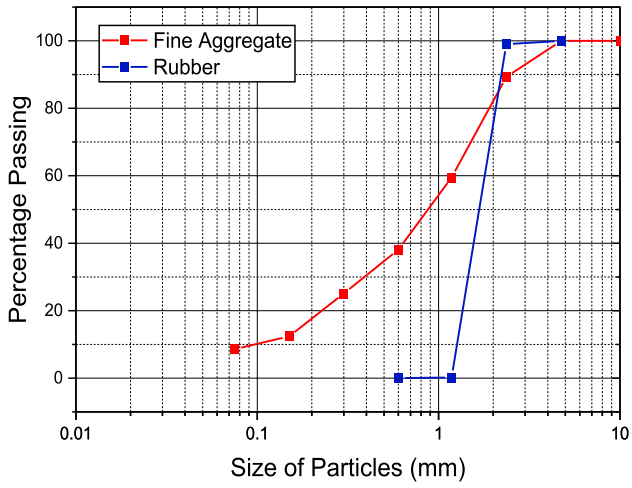


Fig. 2 Grain size distribution curve of aggregate and rubber

crumb sample results, which may be responsible for loss in strength. EDX results of rubber crumbs also show the zinc element present in rubber, which is also accountable for strength loss of concrete according to previous research [9, 11, 32], and it should be cleaned from the rubber crumb. Because of this significant issue, surface treatment of rubber is necessary before adding it to the concrete.

2.5.1 Surface Treatment of Rubber

Strength reduction is the major drawback of the topic of rubberized concrete. Many researchers use untreated rubber crumbs, and they found significant losses in rubberized concrete's mechanical properties. The mechanical strength of RuC decreases by increasing the quantity of rubber in concrete due to the weak bonding between the rubber and the cement matrix. Zinc stearate ($C_{36}H_{70}O_4Zn$) is widely used as a release agent for rubber, polyester processing system, and polyurethane [9, 11, 32]. These applications take advantage of its "nonstick" properties. It is a lubricant and thickening agent and is also used to improve texture. Stearic acid and zinc are component compounds for making zinc stearate. Zinc stearate is present in the rubber tire, which shows hydrophobic properties and tends to repel the water. Due to this property of rubber, particles make a weak bonding with the cement matrix. Due to this significant issue, rubber particles should be treated before mixing them in concrete to enhance RuC's compressive strength. 15 wt% H_2SO_4 solution is used for treating rubber tire particles. In this process, rubber particles are saturated for 2 h in 15% sulfuric acid

Table 1 Chemical composition of PPC cement (J.K Laxmi pro+ cement)

Component	Cao	SiO_2	Al_2O_3	Fe_2O_3	SO_3	MgO	Na_2O	K_2O
J.K Laxmi pro + (PPC)	65.80	22.64	3.80	3.10	1.40	1.20	0.31	1.21



Fig. 3 Process of surface treatment of the rubber crumb

Table 2 Percentage weight of chemical elements without surface treatment

Element	Weight%	Atomic%
C K	62.80	71.52
O K	29.92	25.58
Al K	0.80	0.40
Si K	2.39	1.16
S K	0.64	0.27
Ca K	2.56	0.87
Zn K	0.88	0.18
Total	100	

solution and cleaned with water and then air-dried for one day in an open area. In this rubber treatment process, H_2SO_4 solution reacts with zinc element and converts it into zinc sulfate that can be washed by running tap water because it is insoluble in water [11, 32]. The surface treatment of the rubber crumb process is shown in Fig. 3.

2.5.2 EDX and XRD Analyses of Rubber

For the efficient use of waste tire rubber, microstructural analysis has been following up in this study. Energy-dispersive X-ray (EDX) analysis has been done to determine the chemical composition of rubber. As a result, it is found that zinc is present in a rubber tire, which is responsible for weak bonding with the cement matrix. Without surface treatment, rubber particles show nonstick properties due to zinc stearate [9, 11, 32], and it has to be removed to make good bonding between rubber and cement matrix. After EDX analysis of rubber without surface treatment, Zn is found 0.88% by weight in Table 2. EDX data of untreated rubber in Fig. 4a show all chemical elements present in the rubber. The XRD analysis is done on rubber powder that passes through 150-µm sieve and retains on 45-µm sieve, according to ASTM [33]. The results of XRD are analyzed by the software X'Pert HighScore Plus, and the selected chemical elements in the software are C, O, Al, Si, S, Ca, Zn, N, Mg, Mn, and Fe. XRD analysis indicates rubber's mineralogical properties, and the peaks show silica and calcium carbonate in Fig. 5. The nature of rubber crumbs is found crystalline before the surface treatment by sulfuric acid solution.

Then, 10 wt% sulfuric acid solution is used for rubber's surface treatment, and again EDX analysis is done to find

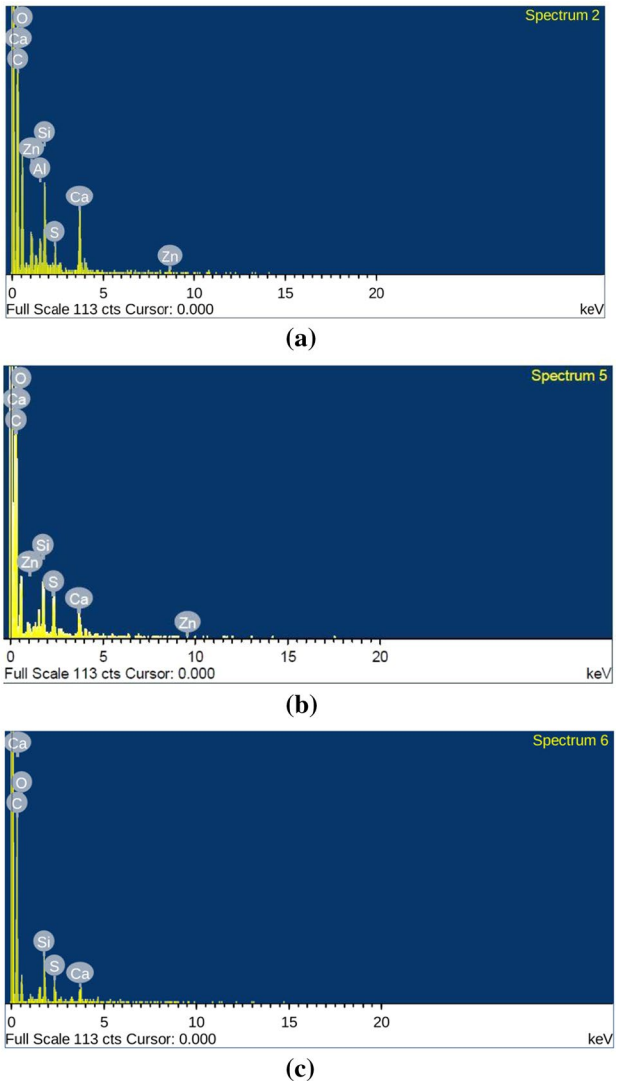


Fig. 4 **a** EDX data of rubber without surface treatment. **b** EDX data of rubber with surface treatment by 10% H_2SO_4. **c** EDX data of rubber with surface treatment by 15% H_2SO_4

the elements. After analyzing, zinc is found 0.09% in EDX analysis (Fig. 4b). All the chemical elements are present in Table 3 after surface treatment of rubber by 10 wt% of H_2SO_4. In this event, the amount of carbon is increased compared to without treated rubber particles. The zinc element is present in rubber crumbs in this process which should be

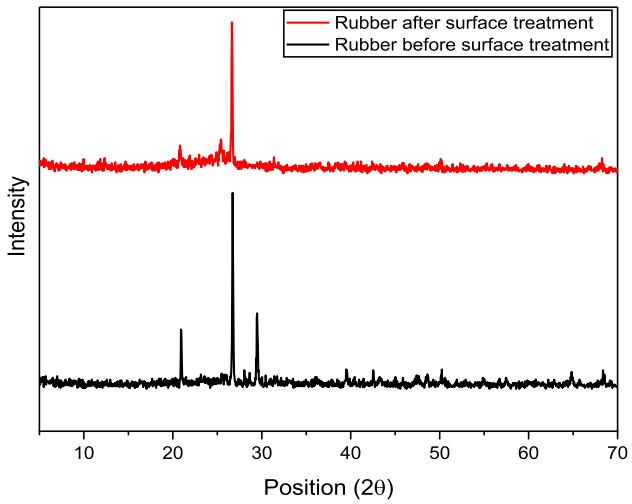


Fig. 5 XRD data of rubber before surface treatment and after surface treatment by 15% H_2SO_4

Table 3 Percentage weight of chemical elements with surface treatment by 10% H_2SO_4

Element	Weight%	Atomic%
C K	73.24	80.13
O K	21.69	17.82
Si K	1.85	0.86
S K	1.67	0.68
Ca K	1.46	0.48
Zn K	0.09	0.02
Total	100	

Table 4 Percentage weight of chemical elements with surface treatment by 15% H_2SO_4

Element	Weight%	Atomic%
C K	70.25	78.10
O K	22.71	18.95
Si K	3.20	1.52
S K	1.74	0.72
Ca K	2.10	0.70
Total	100	

removed from the rubber surface to enhance the bonding property of rubber and cement.

Due to this issue, rubber is again treated with 15 wt% sulfuric acid solution, and then another time EDX analysis is done. After investigating the results of EDX, it is found that the Zn element is completely removed from the rubber particles and is ready to use for making the RuC. All the outcomes of EDX analysis are shown in Table 4 and Fig. 4c, and XRD analysis of rubber particles treated with 15% H_2SO_4 solution offers mineralogical properties in Fig. 5. The significant peaks show silica in the majority, and at this stage, calcium carbonate did not form. All the informative data that belong to XRD peaks, like Miller indices, d-spacing, chemical formula, and crystal system, are present in Table 5. The nature of rubber crumb is found crystalline after the surface treatment by sulfuric acid solution, and no harmful compound is found before mixing it in the concrete.

15 wt% H_2SO_4 solution is used to treat rubber tire particles in this study to remove zinc from the rubber crumbs and enhance the mechanical property of RuC. The H_2SO_4 solution reacts with the zinc element in this treatment process, forming zinc sulfate, which can be washed away with tap water because it is insoluble in water. This process reduces the hydrophobic nature and enhances the hydrophilic nature of rubber crumbs, so the rubber can make a better bond with the cement matrix and enhance the mechanical properties of RuC.

2.5.3 TGA and DTG of Rubber

To analyze the weight–temperature relationship of material, TGA (thermal gravimetric analysis) has been done on two rubber samples, according to ASTM [34]. In this process, a small amount of rubber particles, approx 10 mg, is taken for TGA. The sample is burned from 0 to 550 °C with a nitrogen atmosphere at the rate of 10 °C/min and then cooled down to 350 °C. In the second stage, the sample again burns up to 800 °C with the oxygen atmosphere. These specific settings are selected according to ASTM-E2550 (2017). The weight of the sample is weighed simultaneously with the burning process. For rubber crumbs without surface treatment in TGA, 73.42% weight is lost in the first stage due to evaporation, oils, and an organic substance present in rubber crumbs. In the second stage, 7.29% weight is lost due to carbon black present in the rubber crumb, and 19.28% residue remains. It is challenging to analyze the proper temperature of weight loss from the TGA graph, so its derivative is necessary to examine the exact temperature of phase changes. From the DTG (derivative thermogravimetry) curve, the phase changes with proper temperature can be measured. In the DTG graph, the phase of the rubber crumb changes from 300 to 500 °C (Fig. 6a).

From the second sample of rubber particles after surface treatment by H_2SO_4 solution, 58.33% weight is lost in the first stage due to evaporation, oils, and an organic substance present in the rubber crumb. In the second stage, 14.50% weight is lost due to carbon black, which is present in rubber crumbs, and 27.33% residue remains. In the DTG graph, the first phase of rubber crumb changes from 300 to 500 °C, and the second phase is changed from 600 to 750 °C (Fig. 6b).

From Fig. 6a and b, after the rubber sample's burning process, only 26.58% of the residue is retained in untreated rubber, but the treated rubber remaining residue is 41.67% at the temperature range 300–500 °C which is confirmed by DTG graph. The surface-treated rubber has more stability

Table 5 XRD data of materials and concrete mixes

S. no.	Material	Peaks no.	Peaks angle	h k l	d-Spacing	Chemical formula	Chemical name	Crystal system
1	Rubber without surface treatment	1	20.9467	1 0 0	4.24108	SiO_2	Silicon oxide	Hexagonal
		2	26.7382	1 0 1	3.33417	SiO_2	Silicon oxide	Hexagonal
		3	29.4805	1 0 4	3.02997	$CaCO_3$	Calcium carbonate	Rhombohedral
		4	42.5557	2 0 0	2.12443	SiO_2	Silicon oxide	Hexagonal
		5	50.3433	1 1 2	1.81255	SiO_2	Silicon oxide	Hexagonal
		6	64.854	3 0 0	1.43652	$CaCO_3$	Calcium carbonate	Rhombohedral
2	Rubber after surface treatment	1	20.8198	1 0 0	4.26665	SiO_2	Silicon oxide	Hexagonal
		2	25.4372	1 1 1	3.50166	SO_2	Sulfur oxide	Orthorhombic
		3	26.6557	0 1 1	3.34429	SiO_2	Silicon oxide	Hexagonal
		4	50.1246	0 0 3	1.81995	SiO_2	Silicon oxide	Hexagonal
		5	68.1166	2 0 3	1.37544	SiO_2	Silicon oxide	Hexagonal
3	R1 MIX	1	20.9131	1 0 0	4.24431	SiO_2	Silicon oxide	Hexagonal
		2	22.8911	1 0 1	3.88184	S_6	Sulfur	Hexagonal
		3	26.6719	1 0 1	3.33954	SiO_2	Silicon oxide	Hexagonal
		4	29.08	1 1 2	3.06824	CS_2	Carbon sulfide	Orthorhombic
		5	39.5176	1 0 2	2.27858	SiO_2	Silicon oxide	Hexagonal
		6	50.1565	1 1 2	1.81736	SiO_2	Silicon oxide	Hexagonal
		7	59.9996	1 2 1	1.54061	SiO_2	Silicon oxide	Hexagonal
4	R2 MIX	1	20.9651	1 0 0	4.23391	SiO_2	Silicon oxide	Hexagonal
		2	26.6827	1 0 1	3.33821	SiO_2	Silicon oxide	Hexagonal
		3	27.7885	1 1 1	3.20783	Ca	Calcium	Cubic
		4	29.5673	1 1 0	3.01877	Na	Sodium	Cubic
		5	50.2222	1 1 2	1.81514	SiO_2	Silicon oxide	Hexagonal
		6	54.9122	3 1 1	1.67068	Ca	Calcium	Cubic
5	R3 MIX	1	10.7363	1 1 1	8.24048	C_{60}	Carbon	Cubic
		2	26.9052	1 0 0	3.31386	SiO_2	Silicon oxide	Hexagonal
		3	27.9793	3 3 1	3.18903	C_{60}	Carbon	Cubic
		4	50.3182	1 1 2	1.8134	SiO_2	Silicon oxide	Hexagonal
		5	60.3215	1 2 1	1.53442	SiO_2	Silicon oxide	Hexagonal
		6	66.3602	3 0 0	1.40752	SiO_2	Silicon oxide	Hexagonal
6	R4 MIX	1	8.9086	− 2 0 1	9.92662	SiO_2	Silicon oxide	Monoclinic
		2	21.0413	1 0 0	4.22223	SiO_2	Silicon oxide	Hexagonal
		3	26.7279	1 0 1	3.33543	SiO_2	Silicon oxide	Hexagonal
		4	28.7419	0 1 1	3.10612	SiO_2	Silicon oxide	Hexagonal
		5	29.5532	1 1 0	3.02268	Na	Sodium	Cubic
		6	39.5923	0 1 2	2.27633	SiO_2	Silicon oxide	Hexagonal
		7	50.269	1 1 2	1.81506	SiO_2	Silicon oxide	Hexagonal
7	R5 MIX	1	9.0443	1 0 0	9.77795	Mn_6O_{12}	Manganese oxide	Monoclinic
		2	10.7558	1 1 1	8.22558	C_{60}	Carbon	Cubic
		3	21.1319	1 0 0	4.20434	SiO_2	Silicon oxide	Hexagonal
		4	26.9368	1 0 1	3.31004	SiO_2	Silicon oxide	Hexagonal
		5	32.9272	2 2 2	2.72025	Mn_2O_3	Manganese oxide	Cubic
		6	55.1013	4 4 0	1.66539	Mn_2O_3	Manganese oxide	Cubic
8	R6 MIX	1	20.9947	1 0 0	4.2315	SiO_2	Silicon oxide	Hexagonal
		2	26.7789	0 1 1	3.32919	SiO_2	Silicon oxide	Hexagonal
		3	29.6353	1 1 0	3.01449	Na	Sodium	Cubic
		4	42.5935	2 0 0	2.12263	SiO_2	Silicon oxide	Hexagonal
		5	60.1237	1 2 1	1.539	SiO_2	Silicon oxide	Hexagonal
		6	68.4913	2 0 3	1.36883	SiO_2	Silicon oxide	Hexagonal

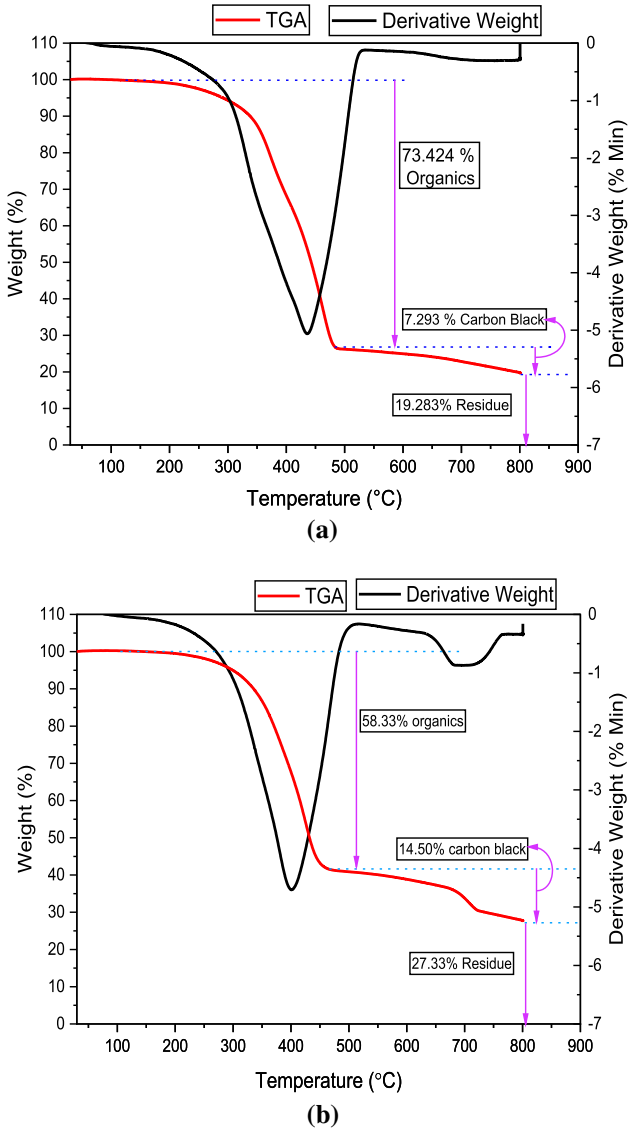


Fig. 6 **a** TGA of rubber without surface treatment. **b** TGA of rubber with surface treatment by 15% H_2SO_4

against temperature compared to untreated rubber. As a result, RuC after acid treatment of rubber crumbs is more temperature stable than RuC without acid treatment of rubber crumbs.

3 Mix Proportion

The concrete mixes are prepared with a w/c ratio of 0.38 with a partial replacement of FA by rubber particles with a 0–30% replacement level. The proportion of mix design for M_{40} grade of concrete is shown in Table 6. It is necessary to maintain the workability of concrete mixes; the amount of SP is used at 1.4%, according to the marsh cone apparatus discussed earlier. Machine mixing is adopted for the homogeneity of the concrete. First, dry mixing of material is done for 2–3 min, and then water and SP are mixed with materials. A mechanical vibrator is used for the compaction and uniformity of the concrete. One of the issues is that the vibration's additional time pushes up the rubber particles on top of the concrete. Due to this problem, RuC is not homogeneous. Vibration time for mix has been adopted 5–10 s to overcome this problem [35], or other hands can be tried for self-compacting RuC.

4 Testing Program

Various tests are performed to find the concrete properties, such as compressive strength, split tensile strength, workability, density, Poisson's ratio, MOE, and flexural strength. Three specimens are cast for each testing of concrete. The density of the concrete mixes is measured according to IS code (Indian Standard code) [36]. The compressive strength and flexural strength of RuC are tested on a universal testing machine with 2000 tons capacity and an automatic controller with specimen size 150X150X150 mm for compressive strength and 100 × 100 × 500 mm [37] for flexural strength. Poisson's ratio, MOE, is also tested with the help of LVDT (linear variable differential transformer) on a specimen's size

Table 6 Mix proportion for concrete

Mix no.	Cement (kg/m^3)	FA (kg/m^3)	CA (kg/m^3)		Rubber crumbs (kg/m^3)	w/c ratio	SP (%)	Slump mm	Compressive strength at 28 days (MPa)
			(20 mm)	(10 mm)					
R1	390	666	769.8	513.2	0	0.38	1.4	99	50.38
R2	390	599.4	769.8	513.2	66.6	0.38	1.4	96	42.22
R3	390	566.1	769.8	513.2	99.9	0.38	1.4	94	40.12
R4	390	532.8	769.8	513.2	133.2	0.38	1.4	91	27.89
R5	390	499.5	769.8	513.2	166.5	0.38	1.4	86	20.08
R6	390	466.2	769.8	513.2	199.8	0.38	1.4	84	18.51

300 mm × 150 mm [37]. LVDTs are also used to find the lateral and longitudinal deformation of RuC and compared with standard concrete. One LVDT is placed in the axial direction of load, and the second LVDT is placed in the lateral position of the axial load. The loading rate is applied gradually to the sample 140 kg/cm^2/min for compressive strength, 180 kg/min for flexural strength, and 140 kg/cm^2/min [38] for the concrete test's modulus of elasticity. SEM (Scanning electron microscope) performed the concrete specimen's microstructural analysis on a small size cut from concrete samples size of approximately 1 × 1 cm. SEM analysis is done to evaluate the bonding between the concrete matrix and the rubber particle.

Modulus of elasticity and Poisson's ratio are measured on cylindrical specimens. For specimen preparation, neat cement capping is done after 2–3 h of casting. The uniformity of capping is checked by a perpendicular tool in Fig. 7a. The specimen's marking should be such that the extensometer's upper and lower ring makes an equal distance from the specimen's upper surface and lower surface. The middle ring of the extensometer for Poisson's ratio is placed at the exact mid-height of the cylinder specimen in Fig. 7b. Two LVDTs are placed for finding the modulus of elasticity and Poisson's ratio: One LVDT is placed in the axial direction of load, and other is placed in the lateral position of the axial load. The arrangements of LVDTs and extensometer of cylinder specimen are shown in Fig. 7c. For measuring and controlling LVDTs, AI 8000+ Data Acquisition and Control System with "Prosof" software is used for MOE and Poisson's ratio (Fig. 7d) according to IS code [38]. Figure 7e shows the testing setup for static MOE and Poisson's ratio.

5 Results and Discussion

5.1 Workability

At the time of batch mixing, to maintain the workability of concrete, 1.4% SP by the weight of cement is used. The slump cone results are shown in Fig. 8. From the results, it is analyzed that RuC's workability decreases with increasing quantity of rubber crumbs. For medium workable concrete, the slump value should be 50–100 mm so that all mixes are in the medium range of workability. The RuC, replacement level of up to 30%, can be used with reinforced concrete members, and no need to add extra SP for making the concrete workable. A similar decreasing pattern of the slump after increasing rubber crumbs in concrete is found in the



Fig. 7 **a** Capping of the cylinder specimen, **b** marking of the cylinder specimen, **c** arrangements of extensometer and LVDTs, **d** data acquisition system for LVDTs, **e** testing setup of the cylinder for modulus of elasticity and Poisson's ratio

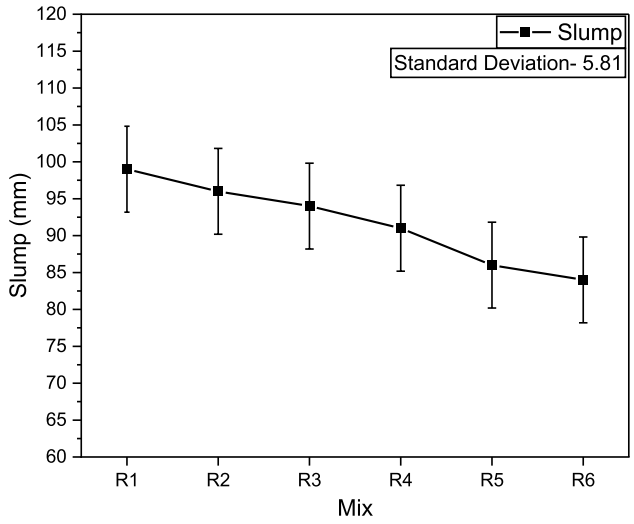


Fig. 8 Results of slump cone for all mixes

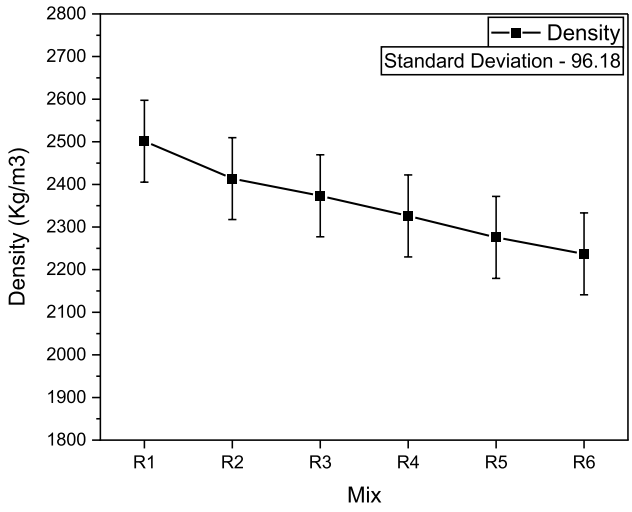


Fig. 9 Density of different mixes

past studies [12, 13]. The rubber particles have a rough surface and lead to large undulations in concrete. The greater the rubber content in the mixture, the reduction in the value of the compaction factor would be the more. The inter-particle resistance between the cement matrix and the rubber particles around them is strong, decreasing the mixture's consistency. The RuC mix density decreases in the fresh state, which is the cause of reducing the degree of compaction of the mix [39].

5.2 Density

The mix's density results are shown in Fig. 9. The results show that the density of concrete mixtures decreases with increasing quantity of rubber crumb in concrete. The rubber crumb has low Sp. Gr. than the FA, which is the reason for the decrement pattern in the density of RuC. A similar decreasing pattern of the density after increasing rubber crumbs in concrete is found in the past studies [12, 13]. "Owing to a low specific gravity value of rubber particles which replaces dense M sand, the decrease in density ought to have observed [40]." The Sp. Gr. of natural FA is much higher than the waste tire rubber crumb, so the density of RuC will decrease. The appearance of entrapped air on rubber particles' rough surfaces could be another explanation for the decrement in the density of RuC [41, 42].

5.3 Compressive Strength

The compressive strength of the concrete mixes at 7, 14, and 28 days after casting is presented in Fig. 10. The decrement in the compressive strength at 15% replacement level of rubber crumb is just 20.36% after 28 days of curing, and the compressive strength of the RuC is 40.12 MPa, which is suitable for the M40 classification. The compressive strength of RuC is not much affected at 15% replacement level of rubber crumb after surface treatment of rubber by H_2SO_4. The low adhesion between the rubber particle and the cement paste reduces the bonding, resulting in a declining trend in compressive strength at all ages. Rubber crumbs are softer and more elastically deformable than the cement matrix around them, indicating that the material stiffness of rubber and cement paste is incompatible. The above factors make it easier for cracks to form at the cement paste and rubber crumb interface, and the cracks' progression speeds up, reducing RuC's compressive strength. Gupta et al. suggested that the rubber replacement level could be used up to 10% without significant loss of compressive strength in RuC because used rubber particles were without treatment [12]. Still, in the present study, 15% replacement level of rubber crumb can be used after the surface treatment. According to

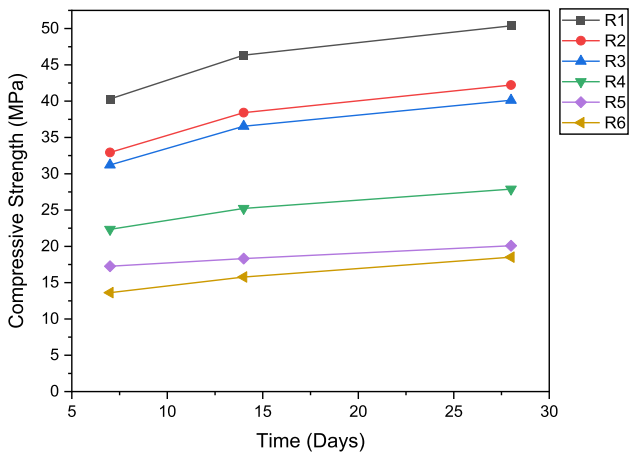


Fig. 10 Compressive strength of standard concrete and RuC

past studies, the compressive strength of concrete decreases by increasing the replacement level of rubber particles in concrete, and they show the same pattern as found in the present study [5, 6, 10–12, 14, 35, 43–47].

On behalf of Uygunoğlu, Topçu, the compressive strength is decreased by 40–64% and 48–58% at 7 and 28 days, respectively, using rubber aggregate in the concrete mix [4], but in this study, after the treatment process, the decrement is 18–22% at 7 days and 16–21% at 28 days up to 15% replacement of rubber which may be considered as positive results. Moreover, according to Qingwen Ma, the reduction in compressive strength at 20% replacement level of rubber after rubber treatment by NaOH is 48.95%. Still, in this study, the reduction is 20.36% at 28 days after surface treatment by H_2SO_4 [32]. The surface treatment of rubber crumbs sulfuric acid reduces the loss in the compressive strength of RuC.

5.4 Split Tensile Strength

The split tensile strength results of the concrete mixes at 7, 14, and 28 days after casting are shown in Fig. 11. In this study, at 15% replacement level of rubber crumbs, the reduction is only 22.48% at 28 days of curing, and the split tensile strength is found to be 3.24 MPa. There is a less substantial loss in split tensile strength of RuC after the surface treatment of the rubber crumb. Because of the weak bonding between rubber crumbs and cement, the RuC does not bear the much tensile load than standard concrete, and the gaps between rubber crumbs and cement become larger quickly. The RuC does not make a strong bond between the rubber crumb and cement as the standard concrete makes the bond between FA and cement. Gurunandan et al. found a 37.43% reduction in split tensile strength [46], but the reduction is only 22.48% at 15% replacement level of rubber crumbs in this study. The split tensile strength is achieved approximately 8–11% of concrete's compressive strength at 28 days of curing. The decrement in split tensile strength is 17–60% at 7 days and 16–58% at 28 days. Researchers exhibit the same pattern of losses of the split tensile strength of concrete after the inclusion of rubber crumb in concrete due to low Sp. Gr. and gap between cement matrix and the rubber crumbs [46, 47]. The outcomes are consistent with the work stated by H.A. Toutanji [3].

5.5 Flexural Strength

The flexural strength results of the concrete mixes at 7, 14, and 28 days after casting are shown in Fig. 12. In this study, at 15% replacement level of rubber crumb, the reduction is 28% at 28 days of curing, and the flexural strength is found to be 4.00 MPa. There is no significant loss in flexural strength of RuC after the surface treatment of the rubber crumb. The flexural strength is achieved approximately 9–14% of concrete's compressive strength at 28 days of curing. The decrement in flexural strength is 18–59% at 7 days and 17–53% at 28 days at replacement level of rubber 0–30%. Since microcrack forms when the post-peak region is reached and propagates after exceeding this stage in flexural strength testing, cracking may begin before the maximum load is applied. A similar decreasing pattern of the flexural strength after increasing rubber crumbs in concrete is found in past studies [13, 48]. According to Najim, flexural strength decrement was 11% for a 5% rubber replacement level, but at 15% rubber replacement level, the decrement was 39% [39]. Still, in this study, the reduction is only 28%, and it can be considered positive results after the surface treatment of rubber crumb by sulfuric acid.

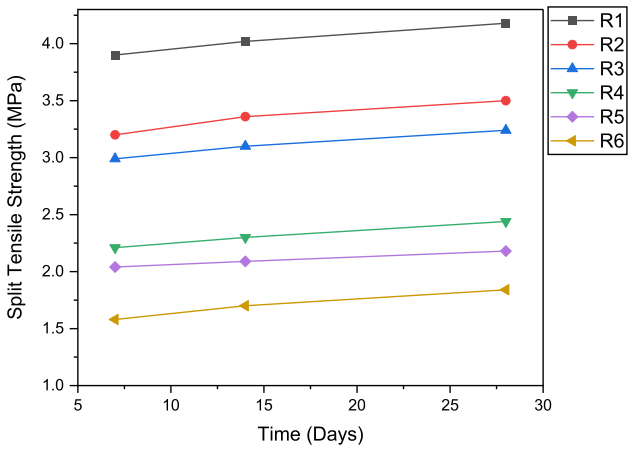


Fig. 11 Split tensile strength of standard concrete and RuC

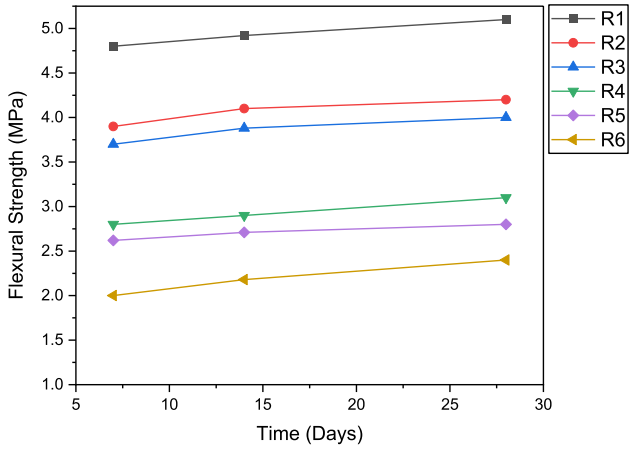


Fig. 12 Flexural strength of standard concrete and RuC

5.6 Compressive Strength of Concrete Mixes by Rebound Hammer

Before the destructive test on the samples, the rebound strength of all mix specimens is measured at 7, 14, and 28 days after casting. Proceq Schmidt Digital Concrete Test Hammer is used to analyze the compressive strength of concrete, according to IS code [49]. The compressive strength of mixes is shown in Fig. 13. After analyzing the results, the compressive strength of concrete mixes by rebound hammer is slightly greater than the hydraulic compression testing machine results. The surface hardness of the specimens influences the rebound strength of the concrete. The rebound test should be performed when the cube specimen is fixed between both faces of the hydraulic machine. This technique gives accurate compressive strength results, but when the specimen is on the floor, the testing is done in a downward direction; it does not provide accuracy.

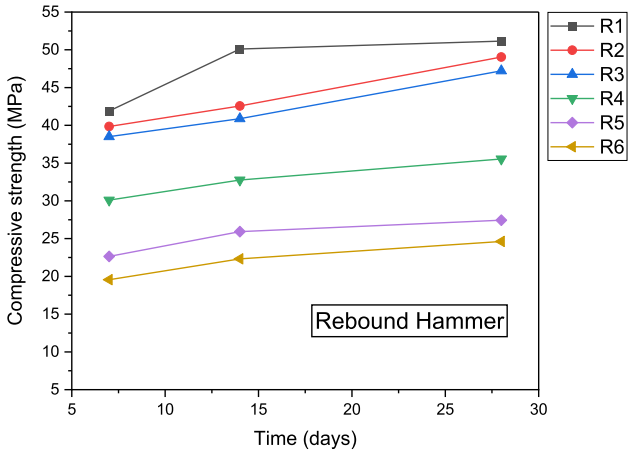


Fig. 13 Compressive strength of standard concrete and RuC by rebound hammer with age

5.7 Ultrasonic Pulse Velocity

A nondestructive test by "MATEST" UPV tester is performed on prism samples to examine wave propagation's velocity in the material and the dynamic MOE of the different concrete mixes, as shown in Fig. 14a according to IS code [50]. The UPV test measures the passage of time from the specimens to determine the performance of the hard material. When the time transit of the UPV wave through the specimen is large, it indicates that the concrete is weaker and vice versa. The velocity for prism specimens for concrete mixes is shown in Fig. 14b. The UPV is also reduced by replacing the FA with rubber crumbs because rubber makes a void-like structure in concrete which increases the time of UPV. At 15% replacement level of rubber, the UPV is found 4.25 km/s, which shows good quality of the concrete. A similar pattern of the UPV results after increasing rubber crumbs in concrete is found in the past studies [12, 13, 39].

5.8 Static Modulus of Elasticity and Poisson's Ratio

The static MOE and Poisson's ratio are evaluated to observe standard concrete and RuC's deformation capacity. Static MOE results at 28 days after casting are shown in Fig. 15 and Table 7. Figure 16a expresses axial stress vs. axial strain pattern. The MOE is decreasing with increasing the rubber crumb in the concrete. In the initial phases of loading, the curves indicate a parabolic form, but as it passes failure loads, it smoothens out and becomes a straight line. Standard concrete's static MOE is found to be as high as 29.48 GPa, indicating that it is a very stiff concrete compared to RuC. Since rubber has an elastically deformable capacity,

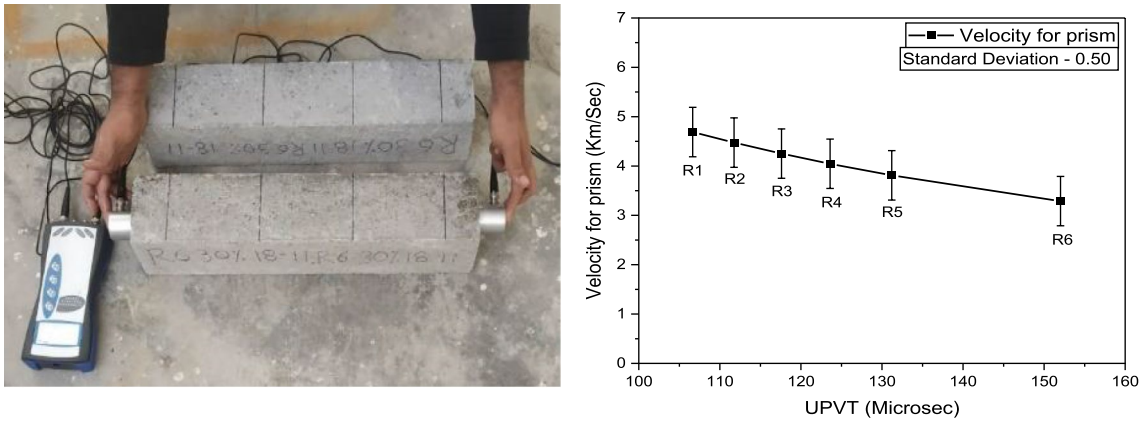


Fig. 14 **a** Ultrasonic pulse velocity test of standard concrete and RuC. **b** Ultrasonic pulse velocity of mixes at 28 days

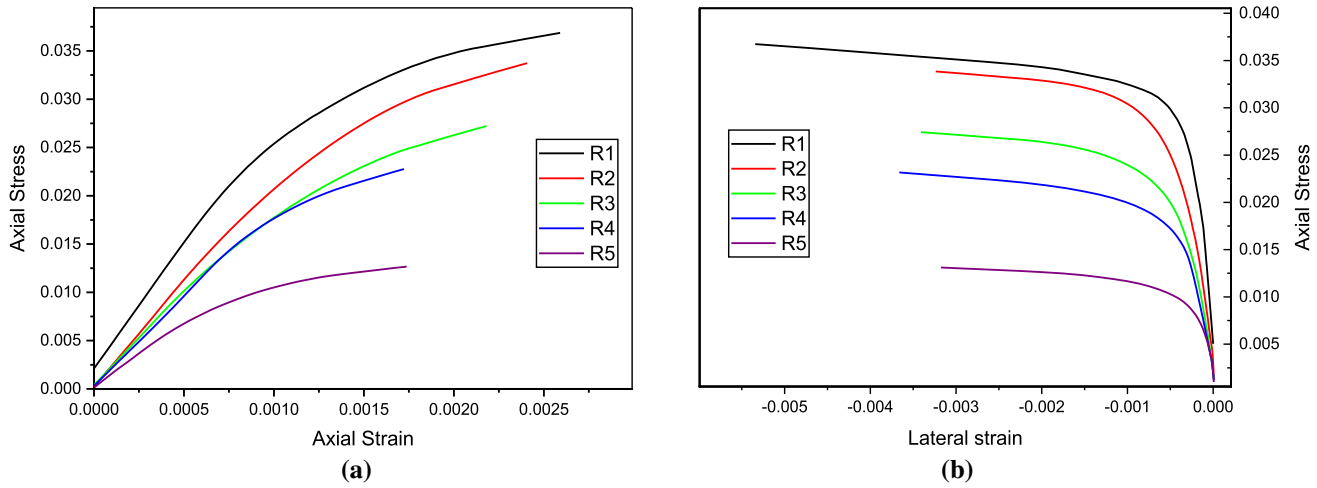


Fig. 15 **a** Axial stress vs. axial strain diagram of mixes at 28 days. **b** Axial stress vs. lateral strain diagram of mixes at 28 days

Table 7 Results of Static MOE, Poisson's ratio, and dynamic MOE

Mix	Compressive strength (MPa)	Static modulus of elasticity (GPa)	Poisson's ratio	Dynamic modulus of elasticity (GPa)
R1	50.38	29.48	0.14	56.04
R2	42.22	22.37	0.28	49.25
R3	40.12	20.53	0.3	43.73
R4	27.89	19.37	0.35	38.79
R5	20.08	12.68	0.39	33.68
R6	18.51	#	#	24.66

[#] Specimen failure

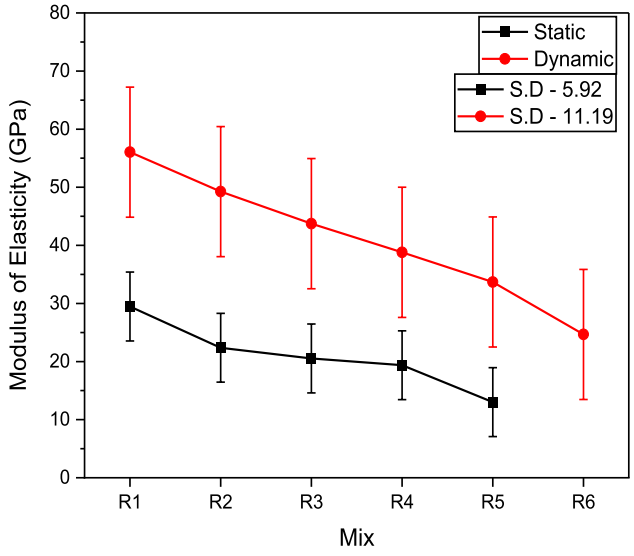


Fig. 16 Static and Dynamic modulus of elasticity of mixes at 28 days

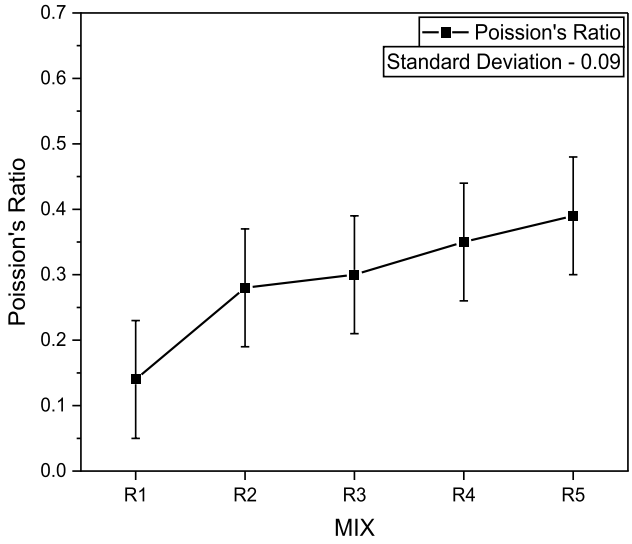


Fig. 17 Poisson's ratio of mixes at 28 days

it improves the ductility of RuC when mixed with concrete. This increased ductility and ability to withstand large deformations before failure due to its low MOE of RuC. MOE of standard concrete and RuC is calculated from the initial slope of the axial stress–strain curves, and a similar decreasing pattern of MOE after increasing the quantity of rubber crumbs in concrete is found in the past studies [44, 51].

The Poisson's ratio results at 28 days after casting are shown in Fig. 17 and Table 7. Figure 16b offers the pattern of axial stress vs. lateral strain. The Poisson's ratio is found at 0.3 at a 15% replacement level of rubber crumb, higher than the standard concrete. The Poisson's ratio increases by increasing rubber crumb in the concrete. Due to the higher value of Poisson's ratio, the RuC has a higher deformable capacity. It is essential to notice that RuC's behavior is not

completely elastic, so Poisson's ratio is not stable in the loading phase. Similar findings of Poisson's ratio are observed in past studies [7, 52, 53].

5.9 Dynamic Modulus of Elasticity

The results of the dynamic MOE for different mixes are shown in Fig. 15. The dynamic MOE decreases with increasing the rubber crumb as a partial replacement in the concrete mix in a similar way observed in concrete's compressive strength because of the porous structure. The dynamic MOE results with a replacement level of 0, 10, 15, 20, 25, and 30% by rubber crumb, as shown in Table 7. The dynamic MOE is found slightly larger than the static MOE, as described by IS code [50]. The decrease in dynamic MOE is due to FA replacement with rubber crumbs, which shows the absorption of ultrasonic waves. This decrement in elastic modulus offers higher flexibility, which can be considered a positive advantage in RuC mixtures. A similar decreasing pattern of the dynamic MOE results after increasing rubber crumbs in concrete is found in the past studies [12, 13, 39, 54].

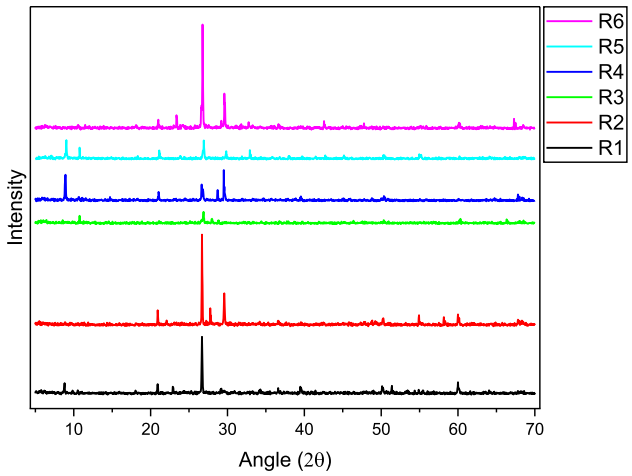


Fig. 18 XRD results of all mixes after 28 days

5.10 Microstructural Analysis (XRD and SEM)

The XRD analysis is done on all concrete mix after 28 days of curing after making the fine powder, according to ASTM [33]. XRD analysis indicates the mineralogical properties of all mixes in Fig. 18. All the mixes are crystalline in nature because they have definite and regular geometry, as shown in Table 5. All the informative data that belong to XRD peaks, like position, Miller indices, d-spacing, chemical formula, chemical name, and crystal system, are present in Table 5.

All the mixes of standard and RuC have the hexagonal shape of crystalline silicon oxide in the majority. Crystalline quartz is significantly less reactive, and it will not react at normal conditions and only react at a very high temperature which is beneficial for RuC. Since quartz is anhydrite of an acid, it would not be attacked by acids in general. Gypsum and ettringite are the two chemical compounds that can harm the RuC by creating internal stress and expanding the inner materials of the concrete. These compounds are generated by a reaction between sulfuric acid and calcium hydroxide. For this reason, rubber is washed with tap water several times in the treatment process to remove the effect of sulfuric acid. In Table 5, in all mixes of RuC, no traces are found of any harmful chemical compounds in XRD analysis that can deteriorate the RuC. Furthermore, the treatment process of rubber crumb will not harm the properties of RuC.

The SEM analysis is done on RuC before and after the surface treatment of rubber particles with H_2SO_4 after 28 days of curing. SEM image of RuC before and after treatment of rubber particles with H_2SO_4 is shown in Fig. 19a and b. According to the SEM study, the rubber particles with no surface treatment show a wide gap between the rubber particle and the concrete compared to the rubber particles with surface treatment by H_2SO_4. The interfacial transition

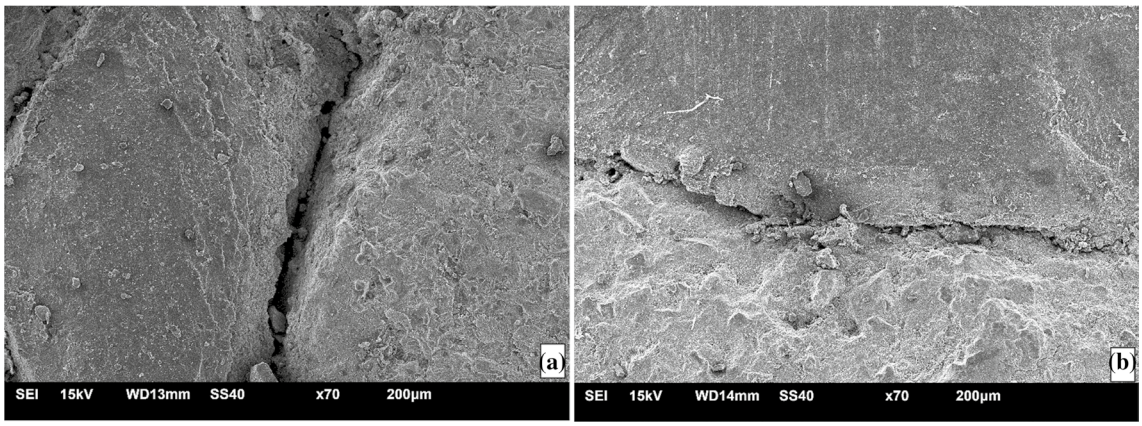


Fig. 19 **a** SEM image of RuC without treatment of rubber particles. **b** SEM image of RuC after treatment of rubber particles with sulfuric acid

zone between the untreated rubber crumbs and concrete is large, making the weak concrete. According to L. P. Rivas-Va'zquez et al., the mechanical properties of RuC are not improved by treating the rubber surface with 10% acetone. With surface treatment by NaOH solution, the SEM picture reveals a poor adhesion between the phases of concrete and rubber [51]. Previous studies [42, 55] found similar results for the gaps between rubber crumbs and cement in tire rubber waste concrete. Because of the rubber particles' hydrophobic nature, the gaps are caused by inadequate compaction of the cement paste around them. Gupta et al. [12, 13] presented a similar argument about the interfacial zone cement and rubber particles.

In this study, after surface treatment of rubber crumbs by sulfuric acid solution, the gaps between rubber and cement are less compared to untreated rubber particles. This process enhances the bonding properties of RuC. The etching of the tire crumb surface and the elimination of hydrophobic admixtures from the tire crumbs are two factors that contribute to the treatment process. Producing improved hydrophilic characteristics can help to improve the bonding between rubber crumbs and cement paste. Before making the RuC, the rubber crumbs should be treated with 15% sulfuric acid solution. The RuC with treated rubber by sulfuric acid reduces the space between the rubber particle and the concrete but cannot eliminate the gaps entirely. So, surface treatment of rubber particles necessitates the use of a more modern process.

5.11 Correlations

5.11.1 Correlation Between Replacement Level of Rubber Crumb and Compressive Strength of RuC

After analyzing the results of RuC and compared with normal concrete, the compressive strength of RuC can be predicted by the following relationship, which is denoted by Eq. 1. It should be noted that this proposed equation is derived from the experimental analysis of RuC. The experimental data of compressive strength of RuC at different replacement levels of rubber crumbs at 28 days of water curing, which are shown in Table 6, are used for establishing this correlation. A linear fitting curve can represent the best-fit curve for the plotted data points for this correlation. The origin software performs this curve fitting. In the correlation analysis, the R^2 is obtained by 0.932, which means that this correlation can predict accurate results, approximately near the experimental results. The scattering in the data from the tests on the specimens influences the accuracy of this correlation. It is a linear correlation between the replacement level of the rubber crumbs and the compressive strength of RuC. The relationship between the replacement level of the rubber crumb and the compressive strength of RuC is shown in Fig. 20a. This relationship is only applicable to rubber crumbs of size 0.6–2.36 mm and 0.38 w/c ratio. This relationship may be different for larger or small size particles of rubber crumbs because, in the past studies, the particle size of rubber crumbs influences the compressive strength of RuC. This correlation should be used to care for different rubber crumbs contents or different w/c ratios that fall outside this study.

$$fcr = 52.65 - 1.16x \tag{1}$$

where f_{cr} is the compressive strength of RuC and $\boldsymbol{x}$ is the percentage of replacement level of rubber crumb.

5.11.2 Correlation Between Compressive Strength of RuC and Static MOE of RuC

After analyzing the results of RuC, the static MOE of RuC can be predicted by the following proposed relationship, which is denoted by Eq. 2. This proposed equation is based on experimental data. The relationship between the replacement level of the rubber crumb and the compressive strength of RuC is shown in Fig. 20b. The experimental data of the compressive strength of RuC of all mixes at 28 days of water curing and SMOE, shown in Table 7, are used to establish this correlation. Many curves are applied to these data, but a nonlinear fitting curve can represent the best-fit curve for the plotted data points for this correlation. The origin software performs this curve fitting. In the correlation analysis, the R^2 is obtained by 0.882, which means this correlation can predict accurate results, approximately near the experimental results. The scattering in the data from the tests on the specimens influences the accuracy of this correlation. This relationship is only applicable for rubber crumbs of size 0.6–2.36 mm and 0.38 w/c ratio. This relationship may be different for larger or small size particles of rubber crumbs because, in the past studies, the particle size of rubber crumbs influences the compressive strength and SMOE of RuC. The relationship between the compressive strength of RuC and the static MOE of RuC is shown in Fig. 20c and also compared with other researchers. The results are being compared with some more relationships given by IS 456-2000 [56] in Eq. 3, ACI 318-05 for structural calculation, applicable to normal weight concrete [57] in Eq. 4, Mensur et al. [58] in Eq. 5, Akhtaruzzaman and Hasnat [59] in Eq. 6, Rashid et al. [60] in Eq. 7, and Milicević et al. [61] in Eq. 8. All the following equations are given based on experimental data:

$$E_c = 1.161 fc^{0.807}\ (\text{GPa}) \tag{2}$$

$$E_c = 5000 fc^{0.5}\ (\text{MPa}) \tag{3}$$

$$E_c = 4.70 fc^{0.5}\ (\text{GPa}) \tag{4}$$

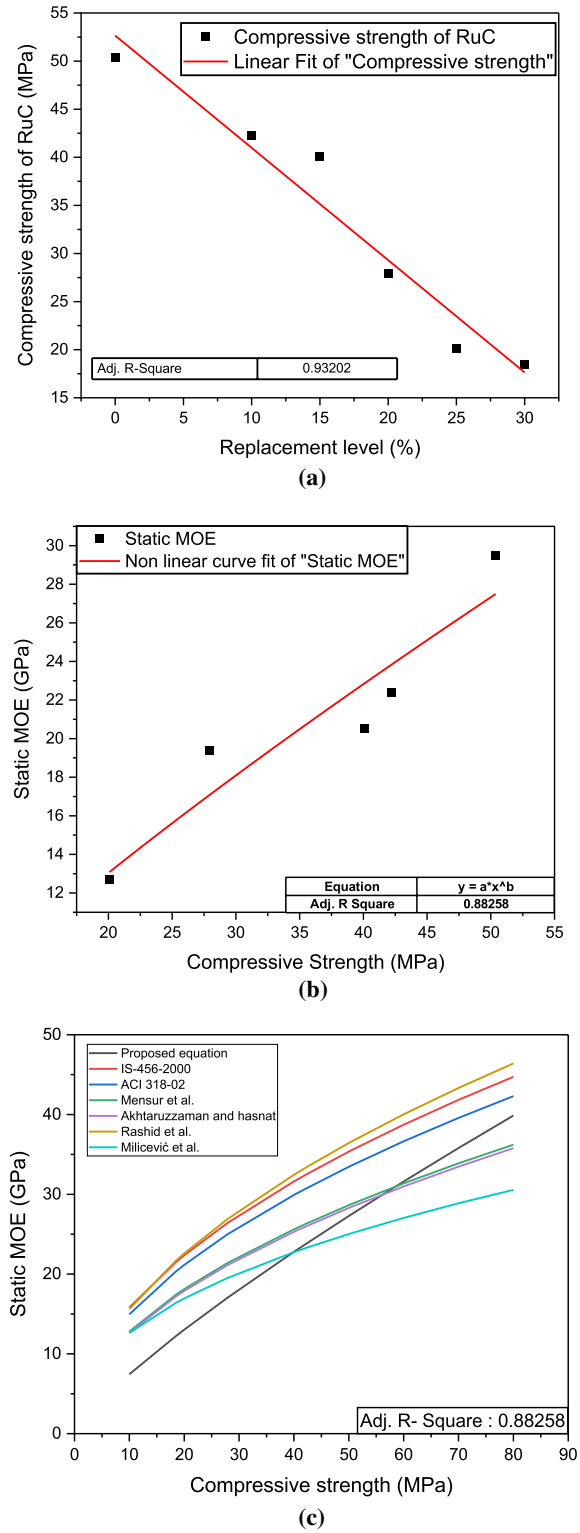


Fig. 20 **a** Correlation between replacement level of rubber crumb and compressive strength of RuC. **b** Correlation between compressive strength and static MOE of RuC. **c** Comparison of correlations between compressive strength of RuC and static MOE

$$E_c = 4050 fc^{0.5}\ (\text{MPa}) \tag{5}$$

$$E_c = 40000 \boldsymbol{fc}^{0.5}\ (\text{Psi}) \tag{6}$$

$$E_c = 5324 \boldsymbol{fc}^{0.5} - 1218\ (\text{MPa}) \tag{7}$$

$$E_c = 4735.7 fc^{0.4255}\ (\text{MPa}) \tag{8}$$

where fc is the compressive strength of concrete.

6 Conclusions

In this research, the fresh, hardened properties and microanalysis of concrete are evaluated. From the results and discussions, the following conclusions are listed:

1. The surface treatment of rubber crumbs by 15% H_2SO_4 solution minimizes the loss of RuC's strength and enhances the interfacial bonding between the rubber crumb and cement paste. Because of the elastic quality and flexibility, the presence of rubber crumb reduces the compressive strength of concrete. Rubber crumb as a partial replacement of FA can be up to 15% without significant compressive strength loss after surface treatment of rubber crumbs by H_2SO_4 solution.
2. The density of RuC decreases with the increasing replacement level and can be considered the semi-lightweight of concrete. The density of RuC decreases due to low Sp. Gr. of rubber crumb.
3. The concrete's fresh and mechanical properties decrease with increasing quantity of rubber crumb in concrete. The replacement level of the rubber crumb in the concrete should not be more than 15% of FA; otherwise, the decrement in strength will be high. When the cube specimen is fixed on both the upper and lower surfaces of the hydraulic machine, the rebound test should be done. This technique provides precise compressive strength results.
4. Replacement of FA by rubber crumb in concrete does not influence so much the workability of concrete, and the RuC can be produced with the same volume of SP as in ordinary concrete. There is no need to add more amount of SP.
5. From the graph of MOE and Poisson's ratio of regular and RuC, the RuC has an adequate deformability capacity, and it can be considered a positive gain in RuC.
6. The SEM analysis found that the rubber particles without surface treatment make a large gap between the rubber particle and concrete compared to the rubber particles with surface treatment. RuC with treated rubber

made less gap but did not obliterate the gap between rubber particle and concrete. Some more advanced method is required for surface treatment of rubber particles.

7. Two correlation equations are proposed to determine the compressive strength of RuC by the replacement level of rubber crumb, and MOE of RuC by the compressive strength of RuC, which are $f_{cr} = 52.65 - 1.16x$ (MPa) and $E_c = 1.161fc^{0.807}$ (GPa), respectively.

In the present study, rubber crumbs' surface is treated with 15% H_2SO_4 solution to minimize the concrete's strength loss and make a strong bond between rubber crumbs and cement compared to past research using rubber without surface treatment. After analyzing the results, it is found that the surface treatment of rubber with 15% H_2SO_4 solution reduces the reduction in mechanical properties and makes a strong bond compared to untreated rubber. Rubber crumb can be used in concrete up to 15% with the replacement of FA.

Acknowledgements This work is supported by the Civil Engineering Department, Delhi Technological University. Mr. Sandeep Mishra, Department of Applied Physics, helped in the XRD analysis related to the central instrumentation facility of the Delhi Technological University.

Funding The Delhi Technological University in Delhi, India, funded this research.

Availability of Data and Materials The available data have been used and discussed in the manuscript.

Declarations

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 agree with the publication.

Conflict of interest This article does not include any conflicts of interest.

References

1. By, O.; Khatib, Z.K.; Bayomy, F.M.: Rubberized portland cement concrete. J. Mater. Civ. Eng. **11**, 206–213 (1999)
2. Gesoğlu, M.; Guneyisi, E.: Strength development and chloride penetration in rubberized concretes with and without silica fume. Mater. Struct. **40**, 953–964 (2007)
3. Toutanji, H.A.: The use of rubber tire particles in concrete to replace mineral aggregates. Cem. Concr. Compos. **18**, 135–139 (1996). https://doi.org/10.1016/0958-9465(95)00010-0
4. Uygunoğlu, T.; Topçu, I.B.: The role of scrap rubber particles on the drying shrinkage and mechanical properties of self-consolidating mortars. Constr. Build. Mater. **24**, 1141–1150 (2010). https://doi.org/10.1016/j.conbuildmat.2009.12.027
5. Gupta, T.; Siddique, S.; Sharma, R.K.; Chaudhary, S.: Effect of aggressive environment on durability of concrete containing fibrous rubber shreds and silica fume. Struct. Concr. (2020). https://doi.org/10.1002/suco.202000043
6. Wang, R.; He, X.; Li, Y.: Evaluation of microcracks in the interfacial transition zone of recycled rubber concrete. Struct. Concr. **20**, 1684–1694 (2019). https://doi.org/10.1002/suco.201900044
7. Aslani, F.; Gedeon, R.: Experimental investigation into the properties of self-compacting rubberised concrete incorporating polypropylene and steel fibers. Struct. Concr. **20**, 267–281 (2019). https://doi.org/10.1002/suco.201800182
8. Verma, M.; Dev, N.: Sodium hydroxide effect on the mechanical properties of fly ash-slag based geopolymer concrete. Struct. Concr. (2020). https://doi.org/10.1002/suco.202000068
9. He, L.; Cai, H.; Huang, Y.; Ma, Y.; Van Den Bergh, W.; Gaspar, L.; Valentin, J.; Vasiliev, Y.E.; Kowalski, K.J.; Zhang, J.: Research on the properties of rubber concrete containing surface-modified rubber powders. J. Build. Eng. **35**, 101991 (2021). https://doi.org/10.1016/j.jobe.2020.101991
10. Ismail, M.K.; Sherir, M.A.A.; Siad, H.; Hassan, A.A.A.; Lachemi, M.: Properties of self-consolidating engineered cementitious composite modified with rubber. J. Mater. Civ. Eng. (2018). https://doi.org/10.1061/(ASCE)MT.1943-5533.0002219
11. Kashani, A.; Ngo, T.D.; Hemachandra, P.; Hajimohammadi, A.: Effects of surface treatments of recycled tyre crumb on cement-rubber bonding in concrete composite foam. Constr. Build. Mater. **171**, 467–473 (2018). https://doi.org/10.1016/j.conbuildmat.2018.03.163
12. Gupta, T.; Chaudhary, S.; Sharma, R.K.: Mechanical and durability properties of waste rubber fiber concrete with and without silica fume. J. Clean. Prod. **112**, 702–711 (2016). https://doi.org/10.1016/j.jclepro.2015.07.081
13. Gupta, T.; Chaudhary, S.; Sharma, R.K.: Assessment of mechanical and durability properties of concrete containing waste rubber tire as fine aggregate. Constr. Build. Mater. **73**, 562–574 (2014). https://doi.org/10.1016/j.conbuildmat.2014.09.102
14. He, L.; Ma, Y.; Liu, Q.; Mu, Y.: Surface modification of crumb rubber and its influence on the mechanical properties of rubber-cement concrete. Constr. Build. Mater. **120**, 403–407 (2016). https://doi.org/10.1016/j.conbuildmat.2016.05.025
15. Raffoul, S.; Garcia, R.; Pilakoutas, K.; Guadagnini, M.; Medina, N.F.: Optimisation of rubberised concrete with high rubber content: an experimental investigation. Constr. Build. Mater. **124**, 391–404 (2016). https://doi.org/10.1016/j.conbuildmat.2016.07.054
16. Ganjian, E.; Khorami, M.; Maghsoudi, A.A.: Scrap-tyre-rubber replacement for aggregate and filler in concrete. Constr. Build. Mater. **23**, 1828–1836 (2009). https://doi.org/10.1016/j.conbuildmat.2008.09.020
17. Tian, S.; Zhang, T.; Li, Y.: Research on modifier and modified process for rubber-particle used in rubberized concrete for road. Adv. Mater. Res. **243–249**, 4125–4130 (2011). https://doi.org/10.4028/www.scientific.net/AMR.243-249.4125
18. Li, G.; Stubblefield, M.A.; Garrick, G.; Eggers, J.; Abadie, C.; Huang, B.: Development of waste tire modified concrete. Cem. Concr. Res. **34**, 2283–2289 (2004). https://doi.org/10.1016/j.cemconres.2004.04.013
19. Richardson, A.E.; Coventry, K.A.; Ward, G.: Freeze/thaw protection of concrete with optimum rubber crumb content. J. Clean. Prod. **23**, 96–103 (2012). https://doi.org/10.1016/j.jclepro.2011.10.013
20. Tarry, S.R.: Effect of partial replacement of coarse aggregates in concrete by untreated and treated tyre rubber aggregates. Int. J. Adv. Sci. Res. **3**, 65–69 (2018)
21. IS: 4031 (Part 6): Methods of Physical Tests For Hydraulic Cement Part 6 Determination of Compressive Strength of Hydraulic Cement

Other Than Masonry Cement (First Revision). Bur. Indian Stand. Delhi. 1–3 (2005)
22. IS 4031- Part IV: Methods of physical tests for hydraulic cement. Part IV- Determination of consistency of standard cement paste. Bur. Indian Stand. New Delhi. Reaffirmed in 2005 (1988)
23. IS 4031 (Part 11) - 1988: METHODS OF PHYSICAL TESTS FOR HYDRAULIC CEMENT 4031 (Part 11) - 1988. Bur. Indian Stand. New Delhi (2005)
24. IS 4031- Part V: Methods of physical tests for hydraulic cement. Part V- Determination of initial and final setting times. Bur. Indian Stand. New Delhi. Reaffirmed in 2005 (1988)
25. IS 4031 - 3 (1988) (Reaffirmed 2005): Methods of Physical Tests for Hydraulic Cement, Part 3: Determination of Soundness [CED 2: Civil Engineering]. Bur. Indian Stand. New Delhi. 1–10 (2005)
26. C1365-18: Standard Test Method for Determination of the Proportion of Phases in Portland Cement and Portland-Cement Clinker Using X-Ray Powder. ASTM Int. 1–11 (1998). https://doi.org/10.1520/C1365-18.2
27. Bureau of Indian Standards: IS 383: 1970 Specification for Coarse and Fine Aggregates From Natural Sources for Concrete. IS 383 1970(Reaffirmed 2002). 1–24 (1970)
28. IS 2386- Part III: Method of Test for aggregate for concrete. Part III- Specific gravity, density, voids, absorption and bulking. Bur. Indian Stand. New Delhi. (Reaffirmed 2002) (1963)
29. Bureau of Indian Standards (BIS): IS : 2366 (Part IV)-1963-Methods of test for Aggregates for Concrete, part 4 : Mechanical properties. Indian Stand. 1–37 (2002)
30. M.S. Shetty: Concrete Technology: Theory and Practice. S.Chand, New Delhi (2013)
31. RANKEM (A AVANTOR BRAND): Sulphuric Acid, https://in.vwr.com/store/product/27465747/sulphuric-acid-98-00-ar-rankem (1889)
32. Ma, Q.W.; Yue, J.C.: Effect on mechanical properties of rubberized concrete due to pretreatment of waste tire rubber with NaOH. Appl. Mech. Mater. **357–360**, 897–904 (2013). https://doi.org/10.4028/www.scientific.net/AMM.357-360.897
33. D934-13: Identification of Crystalline Compounds in Water-Formed Deposits by X-Ray Diffraction. ASTM Int. 11, 1–10 (2014). https://doi.org/10.1520/D0934-13.2
34. ASTM-E2550: Standard Test Method for Thermal Stability by Thermogravimetry. ASTM Int. i, 1–6 (2017). https://doi.org/10.1520/E2550-17.Development
35. Zhu, H.; Rong, B.; Xie, R.; Yang, Z.: Experimental investigation on the floating of rubber particles of crumb rubber concrete. Constr. Build. Mater. **164**, 644–654 (2018). https://doi.org/10.1016/j.conbuildmat.2018.01.001
36. IS: 1199 - 1959 Reaffirmed: Methods of sampling and analysis of concrete. Indian Stand (2004)
37. IS: 10086-1982 Reaffirmed: Specification for moulds for use in tests of cement and concrete. Indian Stand (2008)
38. IS 516:1959: Method of Tests for Strength of Concrete. IS 516 - 1959 (Reaffirmed 2004). New Delhi, India (2004)
39. Najim, K.B.; Hall, M.R.: Mechanical and dynamic properties of self-compacting crumb rubber modified concrete. Constr. Build. Mater. **27**, 521–530 (2012). https://doi.org/10.1016/j.conbuildmat.2011.07.013
40. Marie, I.: Thermal conductivity of hybrid recycled aggregate—Rubberized concrete. Constr. Build. Mater. **133**, 516–524 (2017). https://doi.org/10.1016/j.conbuildmat.2016.12.113
41. Batayneh, M.K.; Marie, I.; Asi, I.: Promoting the use of crumb rubber concrete in developing countries. Waste Manag. **28**, 2171–2176 (2008). https://doi.org/10.1016/j.wasman.2007.09.035
42. Taha, M.M.R.; Asce, M.; El-wahab, M.A.A.: Mechanical. Fracture, and Microstructural Investigations. **20**, 640–649 (2009)
43. Bheel, N.; Memon, F.A.; Meghwar, S.L.: Study of fresh and hardened properties of concrete using cement with modified blend of millet husk ash as secondary cementitious material. SILICON (2020). https://doi.org/10.1007/s12633-020-00794-7
44. Xue, G.; Pei, Z.: Experimental study on axial compressive properties of rubber concrete at low temperature. J. Mater. Civ. Eng. **30**, 1–8 (2018). https://doi.org/10.1061/(ASCE)MT.1943-5533.0002178
45. Liu, F.; Zheng, W.; Li, L.; Feng, W.; Ning, G.: Mechanical and fatigue performance of rubber concrete. Constr. Build. Mater. **47**, 711–719 (2013). https://doi.org/10.1016/j.conbuildmat.2013.05.055
46. Gurunandan, M.; Phalgun, M.; Raghavendra, T.; Udayashankar, B.C.: Mechanical and damping properties of rubberized concrete containing polyester fibers. J. Mater. Civ. Eng. **31**, 1–10 (2019). https://doi.org/10.1061/(ASCE)MT.1943-5533.0002614
47. Su, H.; Yang, J.; Ling, T.C.; Ghataora, G.S.; Dirar, S.: Properties of concrete prepared with waste tyre rubber particles of uniform and varying sizes. J. Clean. Prod. **91**, 288–296 (2015). https://doi.org/10.1016/j.jclepro.2014.12.022
48. Gupta, T.; Tiwari, A.; Siddique, S.; Sharma, R.K.; Chaudhary, S.: Response assessment under dynamic loading and microstructural investigations of rubberized concrete. J. Mater. Civ. Eng. **29**, 1–15 (2017). https://doi.org/10.1061/(ASCE)MT.1943-5533.0001905
49. IS 13311 part 2 - 1992 Reaffirmed: IS 13311-2 (1992): Method of non-destructive testing of concrete-methods of test, Part 2: Rebound hammer. Indian Stand. (2004)
50. IS 13311 Part 1 - 1992 Reaffirmed: Method of Non-destructive testing of concrete, Part 1: Ultrasonic pulse velocity. Indian Stand. 1–7 (2004)
51. Wu, Y.F.; Kazmi, S.M.S.; Munir, M.J.; Zhou, Y.; Xing, F.: Effect of compression casting method on the compressive strength, elastic modulus and microstructure of rubber concrete. J. Clean. Prod. **264**, 121746 (2020). https://doi.org/10.1016/j.jclepro.2020.121746
52. Khaloo, A.R.; Dehestani, M.; Rahmatabadi, P.: Mechanical properties of concrete containing a high volume of tire-rubber particles. Waste Manag. **28**, 2472–2482 (2008). https://doi.org/10.1016/j.wasman.2008.01.015
53. Goulias, D.G.; Ali, A.H.: Evaluation of rubber-filled concrete and correlation between destructive and nondestructive testing results. Cem. Concr. Aggregates. **20**, 140–144 (1998). https://doi.org/10.1520/cca10447j
54. Xue, J.; Shinozuka, M.: Rubberized concrete: A green structural material with enhanced energy-dissipation capability. Constr. Build. Mater. **42**, 196–204 (2013). https://doi.org/10.1016/j.conbuildmat.2013.01.005
55. Najim, K.B.; Hall, M.R.: Crumb rubber aggregate coatings/pretreatments and their effects on interfacial bonding, air entrapment and fracture toughness in self-compacting rubberised concrete (SCRC). Mater. Struct. Constr. **46**, 2029–2043 (2013). https://doi.org/10.1617/s11527-013-0034-4
56. IS 456 - 2000: Indian Standard Plain and Reinforced Concrete - Code of Practice, IS-456:2000. Bur. Indian Stand. New Delhi. New Delhi,India (2000)
57. A.M. Neville: concrete technology, (2013)
58. Mansur, M.A.; Wee, T.H.; Cheran, L.S.: Crushed bricks as coarse aggregate for concrete. ACI Mater. J. **96**, 478–484 (1999). https://doi.org/10.14359/649
59. Akhtaruzzaman, A.A. and hasnat A.: Properties of concrete using crushed brick as aggregate, Concrete International (1983)
60. Rashid, M.A., Salam, M.A., Kumar Shill, S., Hasan, M.K.: Effect of replacing natural coarse aggregate by brick aggregate on the properties of concrete. Dhaka Univ. Eng. Technol. Gazipur. 1, (2012)
61. Miličević, I.; Štirmer, N.; Bjegović, D.: Relation between the compressive strength and modulus of elasticity of concrete with crushed brick and roof tile aggregates. Struct. Concr. **18**, 366–375 (2017). https://doi.org/10.1002/suco.201500207

Multi-modal Fusion using Fine-tuned Self-attention and Transfer Learning for Veracity Analysis of Web Information

Priyanka Meel[1], Dinesh Kumar Vishwakarma[2]
Biometric Research Laboratory, Department of Information Technology
Delhi Technological University, New Delhi, India-110042

Abstract

The nuisance of misinformation and fake news has escalated many folds since the advent of online social networks. Human consciousness and decision-making capabilities are negatively influenced by manipulated, fabricated, biased or unverified news posts. Therefore, there is a high demand for designing veracity analysis systems to detect fake information contents in multiple data modalities. In an attempt to find a sophisticated solution to this critical issue, we proposed an architecture to consider both the textual and visual attributes of the data. After the data pre-processing is done, text and image features are extracted from the training data using separate deep learning models. Feature extraction from text is done using BERT and ALBERT language models that leverage the benefits of bidirectional training of transformers using a deep self-attention mechanism. The Inception-ResNet-v2 deep neural network model is employed for image data to perform the task. The proposed framework focused on two independent multi-modal fusion architectures of BERT and Inception-ResNet-v2 as well as ALBERT and Inception-ResNet-v2. Multi-modal fusion of textual and visual branches is extensively experimented and analysed using concatenation of feature vectors and weighted averaging of probabilities named as Early Fusion and Late Fusion respectively. Three publicly available broadly accepted datasets All Data, Weibo and MediaEval 2016 that incorporates English news articles, Chinese news articles, and Tweets correspondingly are used so that our designed framework's outcomes can be properly tested and compared with previous notable work in the domain. The accuracy varies on different datasets using different fusion methods. The recommended multi-modal technique accomplishes 97.19 % highest accuracy on All Data dataset using ALBERT & Inception-ResNet-v2 late fusion framework, which is notably better than all the previous works in the domain of veracity and credibility assessment of online information contents.

Keywords: Deep Neural Network, Self-attention, Early Fusion, Late Fusion, Transformer, Multimodal, Veracity.

1. Introduction

In the present era, the accessibility and readiness of the internet have made it possible to get real-time news for people worldwide. Nevertheless, these ongoing developments in technology and the distribution of information through various networks, including social media have provocatively amplified the spread of rumors and fake news. The real problem is not disseminating phony details but the fact that people tend to accept whatever is fed to them through these online platforms. Hence, information pollution is one of the most prominent intimidations to ethical journalism, the autonomy of expression, and in a way, the fundamentals of democracy. Just because of the utter difficulty of identifying rumors, they disseminate like wildfire on online platforms, increasing the issue's severeness multiple folds. It is comparatively easy to identify fake service and product reviews by their tenor and intent; fake news can be highly elusive and deceptive. So obviously, it would be pretty naive to expect an average reader, who doesn't always have the time, energy or even the resources, to perform the tricky task of identifying fake news. The explosive growth and deadly repercussions of fake news have made it inevitable to find solutions that can wipe this problem out once and for all.

Though publishing and spreading false information is not something new, the usage of the term "fake news" is. It was first used only at the end of the 19th century in a newspaper, Cincinnati Commercial Tribune in the United States (US) [1]. Since then, many researchers and scholars [2], [3], [4] have tried to outline it in different ways such as: misinformation, disinformation, clickbait, hoax, propaganda, satire, fake news, rumor etc. But what is common in all their explanations is that fake news is a portion of the information presented to be accurate and real, but in reality, it has no facts to support it. In more cases than not, it is spread intentionally to achieve some hidden personal agenda. While many of these fabricated tales are not intended to be harmful and take the form of funny jokes and satire, others indeed have more disturbing implications. Because the accountability factor is missing, anyone in any part of the world can claim whatever he feels like without fearing the consequences, which practically don't exist. This presents a long list of problems, with the most critical of them being the damage to public mental health.

According to Dr. Chrysalis Wright, Associate Lecturer of Psychology at the University of Central Florida [5], the most lethal aspect of fake news is that it makes us only see one aspect of a story, which in fact may be completely fabricated. But it's so captivating that we don't

always care about the complete picture. We get upset and emotional, making us want to share it with our friends and family. That's how the cycle begins. In no time, a small piece of fake news is getting circulated like wildfire worldwide.

Many people believe that the increase in fake news is because of the declining popularity of newspapers and other traditional media sources. The journalist's and social media blogger's carefree attitude of sharing stories without verifying them first adds severity to the issue multiple folds. The erosion of trust in traditional news sources created a big void or an opportunity for the fake news spreaders to unfortunately take advantage. Technological advancements have made it possible to spread misleading stories in every corner of the world. Fake news is being read, processed and forwarded before a person even gives a second to verify its authenticity. Fig. 1 highlights some of the widely circulated fake stories in recent times.



Figure 1: Instances of fake news on web platforms in multiple data modalities

On the surface, it might seem that fake news doesn't have any extreme consequences. But in reality, it has the potential to ruin the lives of people permanently. One such example is the 2017 terror outbreak in London [6] that happened on 22nd March 2017. Just hours after half a dozen people were killed and half a century got injured, an image of a middle-aged lady wearing a hijab and speaking on the phone on Westminster Bridge, the site of the attack, got viral all over the internet. Thousands of people shared it on their profiles claiming the woman, as a Muslim, was unsympathetic to the sufferings of fatalities around her, with the hashtag "#BanIslam" that got trending everywhere as in Fig 1(a). Later, the account which first tweeted the image was suspended after it was found to be a Russian bot and the image was taken off

from the internet. But the mental damage it caused to the woman could never be undone. She issued a statement after being devastated by the negative attention she encountered due to this fake news and spoke about how this news has ruined her identity in the community in which she lives. That's how poisonous a small piece of fake news can be. It can spoil the lives of innocent people in a matter of hours.

During coronavirus lockdown, social media was flooded with animal news, healing nature claims and the images of dolphins in canals of Venice as in Fig 1(b). Sadly enough, the viral images were not of Venice canals, they are from a different region of Italy. But the question that arises is that why is fake news created in the first place. What could a person possibly gain from it? The most common and apparent reason is money. Fake news pays well. On various social media sites, the posts that trend and gather a higher number of views generate revenue. And what better way to get a viral post than to create something controversial and buzzy, something that can be catchy enough to garner a lot of viewers. And it's not only the person who is being targeted in the fake news but also the money that gets paid to the creator comes out of our pockets. Another widespread motive to generate fake news is political propaganda. Often during the elections, a sudden surge in the spread of fake news is observed. One case in point being the 2016 US elections, which got heavily swayed by the impact of fake news [7], [8]. Some political experts even claim that Donald Trump won't have won the elections if it weren't for the massive impact of fake news.

It is well-established by now that fake news is a grave issue that needs to be uprooted from its core. But the task of eradicating it is not as simple as it might look. Some of the significant challenges need to tackle to classify fake news from a real one can be listed as follows:

- The lack of a standard definition of fake news makes it tough to classify one. There's no exact way to know that where we should draw the line. It is a significant concern since it blurs the boundary between real and fake.
- Fake news spreads at lightning speed. By the time a piece of information is doubted for its authenticity, it has already reached thousands of people, feeding them with false claims and rumours.
- A fake news article's tone is not as explicit and direct as in a fake review or recommendation. Since the intent is so subtle and hidden, finding an alternate way of identifying fake news gets tricky.
- Sometimes, a piece of information is not entirely true or false. Some fake news creators mix

fake elements in true stories to misguide and influence people. This makes the whole task extremely complex since the information is not entirely fake but only to an extent.

Multimedia has become an integral part of human life as it has more conclusive, convincing and long-lasting memory effects. Almost every circulating news story is strengthened with accompanying images or videos. Multiple data formats have made forgery identification in online circulated news articles quite complicated because each data format has different characteristics. As the inherent attributes of varied data formats differ considerably, so as the techniques to detect their forgeries. Therefore, we propose the framework of multi-stream fusion to harness the independent benefits of the properties of textual and visual data streams, then merge them to tag the overall instance as real or fake. Transformer-based BERT [9] and ALBERT [10] language models, solely based on attention mechanism and bidirectionally trained to have a more profound sense of language context and flow, are used for textual feature extraction. Inception -Resnet-v2 [11] deep neural network is employed for training the classifier with the visual data. The effectiveness of multi-stream fusion is being explored by designing independent late and early fusion architectures. The work is expected to enrich the research community with the naïve framework of multi-modal fusion using self-attention and deep neural networks. The key contributions of the work are as follows:

- The problem of fraudulent content in text and image multi-modal data format is being identified and addressed for veracity analysis of web information contents.
- Textual feature extraction and model training is being done using BERT and ALBERT bidirectional self-attention-based transformers.
- Visual data is being analysed by harnessing transfer learning capabilities using fine-tuned Inception-ResNet-v2 deep neural network architecture.
- The proposed framework focused on two independent multi-modal fusion frameworks of BERT and Inception-ResNet-V2 as well as ALBERT and Inception-ResNet-V2.
- The multi-modal fusion of textual and visual branches is extensively experimented and analysed using concatenation of feature vectors and weighted averaging of probabilities named Early Fusion and Late Fusion respectively.
- Three publicly available broadly accepted datasets All Data, Weibo and MediaEval 2016 that incorporate English news articles, Chinese news articles, and Tweets correspondingly are used to test and compare our designed framework's outcomes with previous notable work in the domain.

The work is explained in its entirety under five sections and multiple subsections. Section 1 emphasis on the need of addressing the information pollution issue with few widely circulated fake news examples. Literature survey and research gap is being highlighted in section 2. Section 3 details in the technical description of the proposed architecture and its different components. Experimental settings, datasets, system specification, detailed analysis of obtained results and comparative study is being done in section 4. Eventually, section 5 summarizes the work and recommends the prospective directions for upcoming research. Our work's central objective is to propose an effective and efficient framework that can correctly classify fake news. It leverages advanced deep learning models to accomplish this task using both text and image features of the news data. We hope that this work will prove to be a landmark within the province of veracity analysis and encourage other researchers to propose further advancements that would take the scope of the solution to new heights.

2. Related Works

Information circulating online on web platforms has various attributes like headline/title, text/body, author, associated image, URL, etc. Any changes made in any of the features incline news to behave abnormally, making it a piece of fake news or a rumour. Researchers have tried to verify different attributes, webpage URL, author, text, associated image, time series analysis and propagation statistics to design models for veracity analysis. False contents, also popularly termed as information pollution or infodemic, are amplifying at an alarming rate. Efforts have already started for providing solutions to detect and mitigate fraudulent content using up-to-date artificial intelligence technologies. A considerable number of methodologies embrace fake news uncovering textual data. With the rise of multimedia data on users' posts and news, research incorporating the identification of visual fake news has increased and improved the preciseness of algorithms.

During the COVID-19 pandemic, the circulation of fake news or rumors is being highly observed. In an article titled "**Fake News, Real Arrests**" [12] , the author categorically emphasized the increase in the number of arrests by police due to the spread of fake news across different states of India. It has also been quoted that "**Virus of fake news spreads faster than Corona Virus itself** ". Social media is the most accessible tool to spread fake news rapidly and bots are being programmed to make the task easier. Existing literature enumerates several criteria for classifying the veracity analysis approaches. We organize varied methods available in literature according to the feature being analyzed to verify the truthfulness of the information. The following Fig 2 represents the categories of organization in our work.

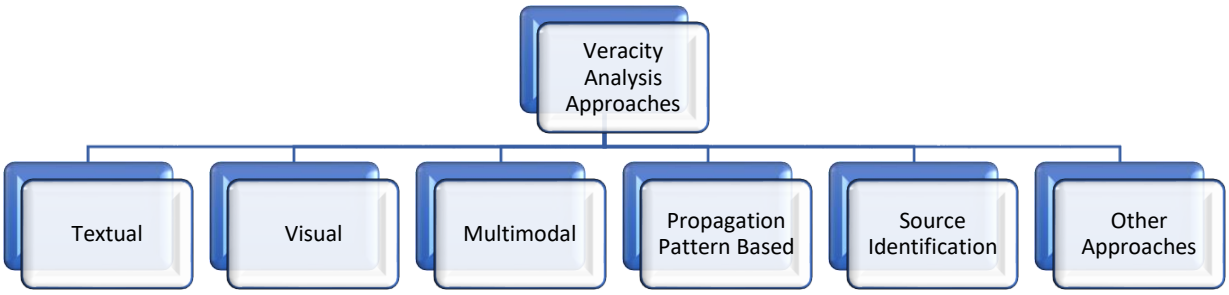


Figure 2: Classification of Veracity Analysis Approaches

2.1 Text-Based Approaches

Many researchers have exploited features from textual contents like textual frequency-based features, semantic features, sentiment analysis, writing style, text pattern, the polarity of contents etc. Pan et al. [13] have studied how the style of writing is responsible for shaping the views of the people. They used a unique structure called knowledge graphs to categorize given news text as fake or real. Gautam and Jerripothula [14] put forward a novel framework for text fake news detection using Random Forest classifier that leverages the capabilities of paraphrasing tool Spinbot, grammar checking tool Grammarly and Glove word-embedding for feature extraction. Li and Zhou [15] employ BERT language model for binary text fake news detection by connecting the dots between the claim and pieces of evidence. Lengthy text news articles are first summarized to extract the claims as well as keywords that are subsequently searched on web to extract the related articles to treat then as evidences or facts to be matched against the claim. Finally, the scores of similarities are fed into a fact verification model to classify a news article as real or fake. To detect ambiguous information on COVID 19 Elhadad et al. [16] collected factual data from WHO, UNICEF, United Nations and epidemiological data from different fact check websites to consolidate into a repository. This accumulated factual data is used to build a fake news classification system based on ensemble of ten different machine learning classifiers using seven feature extraction techniques implemented and tested against twelve performance metrices.

2.2 Visual Approaches

Images are an integral part of print, electronic and digital media news. Visual snapshots alone or mostly combined with text overwritten them when manipulated spread forgery to a great extent. A novel method is proposed [17] to detect photographic splicing detection via

illumination inconsistencies and deep learning. This approach is capable enough of locating forged regions, eliminates the laborious feature engineering process and provides outperforming accuracies on the same datasets paralleled to state-of-the-arts. Elkasrawi et al. [18] devised a semi-automatic approach of verifying the credibility of images in online articles by means of a two phased process embraces clustering and hierarchical image feature analysis, accounts to highest 88% accuracy for a dataset of 50 images. Fake images generated by Generative Adversarial Networks (GANs) by changing context and semantics of original image via. Image-to-image translation looks very realistic. The study proposed by Marra et al. [19] performed on a dataset of 36302 images describes the performance of multiple image forgery detectors both in standard conditions and under in the influence of compression for GAN generated fake images. Jin et al. [20] designed a novel method based on visual and statistical image features implemented on real-world Sina Weibo dataset for microblog news verification.

2.3 Multi-modal Approaches

The modalities of text and image, when used in conjunction, mislead people by spreading fake news enormously. Although there are not many standard open-source datasets available for multi-modal fake news classification, having text and images, but few researchers have utilized the rarely available datasets for the task. Yang et al. [21] used explicit and implicit features from text and images using Convolutional Neural Networks (CNNs) for counterfeit news detection. Wang et al. [22] have created an end-to-end model that uses an event discriminator to classify fake news exploiting modalities of text as well as image. Their model uses CNN for text and VGG-19 to elicit photographic features and concatenating them to discriminate events and classify false news. Inspired by EANN, Khattar et al. [23] proposed a comparable framework using bi-LSTM instead of Text-CNN and formulated an architecture of auto-encoder and decoder. Same latent vectors were used for both encoder and decoder. Along with textual sentiment analysis and image segmentation process, Shah and Kobti [24] implemented cultural algorithm with situational and normative knowledge for multi-modal feature extraction. The final feature vector is passed through SVM classifier layer for veracity analysis experimented on Weibo and Twitter datasets. A neural network based multi-modal fake news detection system is proposed by Giachanou et al. [25] that combines textual, visual and sematic information being implemented on MediaEval , Politifact and GossipCop datasets. Vishwakarma et al. [26] processed multi-modal data in a different format by extracting the text written on images and then web scrapping this text claim in order to a reality parameter based on top 15 google search results. The value of the calculated reality

parameter is compared against a threshold, which if exceeds a set point, the information is categorized as real otherwise fake. Meel and Vishwakarma [27] researched a framework for multi-modal fake news detection based on ensemble of Hierarchical Attention Network, Image Captioning and Forensics.

2.4 Propagation Pattern Analysis

The way rumour and fake news diffuse in a network is also a prominent feature of analysis and mitigation of misinformation. Non-sequential propagation structure of microblog posts is being identified by Ma et al. [28] to acquire discriminative attributes for generating powerful top-down and bottom-up representations using Recursive Neural Networks for rumour identification. They also proposed a kernel-based method of rumour detection termed as propagation tree kernel [29] to capture higher-order features for segregating various types of rumours based on the structure of their dissemination trees. Vu and Jung [30] proposed a propagation graph embedding method for rumour detection based on Graph Convolutional Neural Network. Experimental results illustrate the efficacy of their suggested method by reducing the detection error up to 10% as equated with state-of-the-arts. Other studies of combining attention mechanism with propagation graph structure [31] , an ensemble of user representation learning with news propagation dynamics [32] , implications of searchability on rumour self-correction [33] are also worth considering.

2.5 Source Identification

Propagation of fake news can be stopped up to an extent by identifying the source. Researchers have formulated various approaches to identify fake accounts on online platforms. In [34], authors have suggested a solution to discriminate fake Twitter accounts from genuine ones using dimensionality reduction and feature selection techniques. The authors in [35] use basic machine learning algorithms to classify fake accounts on the social media platform Instagram. It is not always viable to stop the propagation of fake news by recognizing and blocking the fake accounts because a user can create multiple accounts on an online platform or sometimes real accounts could also be a source of fake news. Louni and Subbalakshmi [36] addressed the problem of finding the origin of rumor in large scale social network. The salient feature of their proposed algorithm is probabilistically varying internode relationship strengths achieved by assigning random non-homogeneous edge weights to the original network graph. Query response method [37] using simple batch querying and interactive querying with directions is also out of the box analytical approach for rumor source detection.

2.6 Other Approaches

Authors in [38] propose a browser extension “BRENDA” implemented with Google Chrome browser for automated credibility assessment of claims using deep neural network. This automatic process is fast, real time, produces the evidence along with the classification and does not need to leave the web page. A Clickbait Video Detector (CVD) is designed by Varshney and Vishwakarma [39] to detect the clickbait videos circulating on web based on cognitive evidences extracted out of user profile, video content and human consensus. To extract the hidden patterns from unlabeled data a semi-supervised approach based on the temporal ensembling method is proposed by Meel and Vishwakarma [40] . Analysis of misinformation detection under varying time constraints under the consideration of three of different types of attacks: Evasion attacks, Poison attacks, Blocking attacks is thoroughly studied by Horne et al. [41].

2.7 Motivation and Research Gap

The detailed literature survey done in previous section and shortcomings of the existing systems encourages us to fuse transformer-based self-attention and fine-tuned transfer learning into a combined framework for veracity analysis of multi-modal online information. The key highlights of motivation and research gap are as follows:

- The main advantage of transformer models is that they are non-sequential, meaning that they don’t require that the input sequence be processed in the order. This innovation allows transformers to be parallelized and scaled much more than previous NLP models.
- The Transformer supports multiple folds of parallelization and can achieve a new benchmark in terms of quality and performance.
- Transformer-based approaches BERT & ALBERT architecture are solely based on self-attention also called intra attention mechanism dispensing with recurrence and convolutions completely.
- Attention mechanism without recurrence and convolutions allows to draw global dependencies between input and output which removes all restrictions of sequential processing and motivates more of exploring the benefits of parallelization.
- BERT & ALBERT are bidirectionally trained language model supports a higher level of understanding about language structure and flow than single directional language model.

- Inception-ResNet-v2 pre-trained on ImageNet is being further fine-tuned for our veracity analysis task to harness the advantages of transfer learning which is more suitable for smaller datasets also and saves a lot of time of training the model from scratch.

- Motivation of using transfer learning is that the pre-trained models can be scaled for a variety of application specific tasks just by adding few final layers and fine tuning the weights as well as adjusting the hyperparameters on a comparatively smaller labelled dataset.

- Less work is reported in literature to address multi-modal information authenticity issue that will take the benefits of latest language modelling and transfer learning technologies.

- This work is expected to prove a landmark within the province of veracity analysis and encourage other researchers to propose further advancements to counter multiple visual and textual forgery formats that would take the scope of the solution to new heights.

2.8 Problem Statement

To better explain our work, we can mathematically formulate the problem statement as follows:

Let D is a set of training instances and n is the size of training dataset of news items.

$D: \{x_1, x_2, \ldots \ldots, x_n\}$

Each item, x_i in set D contains the information $x_i: \{t_i, b_i, im_i, l_i\}$. Where,

$t_i \equiv title\ of\ the\ news\ item, x_i$

$b_i \equiv body\ of\ the\ news\ item, x_i$

$im_i \equiv image\ of\ the\ news\ item, x_i$

$l_i \equiv label\ of\ the\ news\ item, x_i$

$and\ , l_i \epsilon \{fake, real\}$

The prime aim is to predict the label l_j of a news item $x_j: \{t_j, b_j, im_j\}$.

3. Proposed Multi-modal Architecture

It has become significant to incorporate images, as nearly all news, fake or real is accompanied by some visual data. A combination of Deep Learning Neural Network Models is used for veracity analysis of multi-modal data. We have used two modalities of textual and visual features in conjunction to classify a piece of news as fake or real. We use implicit features obtained from pre-trained models to train our custom neural network to obtain the final predictions using two independent architectures of Late Fusion and Early Fusion. We have

used Inception-ResNet-v2 to extract visual features. The models BERT and ALBERT have been used to elicit textual attributes. Diverse forms of text input, like English articles, Chinese articles and Tweets have been used to make our model robust and usable across multiple platforms. The architecture of Multimodal Early Fusion and Late Fusion has been detailed in Figures 3 and 4 as well as the working steps are described in Algorithm 1 and 2 respectively. The frameworks and technical specifications used are defined in detail in the sub-sections below.

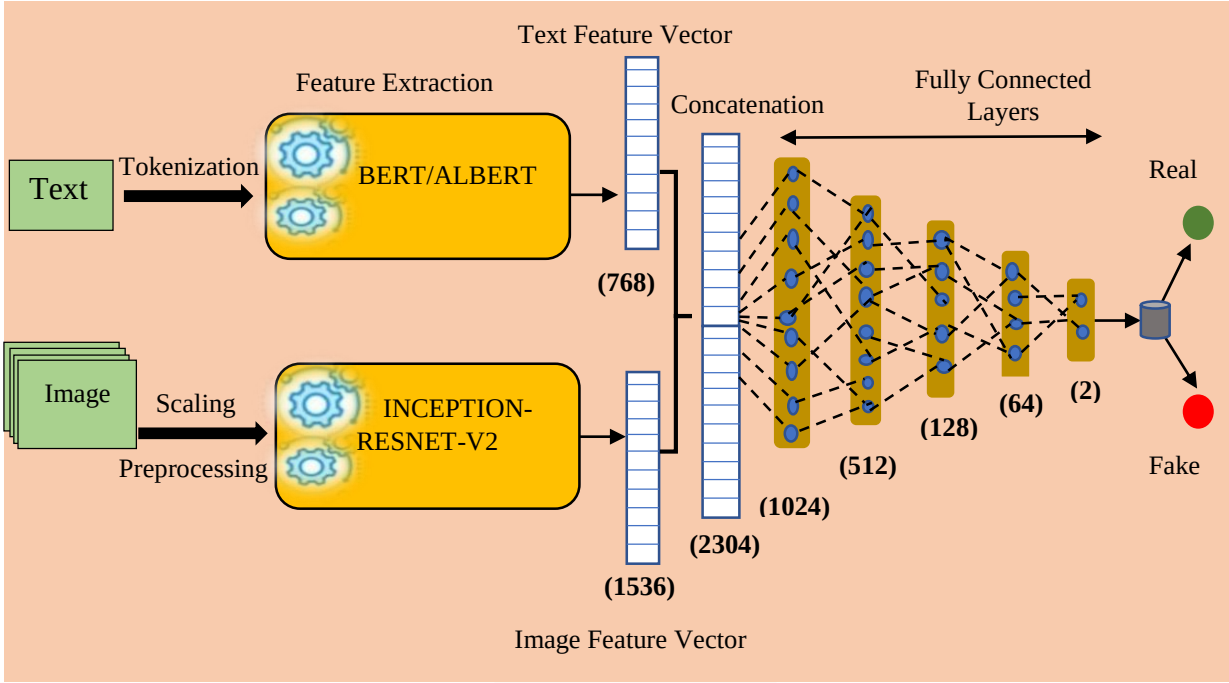


Figure 3: Architecture of Multimodal Early fusion

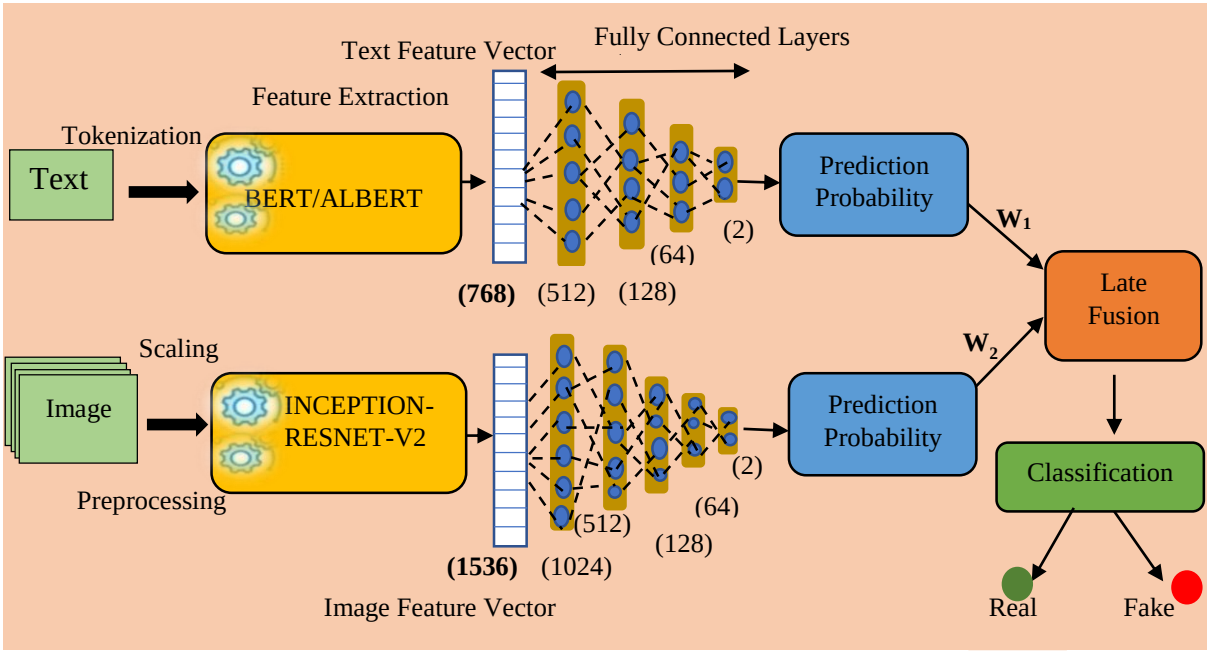


Figure 4: Architecture of Multimodal Late fusion

Algorithm 1: Multimodal Early Fusion

Parameter Initialization

- Input A = $\{a_1, a_2, \ldots a_n\}$ is set of text vectors, B= $\{b_1, b_2, \ldots b_n\}$ is set of corresponding images, C= $\{c_1, c_2, \ldots c_n\}$ is set of labels for A and B.
- n is the size of training set
- Split A, B and C into three subsets for 70% training, 10% validation and 20% testing.

1: **For 1 to 30 epochs do**
2: Extract feature vector using BERT/ALBERT model M_1 for text input A.
3: Extract feature vector using Inception-ResNet-v2 model M_2 for image input B.
4: Convert feature vectors obtained from M_1 and M_2 into unidimensional.
5: Perform concatenation of feature vectors.
6: Add a series of fully connected layers (dense, batch normalization, relu, dropout=0.4)
7: Apply binary sigmoid classifier and calculate prediction probabilities
8: Apply binary cross-entropy loss and Adam optimizer
9: **end for**
10: Evaluate the performance on test set

Algorithm 2: Multimodal Late Fusion

Parameter Initialization

- Input A = $\{a_1, a_2, \ldots a_n\}$ is set of text vectors, B= $\{b_1, b_2, \ldots b_n\}$ is set of corresponding images, C= $\{c_1, c_2, \ldots c_n\}$ is set of labels for A and B.
- n is the size of training set
- Split A, B and C into three subsets for 70% training, 10% validation and 20% testing.

1: **For 1 to 30 epochs do**
2: Extract feature vector using BERT/ALBERT model M_1 for text input A
3: Add fully connected layers (dense, batch normalization, relu, dropout=0.4) to text feature vector
4: Apply binary sigmoid classifier and calculate prediction probability p_1
3: Extract feature vector using Inception-ResNet-v2 model M_2 for image input B
4: Add fully connected layers (dense, batch normalization, relu, dropout=0.4) to image feature vector
5: Apply binary sigmoid classifier and calculate prediction probability p_2
6: Perform final prediction using weighted averaging late fusion ($\sum_{i=1}^{2} w_i p_i \,/\, \sum_{i=1}^{n} w_i$)
7: Apply binary cross-entropy loss and Adam optimizer

8: **end for**
9: Evaluate the performance on test set

3.1 Pre-processing

The multi-modal veracity analysis framework is implemented on three datasets with textual data and image URLs to access the associated images. URLs were loaded using the Beautiful Soup library to access the images. Some of the images were inaccessible, which subsequently reduced the dataset as the corresponding rows were removed. Any attribute having null was handled by removing the corresponding rows. The dataset after all the mentioned processing was saved in the local drives for later access.

As far as preprocessing is concerned, textual data was not preprocessed as the model used was BERT and ALBERT which handles all types of words by its huge vocabulary size. The maximum sequence length for the BERT model for All data [42], Weibo [43] and MediaEval 2016 [44] datasets was fixed at 512, 400 and 20 respectively. The maximum sequence length was decided by taking the limit such that 95-98% of the word length comes under its range. For the dataset having tweets, preprocessing was performed by removing emojis, hyperlinks and tweet mentions from the textual part.

Images for every dataset were resized to (299,299) as this is the default size required by the Inception-ResNet-v2 model for feature extraction. All the images for every dataset were first scaled and converted to NumPy arrays. Then, minor preprocessing was done by using the Keras predefined function to obtain final images for input to our model. Features from textual data are extracted by using BERT and ALBERT language models based on the bidirectional encoding property of transformers and visual features are extracted by using the Inception-ResNet-v2 deep neural pre-trained model. After the feature vectors are obtained from both the modalities, separate standard scaling of both is done by using Sci-Kit Learns module "Standard Scaler".

3.2 BERT

BERT stands for Bidirectional Encoder Representations from Transformers. It was developed by research scientists working at Google AI Language [9]. It can perform various tasks including NLP, Next Sentence Prediction, Question-Answering, Classification, etc. What sets BERT apart from the other similar models is its ability to apply the bidirectional training property of a Transformer, which is a famous attention model, efficiently to language modeling framework. The Transformer is an attention mechanism that is used to learn the

relationship between contexts of different words in text data. Transformers generally comprise two distinct mechanisms —one is an encoder that can read the input text, and another is a decoder that can produce the task prediction. But since BERT is only concerned with generating a language model, the decoder mechanism is not required. BERT is employed for a broad range of language-related tasks. All of them can be successfully performed only by augmenting a small-layer to the standard model. Classification-related tasks, such as sentiment analysis are quite similar in operation to the Next Sentence classification. We only need to add a classification layer on top of the output of the Transformer for the [CLS] token. The input representation for a token in BERT model is created by summating the equivalent token, segment and position embeddings. An apprehension of this structure is highlighted in figure 5.

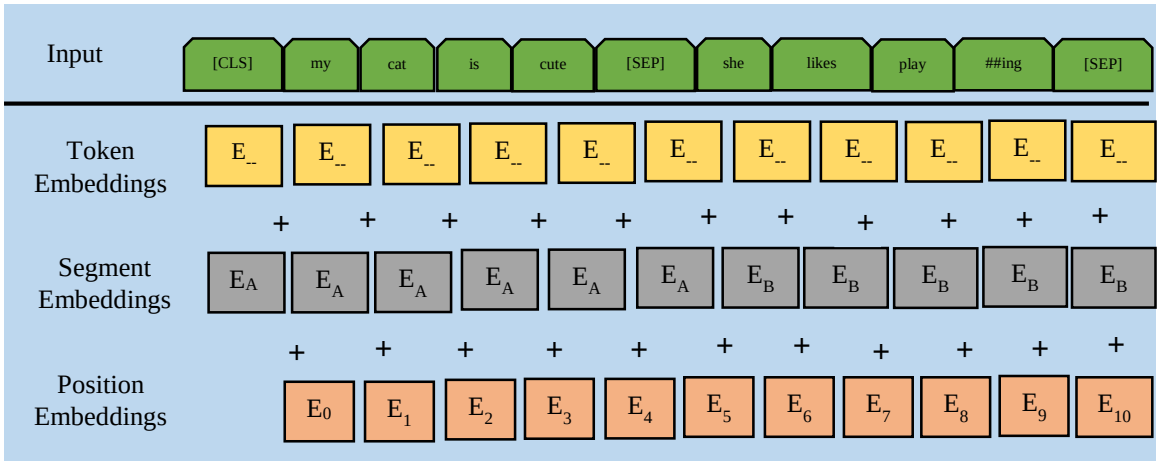


Figure 5: BERT Input Representation

Textual features associated with our dataset were obtained using BERT pre-trained model which was accessed from TensorFlow-Hub. BERT-base version was used for feature extraction as it was found to be suitable for our dataset. BERT-base extracts feature using attention mechanism by several encoding layers. Scaled Dot product attention and Multi Head attention detailed in figure 6 (a) and (b) are the core functions of BERT encoder. An attention function can be defined as mapping a query and a set of key-value pairs to an output, where the query, keys, values and output are all vectors. In scaled dot product attention, the input comprises of keys and queries of dimension d_k and values of dimension d_v.To obtain the weights on the values the dot product of the query with all keys is divided by $\sqrt{d_k}$ and passed on to a SoftMax function as in equation (1). The set of queries, keys and values are arranged together into matrices Q, K and V. Multi-head attention can be described mathematically as in equation (2) and (3) .

$$Attention\ (Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

$$MultiHead(Q,K,V) = Concat(head_1, \ldots\ldots, head_h)W^o \quad (2)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (3)$$

The projections are parameter metrices $W_i^Q \in \mathbb{R}^{d_{model}xd_k}$, $W_i^K \in \mathbb{R}^{d_{model}xd_k}$, $W_i^V \in \mathbb{R}^{d_{model}xd_v}$ and $W^O \in \mathbb{R}^{hd_vxd_{model}}$ where h is the number of parallel attention layers or heads and d_{model} is the dimension of each head. Twelve encoding layers extract implicit features from the text by taking input as input_ids, input_masks, segment_ids. BERT pre-trained model has an associated tokenizer which is built on large-sized vocabularies. In the Weibo dataset, character level tokenization is required as it is a Chinese language dataset where each character represents a word. For every sentence [CLS] and [SEP] tokens are appended, where CLS signifies Classification and SEP signifies special separating token.

- Input_word_ids: Encoded tokens using BERT-tokenizer
- Mask_ids: Separates useful and padded tokens (0 or1)
- Segment_ids: Useful for pairwise training of sentences

These are tokenization inputs to the feed-forward BERT model which has a transformer architecture with 12 attention layers where each layer extracts features by attention mechanism to give output. The model is kept non-trainable to extract the bottleneck features which are obtained by freezing the weights of the entire model. BERT-base gives pooled output and sequence-output and pooled output is used for the further classification task. The vector length associated with everyrow is 768 sized one-dimension vector.

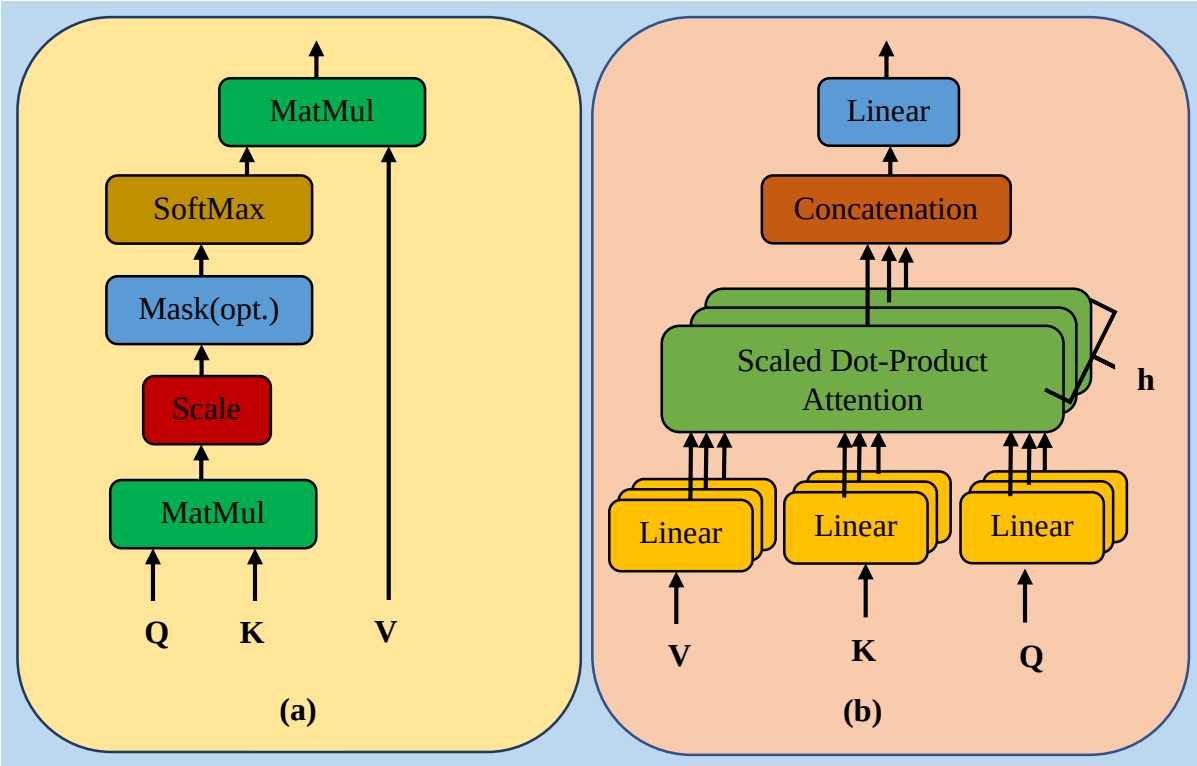


Figure 6: (a) Scaled Dot-Product Attention (b) Multi-Head Attention

In our proposed frameworks we harness the capabilities of BERT's multi-layer bidirectional Transformer encoder-based architecture grounded on the creative implementation of Transformer detailed in [45] by Vaswani et. al. The model is first initialized with the pre-trained parameters and then the last layer is replaced with a customized network to fine-tune the parameters with our labelled training datasets for the downstream task of binary fake news classification.

3.3. ALBERT

ALBERT, which stands for A Lite BERT [10], is a self-supervised learning model that was presented by Google researchers along with some upgrades to the BERT model. They came up with three major innovations that made it an even more refined model than BERT.

The foremost of them is **factorized embedding parameterization**. This played a major role for efficient allocation of the model's capacity. The size of the hidden layers was isolated from the vocab-embeddings size by using one-hot vectorization into first, an embedding space with lower dimensionality, and then to a hidden space altogether. Because of this, it was possible to increase the size of the hidden layer without much tuning of vocab-embeddings' parameter size. Another key feature was the **cross-parameter sharing**. This facilitated the sharing of parameters across all layers, keeping the depth of the model in control. Hence,

ALBERT has 18 times lesser parameters than BERT. Lastly, they introduced the concept of **inter-sentence coherence loss**. The original BERT model was not very reliable when it came to the next sentence prediction tasks. The introduction of SOP loss to model this inter-sentence coherence in ALBERT made it possible to increase the performance in such tasks.

3.4. Inception-ResNet-v2

Combining the best features of both the Inception and ResNet models, Google had proposed the Inception-ResNet-v2 [11].Inception-ResNet-v2 is a type of CNN (convolutional neural network) that is trained on the ImageNet dataset that has more than a million images. The network has a depth of 164 layers and can correctly classify images into about 1000 distinct groups such as pen, tree, box and many fruits. Henceforth, the network has learned sufficient feature representation techniques for a large set of images. The network has a standard size for input images, that is, 299x299.

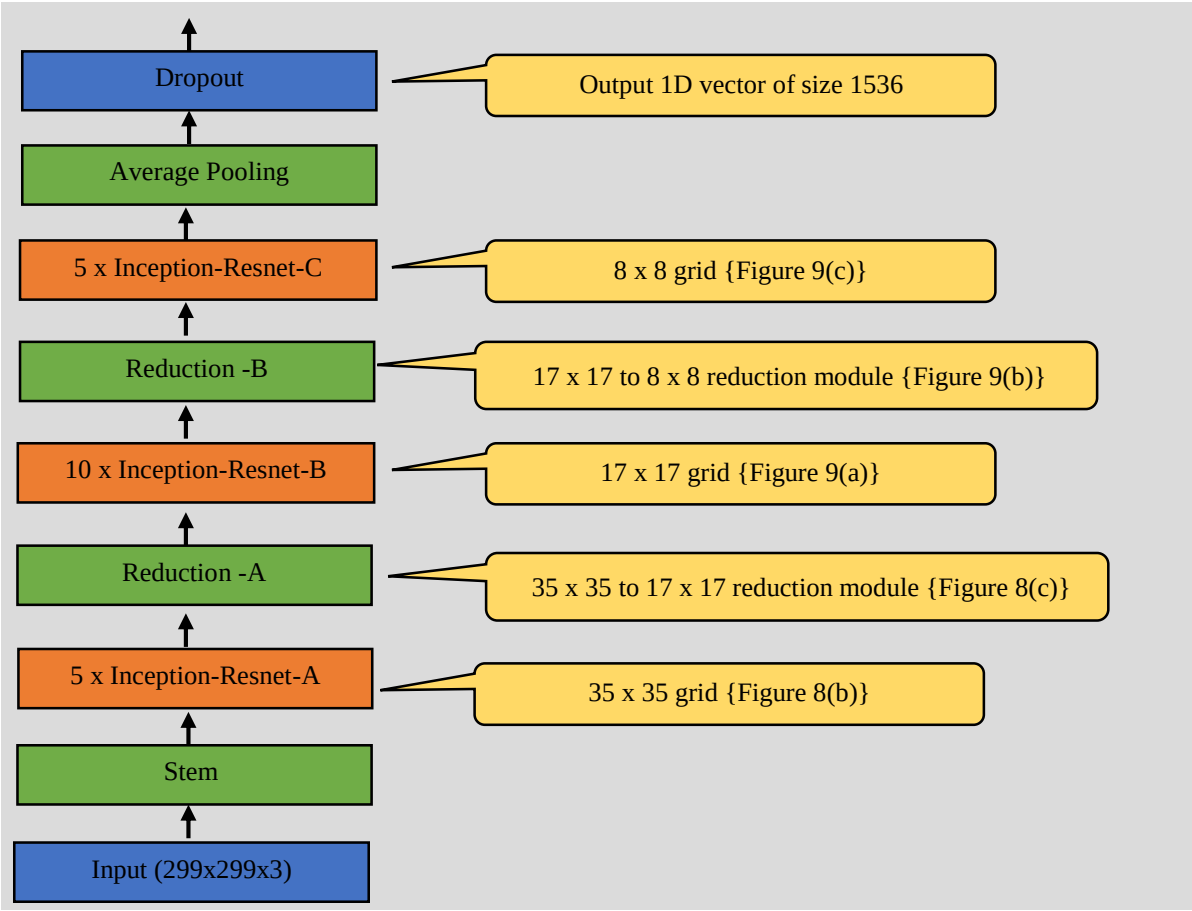


Figure 7: Schema for Inception-ResNet-v2 Network

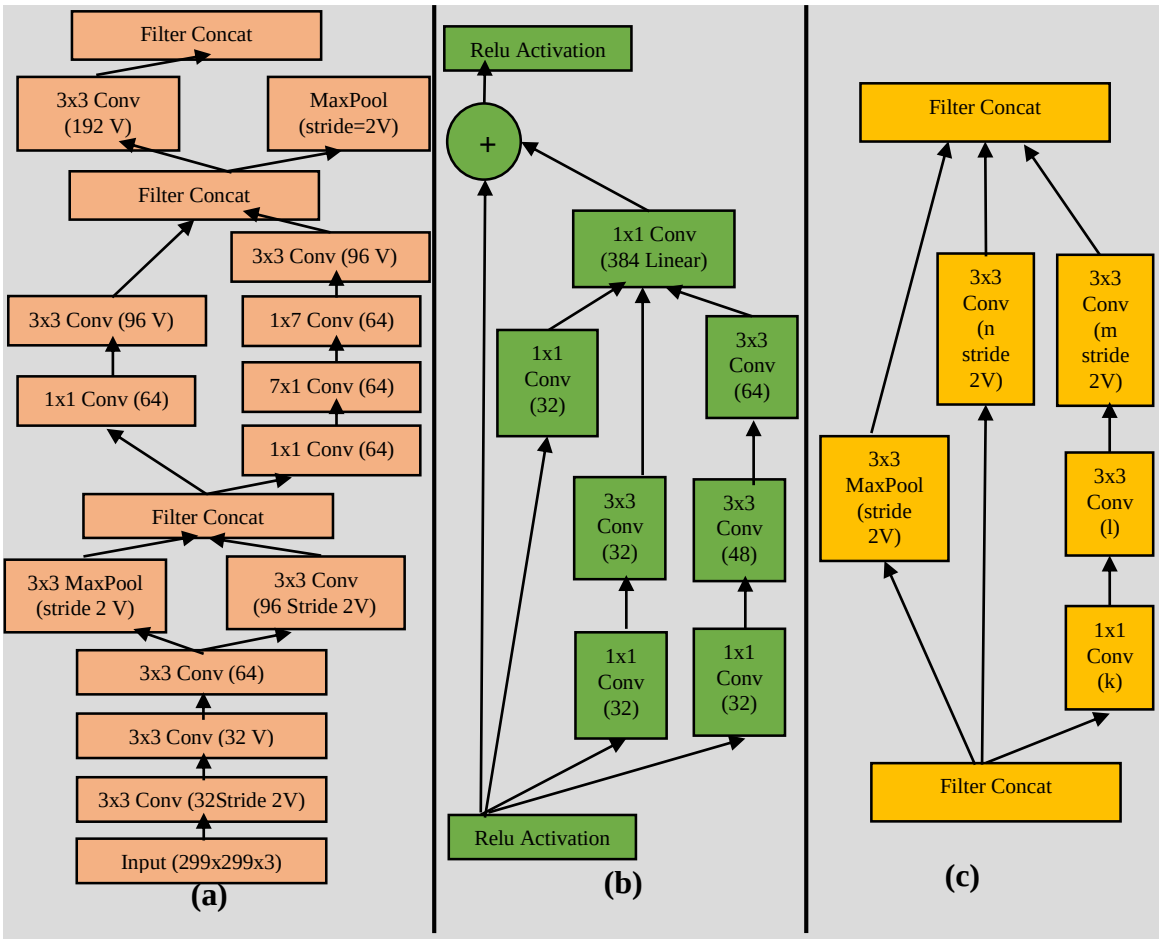


Figure 8: Schema for (a) Stem (b) Inception-Resnet-A (c) Reduction-A blocks of Inception-ResNet-v2 Network

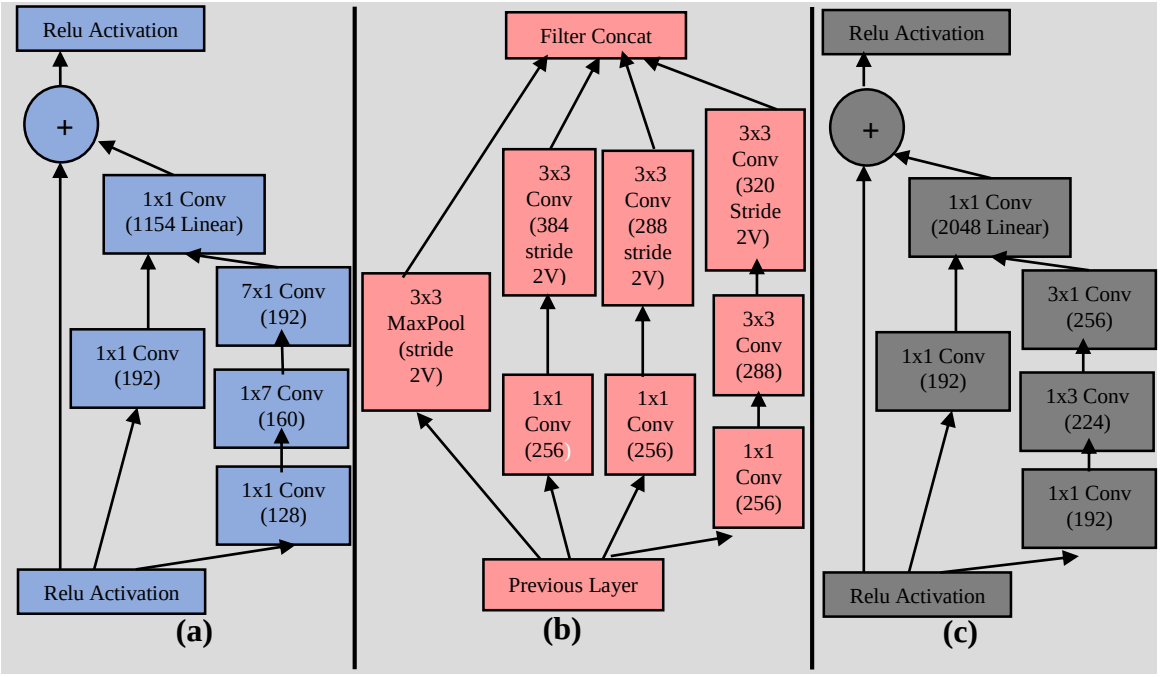


Figure 9: Schema for (a) Inception-Resnet-B (b) Reduction-B (c)Inception-Resnet-C blocks of Inception-ResNet-v2 Network

Figure 7 represents the large-scale schema of Inception-ResNet-v2 network. Figure 8(a), (b), (c) and Figure 9 (a), (b), (c) characterizes the comprehensive structure of its components. All the convolutions not marked with "V" in the figures are same-padded and convolutions marked with "V" are valid padded. The k, l, m and n represent the number of filters of the Reduction-A module for Inception-ResNet-v2 network and their values are 256, 256,384 and 384 respectively.After preprocessing the images, the InceptionResnetv2 model is loaded from the Keras Applications Library. After the model is loaded the final prediction layer is removed which classifies the image to 1000 available classes. The output of the penultimate layer is taken as our feature vector. The vector obtained from the InceptionResNetv2 model is a one-dimensional vector of size 1536.

Inception-Resnet model introduced the concept of residual connections that augment the output produced by the convolution operation of the basic inception model, to the input image. A necessary condition for this residual connection to work, the input image representation and the output produced after the convolution operation should have the same dimensions. Therefore, 1x1 convolutions are used after the original convolutions are performed, to keep the same depth. It is to be noted that after the convolution operation, the depth gets increased. InceptionResNet-v2 model is famous for attaining higher accuracy even at lower epoch cycles.

3.5 Multi-modal Fusion

Multi-modal fusion is the mechanism of combining information from diverse data channels into a single unit. According to the basic architectural levels, there are three basic approaches of fusion: recognition-based, decision based and hybrid multi-level fusion. The recognition-based fusion also popularly known as early fusion merge the feature vectors of each stream by means of integration (concatenation) mechanism. The main requirement of integrating feature vectors from multiple varied modalities is that their dimensions must be identical and this can be achieved by reshape operator. The decision-based fusion, popularly known as late fusion is the process of merging the final probability decisions coming from each branch. Hybrid multi-level fusion comprises of integrating the input modalities by distributing among the decision and recognition levels.

To perform a multi-modal fusion of independent text and image branches, first each of the modality is separately used to train two different models:

- A BERT/ALBERT model followed by custom layers is used to train the text model.

- A pre-trained Inception-ResNet-v2 followed by custom layers is used to train the image model.

Early fusion or recognition-based fusion is achieved by concatenating the feature vectors after converting them into identical dimensions. Late fusion or decision-based fusion is achieved by weighted averaging of the probabilities obtained by the independent predicted arrays from each parallel branch. The overall prediction is performed according to equation (4) by using two normalized weights (w_1, w_2); where p_i is the probability of i^{th} feature and w_i is the corresponding assigned weight. The weights assigned for text classifier is w_1 and for image classifier is w_2 .

$$Final\ prediction = \frac{\sum_{i=1}^{2} w_i p_i}{\sum_{i=1}^{2} w_i} \quad (4)$$

4. Implementation Details

In this sub-section, we extensively detail the experimentation settings, datasets, parameter selection and insights gained from the wide range of experiments performed. The implementation is done on Google Colab which offers up to 13.53 free RAM and 12 GB NVIDIA Tesla K80 GPU. The proposed framework is built and implemented in Python 3 on top of the Keras deep learning framework. All the three datasets are split in the ratio of 7:1:2 for the train, validate and test clusters respectively to make our experimentation better and effective. The performance scores and comparison of the proposed framework are listed in terms of F1-measure, accuracy, recall, precision evaluation metrics. Analysis of the results is being done in numerical as well as in graphical representation using the shapes of accuracy-loss trends with epochs and area under the curve plots individually for each dataset.

4.1 Datasets

To substantiate the efficacy of our proposed framework we experimented with three different datasets comprising unique properties of language and writing style. The first dataset "All Data" contains English news articles, the second one "Weibo" contains Chinese news articles and the third one twitter dataset "MediaEval 2016" comprising of tweets on various topics. All three datasets have text as well as images. The following Table1 highlights the details of the datasets used.

Table1: Dataset Details

Dataset Name	Details	Attributes Used	Total entries	Fake news count	Real news count
All Data [42]	Dataset contains English News Articles with 54 attributes	Title, Text, Image, Label	20015	11941	8074

	including site URL, Image URL, Body, Headline, Label				
Weibo [43]	Dataset contains Chinese News articles with four columns title, text, image URL and label	Title, Text, Image, Label	5250	2767	2438
MediaEval 2016 [44]	Dataset contains tweets related to multiple events with post_id, post_text, user_id, image_ids, username, timestamp and label	Tweet text, Image, Label	15519	6875	8644

4.1.1 All Data

All Data dataset has news articles in English language. It contains title, title-length, body/text, text_length, site_url, sentiment words count, title-length, image_url, and other metadata of a total of 54 attributes for almost 25000 news articles with binary labels of real and fake. The attributes which were used are the title of the news post, the body of the post, associated image and label. Cleaning and preprocessing were performed on the dataset before use. Entire rows that had null values for any of these four attributes were removed. Also, entire rows were removed if the image from image_url could not be fetched from the given source. The Title and Text attributes were concatenated into a single attribute for each row. After cleaning, the dataset had 20,015 rows of data with 11 914 fake news and 8 047 real news count.

4.1.2 Weibo Dataset

Weibo dataset is the standard Chinese dataset having microblogs from the Chinese microblogging site Sina Weibo. It is one of the most popularly used dataset and has news articles in the Chinese Language (Mandarin). News articles are accompanied by an image URL, from which images were fetched and stored before use. The dataset contains title, text, image and binary label (Real/Fake) associated with each news article. Entire rows that had null values for any of these four attributes were removed. Also, entire rows were removed if the image from image_url could not be fetched from the given source. After initial preprocessing, the dataset contains 5250 total news articles out of which 2767 are fake and 2438 are real news articles. The Title and Text attributes were concatenated into a single attribute for each row. For preprocessing, tokenization was performed using a Chinese character level tokenizer. BERT model has a tokenizer which is the most suitable for this task. Tokenization was done on a character level as the Chinese language is to character in the same way as the English language is to words. After this, the dataset was ready to be fed as input to our model.

4.1.3 MediaEval 2016

This dataset is part of the MediaEval task of 2016 and is available online at the Image Verification corpus [44]. The dataset contains tweets related to multiple events. It has been

extensively used since Twitter is amongst the most commonly used social media platforms and is highly susceptible to the creation as well as propagation of fake news. It has several attributes: post_id, post_text, user_id, image_ids, username, timestamp and label. The attributes which were used are Post_text - The body of the post, Image_ids – link to an image related to the post and label - Classified as real or fake. Cleaning and preprocessing were performed on the dataset before use. Entire rows that had null values for any of these three attributes were removed. Also, entire rows were removed if the image from image_ids could not be fetched from the given source. After cleaning the dataset contains a total of 15 519 tweets with images and binary labels out of which 6875 tweets are fake and 8644 tweets are real.

4.2 Model Parameter Description

The network is being trained and tested for three datasets. BERT-base and ALBERT-base pretrained networks are used for text feature extraction and Inception-ResNet-v2 pretrained model is used for image feature vector extraction. All the model parameters are kept identical for these three datasets except the maximum sequence length which is set to 512, 400 and 20 respectively for All Data, Weibo and MediaEval dataset. The network is trained with Adam optimizer having learning rate 0.0001, momentum parameter $\beta_1 = 0.9$ and $\beta_2=0.980$ up to 30 epochs with a batch size of 128. To train Early Fusion architecture text and image feature vectors of length 768 and 1536 are concatenated together followed by five dense layers for smoother dimensionality reduction. The neurons in dense layers are 1024, 512, 128, 64 and 2 correspondingly. Two neuron final dense layer with sigmoid activation supports binary classification. Throughout the custom layers dropout probability is 0.4 with batch normalization. The initial four dense layers are equipped with ReLU activation function and loss is calculated with binary cross entropy loss. The Late fusion multi-modal architecture calculates independent prediction probability p_1 for text and p_2 for image stream. These probabilities are fused with weighted averaging late fusion framework analysed extensively for four different weight combinations $(w_1, w_2) = \{(0.4,0.6), (0.5, 0.5), (0.6,0.4), (0.7, 0.3)\}$. Feature vector extracted from text is followed by four fully connected dense layers of 512, 128, 64 and 2 neurons. Feature vector extracted from image part is followed by five fully connected dense layers of 1024, 512, 128, 64 and 2 neurons. All other hyperparameters are same for both the architectures to have an efficient analysis and fair comparison of the prediction preciseness.

4.3 Result Analysis

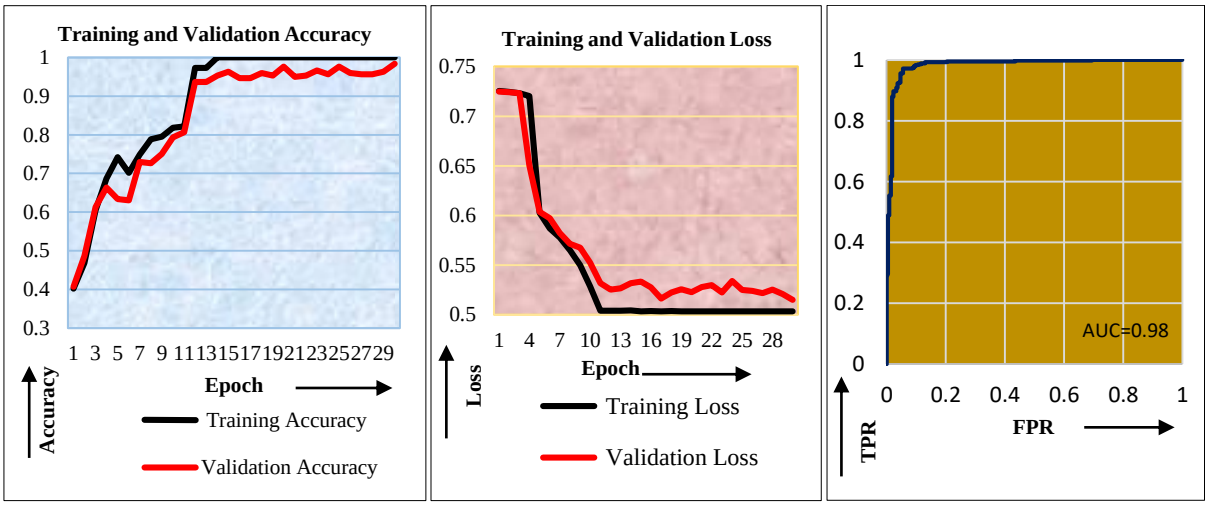
We employed performance measures like Accuracy, Precision, Recall, F1Score,

Accuracy-epoch curve, loss-epoch curve and ROC Curve to analyse and assess the outcomes attained by performing the experimentations on three datasets. The results are presented in Table 2, Table 3 and Table 4 for each dataset sequentially. On each dataset, we applied multi-modal early fusion and late fusion with four different weight combinations for weighted averaging the results using both variants of the framework i.e. BERT and Inception-ResNet-v2 as well as ALBERT and Inception-ResNet-v2. For a clear understanding and insights gained through these results, pictorial representation in terms of graphs is also presented.

The highest accuracy achieved on All Data dataset is 97.19% with ALBERT and Inception-ResNet-v2 framework. Table 2 details in with the outcomes of applying early fusion and late fusion on the dataset. Figure 10 (a), (b) and (c) concentrates on the accuracy versus epoch curve, loss versus epoch curve and ROC curve.

Table 2: Result analysis on All Data Dataset

Fusion	Weightage of text & image streams	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
BERT+Inception Resnet-v2					
Early Fusion	Feature vector concatenation	94.10	95.02	94.88	94.95
Late Fusion	(w1, w2) = (0.4,0.6)	85.75	78.33	82.17	80.20
	(w1, w2) = (0.5,0.5)	90.33	88.06	92.64	90.29
	(w1, w2) = (0.6,0.4)	95.76	92.20	97.45	94.75
	(w1, w2) = (0.7,0.3)	91.33	91.20	89.74	90.46
ALBERT+Inception Resnet-v2					
Early Fusion	Feature vector concatenation	96.41	95.48	94.38	94.93
Late Fusion	(w1, w2) = (0.4,0.6)	87.06	85.77	89.23	87.47
	(w1, w2) = (0.5,0.5)	92.70	90.26	92.11	91.17
	(w1, w2) = (0.6,0.4)	**97.19**	**97.00**	**99.00**	**97.99**
	(w1, w2) = (0.7,0.3)	91.29	92.66	88.24	90.39

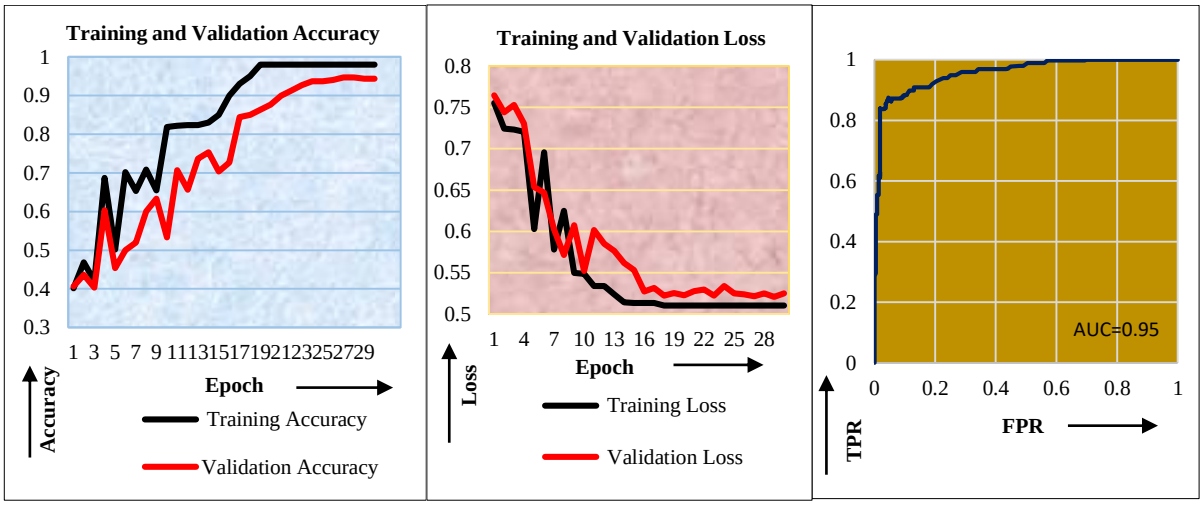


(a) **(b)** **(c)**

Figure 10: (a) Accuracy-Epoch curve (b) Loss-Epoch curve (c) ROC curve for All Data dataset

Table 3: Result analysis on Weibo Dataset

Fusion	Weightage of text & image streams	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
BERT+Inception Resnet-v2					
Early Fusion	Feature vector concatenation	89.61	78.43	76.34	77.38
Late Fusion	(w1, w2) = (0.4,0.6)	80.55	79.42	83.17	81.25
	(w1, w2) = (0.5,0.5)	83.40	86.27	78.90	82.42
	(w1, w2) = (0.6,0.4)	89.92	90.20	93.42	91.78
	(w1, w2) = (0.7,0.3)	82.45	87.32	74.65	80.49
ALBERT+Inception Resnet-v2					
Early Fusion	Feature vector concatenation	91.00	91.27	89.79	90.52
Late Fusion	(w1, w2) = (0.4,0.6)	83.24	86.47	94.33	90.23
	(w1, w2) = (0.5,0.5)	83.67	87.24	92.22	89.66
	(w1, w2) = (0.6,0.4)	**94.28**	**94.07**	**95.82**	**94.94**
	(w1, w2) = (0.7,0.3)	88.04	87.00	90.22	88.58

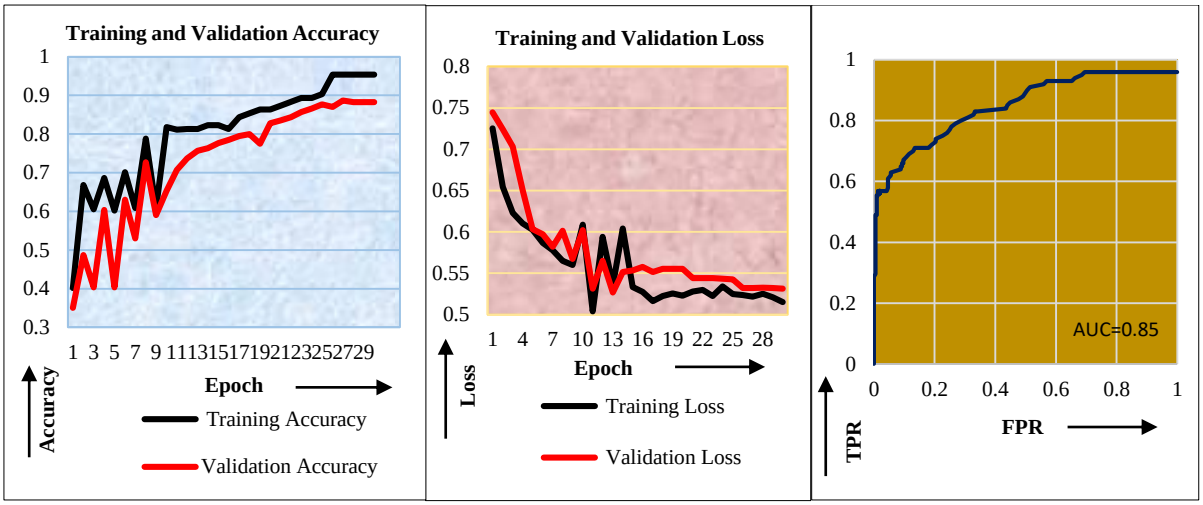


(a) **(b)** **(c)**

Figure 11: (a) Accuracy-Epoch curve (b) Loss-Epoch curve (c) ROC curve for Weibo dataset

Table 4: Result analysis on MediaEval Dataset

Fusion	Weightage of text & image streams	Accuracy	Precision	Recall	F1-score
BERT+Inception Resnet-v2					
Early Fusion	Feature vector concatenation	66.69	59.52	40.86	48.46
Late Fusion	(w1, w2) = (0.4,0.6)	54.22	59.80	54.17	56.84
	(w1, w2) = (0.5,0.5)	59.67	68.52	52.86	59.68
	(w1, w2) = (0.6,0.4)	67.04	74.91	62.50	68.14
	(w1, w2) = (0.7,0.3)	61.51	61.27	65.77	63.44
ALBERT+Inception Resnet-v2					
Early Fusion	Feature vector concatenation	67.79	60.30	76.00	67.24
Late Fusion	(w1, w2) = (0.4,0.6)	57.61	67.25	52.28	58.82
	(w1, w2) = (0.5,0.5)	69.07	70.86	52.47	60.29
	(w1, w2) = (0.6,0.4)	**75.33**	**78.00**	**62.50**	**69.39**
	(w1, w2) = (0.7,0.3)	63.01	74.52	52.09	61.32



(a) **(b)** **(c)**

Figure 12: (a) Accuracy-Epoch curve (b) Loss-Epoch curve (c) ROC curve for MediaEval dataset

It is evident from Table 3 and Table 4 that highest accuracy achieved on Weibo and MediaEval dataset are 94.28% and 75.33% respectively. Figure 11 and Figure 12 elaborates the effectiveness of the training and validation process as well as the robustness of the framework in terms of ROC curve. The above discussion reinforced with Tables 2, 3, and 4 along with Fig. 10,11 and 12 verifies that our proposed multi-modal fusion architecture using fine-tuned self-attention and transfer learning accomplishes promising performance and outstanding precisions for veracity analysis of web information.

4.4 State-of-the-art Comparison

To compare the efficacy of our designed architecture with contemporary methods, we calculated the performance of five different baseline methods in terms of accuracy, precision, recall and F1 score with the dataset split as 7:1:2 for training, validation and testing. State-of-the-art juxtaposition on each of the All Data, Weibo and MediaEval datasets are outlined in Table 5, Table 6 and Table 7, correspondingly. The approaches used as baselines for state-of-the-art comparison are as follows:

- Vishwakarma et al. [26] implemented reverse search method of extracted text from images on internet to calculate a reality parameter by using top search results.
- Wang et al [22] devised multi-modal fake news detector by incorporating text CNN and VGG19 along with event discriminator module. To have a fair comparison, we removed the event discriminator module and compared on binary classifier with our datasets.

- Early fusion of textual features extracted by LSTM and visual features extracted by VGG19 is being proposed and implemented by Jin et al. [46].
- Lago et al. [47] proposed fake news detection multi-modal framework by a combination of classical and advanced techniques such as Mean-filter noise residue inconsistency, Splice buster, JPEG ghosts, TF-IDF, cosine similarity, Logistic Regression, CNN etc.
- Khattar et al. [23] designed variational autoencoder system with three core parts encoders, decoder and a binary fake news detection using Bi-LSTM and VGG19 methods for veracity analysis of online articles.

Table 5: State-of-the-art comparison on All Data Dataset

Reference	Input Modality	Proposed Method /Classifier	Acc. (%)	P (%)	R (%)	F (%)
Vishwakarma et al. [26]	Text +Image	Rule based classifier, Reality Parameter	88.00	87.90	88.80	88.34
Wang et al. [22]	Text +Image	CNN + VGG19	84.27	86.72	83.28	84.96
Jin et al. [46]	Text +Image	LSTM + VGG 19	79.66	87.19	79.61	83.22
Lago et al. [47]	Text +Image	Classical and advanced image forensics, CNN, LR, RF, Jaccard's similarity, TF-IDF, cosine similarity	92.34	93.06	94.06	93.56
Khattar et al. [23]	Text +Image	BiLSTM + VGG19	85.70	87.81	83.03	85.34
Proposed Method	**Text + Image**	**ALBERT + Inception-ResNet-v2**	**97.19**	**97.00**	**99.00**	**97.99**

Table 6: State-of-the-art comparison on weibo Dataset

Reference	Input Modality	Proposed Method /Classifier	Acc. (%)	P (%)	R (%)	F (%)
Vishwakarma et al. [26]	Text +Image	Rule based classifier, Reality Parameter	90.88	91.24	89.16	90.19
Wang et al. [22]	Text +Image	CNN + VGG19	82.73	84.70	81.22	82.92
Jin et al. [46]	Text +Image	LSTM + VGG 19	78.80	86.20	68.60	76.40
Lago et al. [47]	Text +Image	Classical and advanced image forensics, CNN, LR, RF, Jaccard's similarity, TF-IDF, cosine similarity	80.02	79.66	83.24	81.41
Khattar et al. [23]	Text +Image	BiLSTM + VGG19	82.41	85.42	76.93	80.95
Proposed Method	**Text + Image**	**ALBERT + Inception-ResNet-v2**	**94.28**	**94.07**	**95.82**	**94.94**

Table 7: State-of-the-art comparison on MediaEval Dataset

Reference	Input Modality	Proposed Method /Classifier	Acc. (%)	P (%)	R (%)	F (%)
Vishwakarma et al. [26]	Text +Image	Rule based classifier, Reality Parameter	73.45	75.20	71.00	73.04
Wang et al. [22]	Text +Image	CNN + VGG19	71.50	82.20	63.80	71.84
Jin et al. [46]	Text +Image	LSTM + VGG 19	68.20	78.00	61.50	68.77
Lago et al. [47]	Text +Image	Classical and advanced image forensics, CNN, LR, RF, Jaccard's	75.00	76.32	71.67	73.92

		similarity, TF-IDF, cosine similarity				
Khattar et al. [23]	Text +Image	BiLSTM + VGG19	74.52	80.10	71.90	75.77
Proposed Method	**Text + Image**	**ALBERT+ Inception-ResNet-v2**	**75.33**	**78.00**	**62.50**	**69.39**

The above discussion concludes that our proposed model for veracity analysis of web information is reasonably promising. The accuracy over all the three datasets All Data, Weibo and MediaEval is a decent development over the parallel methods. Fine-tuned self-attention and transfer learning are the prominent technologies that have facilitated refining the preciseness of our results. Finally, merging multiple data streams with late fusion helps us achieve 97.19% highest fake news detection accuracy.

5. Conclusion and Future Works

With the exponential emergence of social media as a globally used platform for information circulation and social interactions in the recent decade, we have encountered massive escalation in fake news dispersion. Therefore, the need to establish baselines for information-assessment across all information flow channels has demanded the research community's attention. In this work, we have considered both the textual and visual characteristics of the input news instances. To use these features in veracity analysis, the latest pre-trained deep learning models BERT, ALBERT and Inception-ResNet-v2 have been used. These models are fine-tuned according to our application and dataset requirements to get precise results. Fine-tuning for leveraging existing knowledge in a deep learning model can be potentially beneficial for automatic detection and analysis of fake news. Extensive evaluation of the proposed Multimodal Early and Late fusion framework on three renowned datasets of English news articles, Chinese news articles, and Tweets demonstrated our proposed model's commendable performance in identifying the fake news articles. Its ability to work on large and diverse datasets makes the framework a strong contender for better fake news detection algorithms than the current ones in practice.

There is still a lot of scope for further innovation and future improvements in the proposed architecture. Deep Learning is an ever-evolving field with new and better algorithms being created now and then. Self-attention-based models leveraging the benefits of transformers, dispensing with convolutions and recurrence entirely, will be the future scope of a wide range of language modelling-based application areas. The Transformers can be progressed to figure out the solutions of input and output modalities other than text such as video, audio and images to investigate local, constrained attention mechanisms for handling

massive inputs and outputs proficiently. Making processing less sequential could be another approaching research objective. We hope that this research would work as a baseline for future researchers who can further optimize this solution and make it more relevant with time. Eventually, the proposed work facilitates some reliable mechanism to help in the veracity analysis of web information and more importantly, apprehends the much-needed awareness that everything we read online is not always true.

There is no conflict of interest

References

[1] "Merriam-Webster sine 1828 - The Real Story of 'Fake News'," [Online]. Available: https://www.merriam-webster.com/words-at-play/the-real-story-of-fake-news.

[2] K. Sharma, F. Qian, H. Jiang and N. Ruchansky, "Combating fake news: A survey on identification and mitigation techniques," *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 10, no. 3, pp. 1-42, 2019.

[3] A. Bondielli and F. Marcelloni, "A survey on fake news and rumour detection techniques," *Information Sciences,* vol. 497, pp. 38-55, 2019.

[4] P. Meel and D. K. Vishwakarma, "Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities," *Expert Systems with Applications,* p. 112986, 2019.

[5] C. Wright, "American Psychological Association," 2019. [Online]. Available: https://www.apa.org/research/action/speaking-of-psychology/fake-news.

[6] G. Rannard, "BBC NEWS," December , 2017. [Online]. Available: https://www.bbc.com/news/world-42487425.

[7] A. Bovet and H. A. Makse, "Influence of fake news in Twitter during the 2016 US presidential election," *Nature communications,* vol. 10, no. 7, 2019.

[8] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives,* vol. 31, no. 2, pp. 211-236, 2017.

[9] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805,* 2018.

[10] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942 ,* 2019.

[11] C. Szegedy, S. Ioffe, V. Vanhoucke and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, California, USA, 2017.

[12] B. Dore, "Foreign Policy-Fake News , Real Arrests," April 2020. [Online]. Available: https://foreignpolicy.com/2020/04/17/fake-news-real-arrests/.

[13] J. Z. Pan, S. Pavlova, C. Li, N. Li, Y. Li and J. Liu, "Content based fake news detection using knowledge graphs," in *International semantic web conference*, Monterey, California, USA, 2018.

[14] A. Gautam and K. R. Jerripothula, "SGG: Spinbot, Grammarly and GloVe based Fake News Detection," in *IEEE Sixth International Conference on Multimedia Big Data (BigMM). ,* New Delhi, India, 2020.

[15] Q. Li and W. Zhou, "Connecting the Dots Between Fact Verification and Fake News Detection," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain, 2020.

[16] M. K. Elhadad, K. F. Li and F. Gebali, "Detecting Misleading Information on COVID-19," *IEEE Access*, vol. 8, pp. 165201-165215, 2020.

[17] T. Pomari, G. Ruppert, E. Rezende, A. Rocha and T. Carvalho, "Image splicing detection through illumination inconsistencies and deep learning," in *25th IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, 2018.

[18] S. Elkasrawi, S. S. Bukhari, A. Abdelsamad and A. Dengel, "What you see is what you get? Automatic Image Verification for Online News Content," in *12th IAPR Workshop on Document Analysis Systems (DAS)*, Santorini, Greece, 2016.

[19] F. Marra, D. Gragnaniello, D. Cozzolino and L. Verdoliva, "Detection of GAN-generated Fake Images over Social Networks," in *IEEE Conference on Multimedia Information Processing and Retrieval*, Florida , USA, 2018.

[20] Z. Jin, J. Cao, Y. Zhang, J. Zhou and Q. Tian, "Novel Visual and Statistical Image Features for Microblogs News Verification," *IEEE Transactiona on Multimedia*, vol. 19, no. 3, pp. 598-608, 2017.

[21] Y. Yang, L. Zheng, J. Zhang, Q. Cui, X. Zhang, Z. Li and P. S. Yu, "TI-CNN: Convolutional Neural Networks for Fake News Detection," *arXiv preprint arXiv:1806.00749*, 2018.

[22] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su and J. Gao, "EANN: Event Adversarial Neural Networks for Multi-Modal," in *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, London, United Kingdom, 2018.

[23] D. Khattar, J. S. Goud, M. Gupta and V. Varma, "MVAE: Multi-modal variational autoencoder for fake news detection," in *The World Wide Web Conference*, San Francisco , USA, 2019.

[24] P. Shah and Z. Kobti, "Multi-modal fake news detection using a Cultural Algorithm with situational and normative knowledge," in *IEEE Congress on Evolutionary Computation (CEC)*, Glasgow (UK), 2020.

[25] A. Giachanou, G. Zhang and P. Rosso, "Multimodal Fake News Detection with Textual, Visual and Semantic Information," in *International Conference on Text, Speech, and Dialogue*, Brno, Czech Republic, 2020.

[26] D. K. Vishwakarma, D. Varshney and A. Yadav, "Detection and veracity analysis of fake news via scrapping and authenticating the web search," *Cognitive Systems Research*, vol. 58, pp. 217-229, 2019.

[27] P. Meel and D. K. Vishwakarma, "HAN, image captioning, and forensics ensemble multimodalfake news detection," *Information Sciences*, vol. 567, pp. 23-41, 2021.

[28] J. Ma, W. gao, S. Joty and K.-F. Wong, "An Attention-based Rumor Detection Model with Tree-structured Recursive Neural Networks," *ACM Transactions on Intelligent Systems and Technology*, vol. 11, no. 42, 2020.

[29] J. Ma, W. Gao and K.-F. Wong, "Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* , Vancouver, Canada, 2017.

[30] D. T. Vu and J. J. Jung, "Rumor Detection by Propagation Embedding Based on Graph Convolutional Network," *International Journal of Computational Intelligence Systems*, vol. 14, no. 1, pp. 1053 - 1065, 2021.

[31] Z. Wu, D. Pi, J. Chen, M. Xie and J. Cao, "Rumor detection based on propagation graph neural network withattention mechanism," *Expert Systems with Applications*, vol. 158, pp. 1-16, 2020.

[32] T. H. Do, X. Luo, D. M. Nguyen and N. Deligiannis, "Rumour Detection Via News Propagation Dynamics and User Representation Learning," in *IEEE Data Science Workshop (DSW)*, Minneapolis, MN, USA, 2019.

[33] K. Eismann1, "Difusion and persistence of false rumors in social media networks: implications of searchability on rumor self-correction on Twitter," *Journal of Business Economics*, pp. 1-31, 2020.

[34] S. Khaled, N. El-Tazi and H. M. O. Mokhtar, "Detecting fake accounts on social media," in *IEEE International Conference on Big Data*, Seattle, Washington, USA, 2018.

[35] F. C. Akyon and M. E. Kalfaoglu, "Instagram fake and automated account detection," in *Innovations in Intelligent Systems and Applications Conference (ASYU)*, Izmir, Turkey, 2019.

[36] A. Louni and K. P. Subbalakshmi, “Who Spread That Rumor: Finding the Source of Information in Large Online Social Networks With Probabilistically Varying Internode Relationship Strengths,” *IEEE Transactions on Computational Social Systems,* vol. 5, no. 2, pp. 335-343, 2018.

[37] J. Choi, S. Moon, J. Woo, K. Son, J. Shin and Y. Yi, “Rumor source detection under querying with untruthful answers.,” in *IEEE Conference on Computer Communications*, Atlanta, USA, 2017.

[38] B. Botnevik, E. Sakariassen and V. Setty, “BRENDA: Browser Extension for Fake News Detection,” in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, China, 2020.

[39] D. Varshney and D. K. Vishwakarma, “A unified approach for detection of Clickbait videos on YouTube using cognitive evidences,” *Applied Intelligence,* pp. 1-22, 2021.

[40] P. Meel and D. K. Vishwakarma, “A temporal ensembling based semi-supervised ConvNet for the detection of fake news articles,” *Expert Systems with Applications,* vol. 177, pp. 1-14, 2021.

[41] B. D. Horne, J. Nørregaard and S. Adali, “Robust Fake News Detection Over Time and Attack,” *ACM Transactions on Intelligent Systems and Technology ,* vol. 11, no. 1, pp. 1-23, 2019.

[42] “All Data,” [Online]. Available: https://drive.google.com/file/d/0B3e3qZpPtccsMFo5bk9Ib3VCc2c/view.

[43] Z. Jin, J. Cao, H. Guo, Y. Zhang and J. Luo, “Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs,” in *Proceedings of the 25th ACM international conference on Multimedia,* Mountain View California ,USA, 2017.

[44] “MediaEval 2016,” 2016. [Online]. Available: https://github.com/MKLab-ITI/image-verification-corpus/tree/master/mediaeval2016.

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, “Attention is all you need,” *arXiv preprint arXiv:1706.03762,* 2017.

[46] Z. Jin, J. Cao, H. Guo, Y. Zhang and J. Luo, “Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs,” in *Proceedings of the 25th ACM international conference on Multimedia,* Mountain View California ,USA, 2017.

[47] F. Lago, Q.-T. Phan and G. Boato, “Visual and Textual Analysis for Image Trustworthiness Assessment within Online News,” *Security and Communication Networks,* pp. 1-14, 2019.

Digital Signal Processing 120 (2022) 103253

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

NOMA-based UAV system under finite blocklength regime with analysis in Rician fading channel

Shardul Thapliyal *, Rajoo Pandey, Chhagan Charan

Electronics and Communication Department, National Institute of Technology (NIT), Kurukshetra, India

ARTICLE INFO

Article history:
Available online 27 September 2021

Keywords:
UAV
Finite block-length
NOMA
Power allocation
Rician fading channel
Optimization

ABSTRACT

The finite blocklength (FBL) communication has been studied for quite some time for the fifth generation of cellular networks (5G). Though, the research in FBL is still in early stage due to the fact that 5G requires high reliability of almost 99.9999% with latency of less than 1 ms, simultaneously. The FBL has been worked upon to reduce latency such that Shannon's capacity expression is no longer applicable. The 5G claims to provide connectivity everywhere, hence to increase the coverage to remote places Unmanned Ariel Vehicle (UAV) has gained attention due to its low cost, mobility, and air time. In this paper, a scenario where a UAV relays data from a base station (BS) to remote users (RU), which are otherwise not in direct contact, is analyzed. All the communications among BS, UAV, and RU's are considered to be non-orthogonal multiple access (NOMA) based with Finite blocklength (FBL). An exact closed form expression for end-to-end average block error rate (BLER) is derived for above mentioned scenario and verified by extensive simulations. In addition, an asymptotic BLER for remote users at high transmission Signal to noise ratio is also derived. Further, the BLER is minimized to jointly optimize the power allocation coefficients, position of UAV and Blocklength for the communication system. The effects of different parameters on average BLER, position, and power are studied and the coverage provided by UAV for given probability of error is presented.

© 2021 Elsevier Inc. All rights reserved.

1. Introduction

The fifth generation (5G) cellular networks face great challenges for supporting massive connectivity and seamless connection under limited bandwidth. To handle this problem a variety of technologies have emerged, like Non orthogonal multiple access (NOMA) [1,2], multiple input multiple output (MIMO) [3,4], millimetre wave (mm Wave) [5,6], small cell networks (SCN) [7,8], and UAVs [9,10]. Recently, unmanned aerial vehicle (UAV) has been used in different applications ranging from surveillance, wireless coverage, oil rig monitoring to agriculture, and smart cities [11–15]. The UAV's inherent properties like fast mobility, lower cost, and air time make it suitable for many 5G applications [16]. Hence, UAVs have been quickly adopted for commercial purposes [17,18]. Therefore, to meet the demands of high capacity and high reliability applications, the networks should also ensure reliable communication over the links from base station to UAV and UAV to ground users so that the required quality of service (QoS) can be maintained [19,20].

The non orthogonal multiple access (NOMA) has been an interesting prospect for 5G and beyond applications [21]. The NOMA has been studied in depth with relays and has shown promising results in increasing the performance of overloaded network due to its massive connectivity, high reliability, and efficient spectrum utilization [22–24]. Since, 5G communication promises connectivity everywhere with user density of 1 million devices per square Kilometer, and NOMA has been shown to be efficient in dense user network. Therefore, it has become essential that NOMA should be studied with UAV systems so that wide coverage over a large number of users can be achieved [25]. In [23], UAV based NOMA has been studied where UAV adjusts its position according to the roaming users to maintain a minimum throughput, which is performed by tracking multiple beams with the moving users. A joint beamforming and power allocation under NOMA scheme for satellite communication system has been performed in [24]. Authors in [25], optimize the position and power allocation of UAV with NOMA to maximize the sum rate. In this work, first, the position of UAV is optimized by minimizing the line of sight loss and then the power allocation is optimized for further maximization of sum rate. In [26], the power of UAV is minimized by

* Corresponding author.
E-mail address: shardul_61900075@nitkkr.ac.in (S. Thapliyal).

https://doi.org/10.1016/j.dsp.2021.103253
1051-2004/© 2021 Elsevier Inc. All rights reserved.

jointly optimizing power allocation and position of UAV. In this model, UAV serves distant users that are not directly connected to BS and a power minimization problem is formulated with linear and nonlinear constraints, which is later optimized using penalty and block descent method.

Many of the studies on UAV-NOMA have been performed under infinite block length (IBL) [15–17,24]. Thus, these studies are applicable only when the block-length is considerably large which is not desirable for latency constrained scenarios like real-time applications. As 5G communication demands a high reliability and low latency simultaneously, it has become necessary to study communication system under finite blocklength regime. Therefore, to address this problem, finite block-length (FBL) theoretics has been developed and refined in [27]. Furthermore, this model has been extended to NOMA [28–30], and is thoroughly studied under different scenarios. In addition, FBL model is also extended to channels which are quasi-static fading in nature [31,32] and also with cooperative relay communications [33]. In [34], blocklength optimization under orthogonal multiple access (OMA) with UAV has been studied, where a perturbation based algorithm is used to minimize the probability of decoding error of the system. In [35], a joint blocklength and power allocation has been studied under factory automation scenario, in which authors compare different transmission schemes under FBL. Under FBL regime, another study presented in [36] considers an optimal resource allocation in UAV based communication system. Authors in [37], consider a UAV network and maximize the range for which a BS supports FBL communications for these UAV users.

Among the aforementioned works, most of the studies have considered heterogenous links i.e. UAV to BS employs OMA while UAV to users are served with NOMA. Further, most of the works mainly focus on UAV to ground users scenario, i.e. they assume BS to UAV is a strong link and hence perfect successive interference cancellation (SIC) can be achieved at UAV. In addition, very few works available in literature extend their studies to FBL domain and also do not provide a closed-form expression for average decoding error. These studies use instantaneous error probability with Monte Carlo simulation to perform most of the optimizations.

In NOMA, SIC is a very crucial factor in error analysis, as SIC error propagates from UAV to users. Hence, it becomes essential to study the system from base-station to UAV to ground users as single unit. Also, a closed-form expression is always useful to optimize the complex system. However, to the best of our knowledge, no closed-form expression of this particular scenario has been derived and studied. Hence, in this paper, we have considered a UAV relay system, which carries data from base-station to remote user. UAV as well as users both are served with NOMA. Also, the entire analysis is done under FBL domain which can reduce latency of the system.

Our Contribution

- A closed-form expression for average block error for two remote users over Rician fading channel under finite blocklength regime is derived. In the literature, studies are limited to power minimization and throughput maximization analysis and the models studied have not considered NOMA between BS and UAV. Further, these studies are only available for Rayleigh fading channel. The analysis presented in our work assumes the decode and forward relay protocol [33,35].
- The analysis is further extended by deriving a high SNR asymptotic closed-form expressions for the average block error of remote users.
- An error minimization problem is formulated which jointly optimizes power, position, and blocklength of the communication system. No study is available in literature which considers joint optimization of all these parameters.
- Simulation results are provided to verify the derived expressions and study the effects of different parameters on error performance, and the coverage obtained.

The rest of the paper is organized as follows: Section 2 comprises of system model, transmission scheme, and derivation of average block length error (BLER) for two remote users. In Section 3, a high signal to noise ratio (SNR) approximation for both the users is derived and optimization problem is formulated. Simulation results are analysed and discussed in Section 4. Finally, Section 5 concludes the overall findings of the study.

2. System model and average block error rate

A system model with a base station (BS), UAV and ground users (GU) with coordinates $(x_b,\ y_b)$, $(x_u,\ y_u)$, and $(x_k,\ y_k)$, respectively, where $k \in \{1, 2\}$, is considered as shown in Fig. 1. Here, there are two sets of ground users: near ground users (NGU), and far ground users (FGU). NGU's are served by BS itself while the FGU's or remote users are served by UAV. The analysis of Finite Block-length communication is inherently mathematically complex hence several studies involving FBL have usually considered only two users [38–40]. Thus, in the present work also, the analysis is done for only two FGUs while allotting a fixed power to NGU's. The UAV is assumed to be flying at a fixed height of H meters, while it can move in other two dimensions so that it occupies the optimum position. As assumed in several other studies [41,42], we also consider that FGUs remain quasi static during a communication slot. Thus, the effect of movement of FGUs is considered negligible during a transmission slot so that parameters: power, position, and blocklength optimized by the proposed method can be used in the current transmission slot. Further, the UAV only needs to move when a new optimized position is calculated by the base station.

The present model considers that channels between users and the UAV as well as from BS to UAV remain unchanged for a single time slot. Therefore, many blocks can be transmitted before the next channel estimation.

We use deterministic path loss channel where channel coefficients of different links can be written as:

$$h_k = \sqrt{\beta_k} * \ \psi_k \text{ where } k \in \{b, 1, 2\}$$

where 'b' indicates the coefficient of the link from BS to UAV. Here β_k denotes the large scale fading coefficient while ψ_k denotes the small scale fading factor, for a Rician fading channel. The large scale fading term is only dependent on the distance of UAV from the BS and FGU's, denoted as D_b and D_k, respectively. So from [43] we have,

$$\beta_k = D_k^{-\alpha}\ \beta_0 = \frac{\beta_0}{\left((x_u - x_k)^2 + (y_u - y_k)^2 + H^2\right)^{-\frac{\alpha}{2}}} \text{ where } k \in \{b, 1, 2\} \tag{1}$$

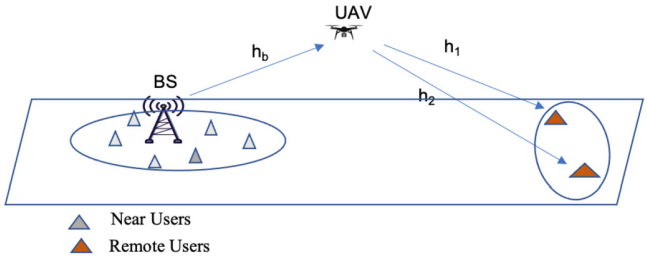


Fig. 1. Data relaying through UAV to remote users. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

where β_0 is the channel power gain at the distance of one meter, and $\alpha \geq 2$ is the path loss exponent. The parameters $\{\psi_k\}$ are assumed to be independent and identically distributed (i.i.d.) random variables with $E[|\psi_k|^2] = 1$. As assumed in other studies [20,21], we consider that the BS only has the partial channel state information (CSI) i.e. distances and Rician factors of the links are known. The UAV is assumed to work in decode and forward (DF) fashion. The communication takes place in two phases: in phase one, BS transfers the data of far users to UAV while in phase two, UAV delivers the data to FGUs. All the terminals possess single antenna. NGU's are served together with FGU's in NOMA scheme but FGU's are allotted more power than NGU's and near users can get their signal by performing SIC. Most of our focus will be on FGU's.

According to Shannon, the capacity is explained as the highest coding rate that can exist between an encoder-decoder pair when the probability of decoding error becomes almost zero when blocklength approaches infinity [28]. With the emergence of applications requiring low latency, it has become necessary to limit and reduce the blocklength in each frame. So, now the probability of error of decoding a message at receiver cannot be ignored. In low latency applications, the specified transmission delay is much smaller than the channel coherence time [37]. According to the work done in [27] for a point to point communication system, the coding rate and probability of error are given by

$$R \approx \log(1+\gamma) - \sqrt{\frac{v}{m}}\frac{Q^{-1}(\epsilon)}{ln2} \tag{2}$$

$$\epsilon = Q\left(F(\gamma, m, N)\right) \tag{3}$$

where, $F(\gamma, m, N) = ln2\sqrt{\frac{m}{v}}\left(\log_2(1+\gamma) - \frac{N}{m}\right)$, ϵ is the decoding error probability, γ denotes the signal to noise ratio (SNR), m is the transmission blocklength, N is the number of bits to be sent to the user, Q^{-1} denotes the inverse of Gaussian Q-function and is given by $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-\frac{t^2}{2}}\,dt$. The parameter v denotes the channel dispersion and is given by $v = 1 - (1+\gamma)^{-2}$. The above expression for error is applicable if $m \geq 100$ [27].

2.1. Transmission scheme

In the scenario considered here, the UAV acts as a relay between BS and FGU's. Hence, transmission takes place in two phases. In the first phase, UAV receives a NOMA signal y_1 from BS, and decodes the messages of FGUs. In the second phase, FGU1 and FGU2 receive the NOMA signal y_2^1 and y_2^2 from the UAV, respectively. The total blocklength is also divided in two parts: m_1 for BS to UAV phase, and m_2 for UAV to FGU phase. To keep the latency in check, the constraint of total block length $m_1 + m_2 \leq M$, where M is total blocklength to be shared in two phases, is used.

We assume that the power allocated to near users is P_n, the power allotted to the link between BS to UAV is P_b, whereas P_{uav} is the power allotted to UAV for allocation to FGU's in the second phase. In the first phase, signal y_1 received by UAV from BS, consists of FGU's messages as well as message of near users and is expressed as:

$$y_1 = \sqrt{P_b\,P_1}\;h_b\,x_1 + \sqrt{P_b\,P_2}\;h_b\,x_2 + (\text{messages of } NGU's \text{ adding to power } P_n) \;+\; \eta_1 \tag{4}$$

where, x_1, x_2 denote the messages of FGU1 and FGU2, respectively. P_1 and P_2, represent power allocation factors of first and second FGU signals, respectively. η_1 denotes the additive white Gaussian noise with distribution $\mathcal{N}(0,1)$ in phase 1, and h_b is channel coefficient between BS and UAV.

In the second phase, UAV decodes the messages from the received signal y_1 and forwards to FGUs. The signals received by FGUs from the UAV are expressed as:

$$y_2^i = \sqrt{P_{uav}\,a_1}\,h_i\,x_1 + \sqrt{P_{uav}a_2}\,h_i\,x_2 + \eta_2^i \tag{5}$$

where, a_1 and a_2 denote the power allocation factors of phase two for FGU1 and FGU2, respectively. The channel coefficients between UAV and FGU i are denoted by h_i, $i \in \{1,2\}$. η_2^i denotes the additive white gaussian noise with distribution $\mathcal{N}(0,1)$ in phase 2.

Now, by using equation (4), the signal to interference and noise ratio (SINR) between BS and UAV for two message signals, denoted as γ_1^{uav} and γ_2^{uav}, are given by

$$\gamma_1^{uav} = \frac{P_b P_1 |h_b|^2}{P_n |h_b|^2 + P_2 P_b |h_b|^2 + \sigma^2} \quad \text{and} \quad \gamma_2^{uav} = \frac{P_b P_2 |h_b|^2}{P_n |h_b|^2 + \sigma^2} \tag{6}$$

where, γ_i^d represents the SINR of 'ith' message signal at location d, with, $i \in \{1,2\}$ and $d \in \{FGUs, UAV\}$. Now SINRs of equation (6) can be normalized by σ^2 to give

$$\gamma_1^{uav} = \frac{\rho_b P_1 |h_b|^2}{\rho_n |h_b|^2 + P_2 \rho_b |h_b|^2 + 1} \quad \text{and} \quad \gamma_2^{uav} = \frac{\rho_b P_2 |h_b|^2}{\rho_n |h_b|^2 + 1} \tag{7}$$

where, $\rho_b = \frac{P_b}{\sigma^2}$ denotes the transmission SNR of the base station and is used to indicate the transmitted power, and $\rho_n = \frac{P_n}{\sigma^2}$. Similarly, using equation (5) SINRs related to links between FGU1 to UAV, and FGU2 to UAV are obtained. Since, FGU1 is a strong user, it decodes its message directly. Hence, SINR for decoding message x_1 at FGU1 is given as

$$\gamma_1^{u1} = \frac{\rho_{uav} a_1 |h_1|^2}{a_2 \rho_{uav} |h_1|^2 + 1} \tag{8a}$$

where, $\rho_{uav} = \frac{P_{uav}}{\sigma^2}$ denotes transmission SNR of UAV. Since, FGU2 has to perform SIC to decode its message, it decodes FGU1's message first. Thus, SNIR for decoding message x_1 at FGU2 is expressed as

$$\gamma_1^{u2} = \frac{\rho_{uav} a_1 |h_2|^2}{a_2 \rho_{uav} |h_2|^2 + 1} \tag{8b}$$

Similarly, after subtracting FGU1's message from signal y_2, the SNIR for decoding message x_2 at FGU2 is given as

$$\gamma_2^{u2} = \frac{\rho_{uav} a_2 |h_2|^2}{1} \tag{8c}$$

Here, γ_i^{uk} implies SINR of ith signal at user uk where $k \epsilon \{1,2\}$. From equations (1) and (6), it can be observed that SINR γ is a function of position of UAV. Hence, in this paper we propose to jointly optimize position, blocklength and power allocation of the NOMA signals.

2.2. Instantaneous errors

The instantaneous errors computed in this section imply the error obtained for a given realisation of channel coefficients. Let ϵr_i^k denote the probability of an instantaneous error, as given by equation (3) for a point to point communication, of message i at location k, where $i \in \{1,2\}$ and $k \in \{uav,\ u1, u2\}$ and $\overline{\epsilon r}_i^k = 1 - \epsilon r_i^k$.

The total probability of instantaneous error of decoding message x_1 at UAV is given as:

$$\epsilon_1^{uav} = \epsilon r_1^{uav} \tag{9}$$

However, the error in decoding message x_2 at UAV can occur due to following events:

(i) The message of second user will be in error, if the first user's message is in error, because SIC will not be perfect. This probability of error can be expressed as:

$$P' = P\,(\text{decoding } x_2 \text{ at UAV in error } \mid x_1 \text{ was decoded in error}) * \epsilon r_1^{uav}$$

where, $P(.|.)$ represents the conditional probability. When successive interference cancellation (SIC) is used, message x_2 can only be decoded correctly if message x_1 was decoded correctly. Therefore, due to large interference present in signal, x_2 will always be in error, if x_1 is incorrectly decoded. Hence, P (decoding x_2 at UAV in error $\mid$ x_1 was decoded in error) is assumed to be 1.

(ii) The UAV does not decode second message correctly, while first one was decoded with no error, i.e.

$$P'' = P\,(\text{decoding } x_2 \text{ at UAV in error} \mid x_1 \text{ was decoded in correctly}) * P\,(\text{decoding } x_1 \text{ correctly})$$

Therefore, total probability of instantaneous error of message x_2 at UAV is given by

$$\epsilon_2^{uav} = P' + P''$$
$$\epsilon_2^{uav} = 1.\epsilon r_1^{uav} + \left(1 - \epsilon r_1^{uav}\right) \epsilon r_2^{uav}$$

where, ϵ_i^d denotes the total probability of instantaneous error on message i at d, where $d \in \{UAV, FGU1, FGU2\}$ and ϵr_i^d indicates the probability instantaneous error on message i at location d. Since in an ultra-reliable communication errors are of the order of 10^{-4} to 10^{-5}, therefore, the term ϵr_2^{uav}. ϵr_1^{uav} can be neglected, and ϵ_2^{uav} can be expressed as:

$$\epsilon_2^{uav} \approx \epsilon r_1^{uav} + \epsilon r_2^{uav} \tag{10}$$

By using equation (3), ϵ_1^{uav} and ϵ_2^{uav} can be expressed as

$$\epsilon_1^{uav} = Q\left(F\left(\gamma_1^{uav}, m_1, N_1\right)\right) = Q\left[ln2\sqrt{\frac{m_1}{v_1^{uav}}}\left(\left(\log_2 1 + \gamma_1^{uav}\right) - \frac{N_1}{m_1}\right)\right] \tag{11}$$

and,

$$\begin{aligned}\epsilon_2^{uav} &\approx Q\left(F\left(\gamma_1^{uav}, m_1, N_1\right)\right) + Q\left(F\left(\gamma_2^{uav}, m_1, N_2\right)\right)\\ \epsilon_2^{uav} &\approx Q\left[ln2\sqrt{\frac{m_1}{v_1^{uav}}}\left(\log_2(1+\gamma_1^{uav}) - \frac{N_1}{m_1}\right)\right]\\ &\quad + Q\left[ln2\sqrt{\frac{m_1}{v_2^{uav}}}\left(\log_2(1+\gamma_2^{uav}) - \frac{N_2}{m_1}\right)\right]\end{aligned} \tag{12}$$

Similarly, error at FGU1 and FGU2 can be found. The message x_1 will be in error at FGU1 due to following events:

(i) The FGU1's message was incorrectly decoded at the UAV.
(ii) The FGU1's message is decoded incorrectly at the user itself.
Therefore,

$$\begin{aligned}\epsilon^{u1} &= \epsilon r_1^{uav} + \epsilon r_1^{u1}\\ \epsilon^{u1} &= Q\left[ln2\sqrt{\frac{m_1}{v_1^{uav}}}\left(\log_2(1+\gamma_1^{uav}) - \frac{N_1}{m_1}\right)\right]\\ &\quad + Q\left[ln2\sqrt{\frac{m_2}{v_1^{u1}}}\left(\log_2(1+\gamma_1^{u1}) - \frac{N_1}{m_2}\right)\right]\end{aligned} \tag{13}$$

In the case of FGU2, user has to perform SIC to decode its own signal. Now, x_2 will be in error at FGU2 due to following events:

(i) There is error at UAV in decoding FGU2's message.
(ii) There is error in decoding the FGU1's message at FGU2.
(iii) There is error in decoding FGU2's message at the user itself.

Again, the probability of event that FGU2's message is in error while the FGU1's message is decoded in error at the FGU2 is taken as 1. Hence

$$\epsilon^{u2} = \epsilon_2^{uav} + \epsilon r_1^{u2} + \epsilon r_2^{u2} \tag{14}$$

$$\begin{aligned}\epsilon^{u2} &= Q\left[ln2\sqrt{\frac{m_1}{v_1^{uav}}}\left(\log_2(1+\gamma_1^{uav}) - \frac{N_1}{m_1}\right)\right]\\ &\quad + Q\left[ln2\sqrt{\frac{m_1}{v_2^{uav}}}\left(\log_2(1+\gamma_2^{uav}) - \frac{N_2}{m_1}\right)\right]\\ &\quad + Q\left[ln2\sqrt{\frac{m_2}{v_1^{u2}}}\left(\log_2(1+\gamma_1^{u2}) - \frac{N_1}{m_2}\right)\right]\\ &\quad + Q\left[ln2\sqrt{\frac{m_2}{v_2^{u2}}}\left(\log_2(1+\gamma_2^{u2}) - \frac{N_2}{m_2}\right)\right]\end{aligned} \tag{15}$$

To obtain the average error, the expected values of total instantaneous error probabilities ϵ^{u1} and ϵ^{u2} are to be computed.

2.3. Average error in decoding a block

Average error at FGUs can be written as

$$\mathbb{E}\left(\epsilon^{u1}\right) = \mathbb{E}\left(\epsilon r_1^{uav}\right) + \mathbb{E}\left(\epsilon r_1^{u1}\right) \tag{16}$$

$$\mathbb{E}(\epsilon^{u2}) = \mathbb{E}(\epsilon_2^{uav}) + \mathbb{E}(\epsilon r_1^{u2}) + \mathbb{E}(\epsilon r_2^{u2}) \tag{17}$$

Here, $\mathbb{E}(x)$ denotes expected value of random variable x. To evaluate $\mathbb{E}(\epsilon r_i^d)$ in the above expressions integral shown below has to be calculated, which is intractable.

$$\mathbb{E}\left(\epsilon r_i^d\right) = \int Q\left(F\left(\gamma_i^d, m_i, N_i\right)\right).f_{\gamma_i^d} d\gamma \tag{18}$$

where $f_{\gamma_i^d}$ is probability density function of SINR. To evaluate the above expression, the approximation derived in [44] is used which considers that $Q\left(F\left(\gamma_i^d, m_i, N_i\right)\right) \approx \mathbb{F}(\gamma_i)$, with $\mathbb{F}(\gamma_i)$, given by

$$\mathbb{F}\left(\gamma_i^d\right) = \begin{cases} 1, & \gamma_i^d < \phi_i^d \\ \frac{1}{2} - v_i^d\left(\gamma_i^d - \tau_i^d\right), & \phi_i^d < \gamma_i^d < \Delta_i^d \\ 0, & \gamma_i^d \geq \Delta_i^d \end{cases} \tag{19}$$

where, $\phi_i^d = \tau_i^d - \frac{1}{2v_i^d}$; $v_i^d = \sqrt{\frac{m_i}{2\pi\left(2^{\frac{2N_i}{m_i}}\right)-1}}$; $\Delta_i^d = \tau_i^d + \frac{1}{2v_i^d}$; $\tau_i^d = 2^{\frac{N_i}{m_i}} - 1$ and $i \in \{1,2\}$ Here, subscript i denotes the ith message and superscript d can be UAV, or far ground users, $u1$ and $u2$. N_i denotes number of bits to be sent to ith user, and m_i shows the blocklength of ith phase. Hence, using equations (18) and (19). $\mathbb{E}\left(\epsilon_i^d\right)$ can be written as

$$\mathbb{E}\left(\epsilon_i^d\right) = v_i^d \int_{\phi_i^d}^{\Delta_i^d} F_{\gamma_i^d}(x)\, dx \tag{20}$$

where, $F_{\gamma_i^d}(x)$ is cumulative density function of the SINR γ_i^d. Using equation (20) one can find the average BLER at FGU1 and FGU2. The CDF $F_{\gamma_1^{uav}}(z)$ is given by

$$\begin{aligned} F_{\gamma_1^{uav}}(z) = P\left(\gamma_1^{uav} \leq z\right) &= P\left[\frac{\rho_b P_1 |h_b|^2}{\rho_n |h_b|^2 + P_2\rho_b |h_b|^2 + 1} \leq z\right] \\ &= P\left[|h_b|^2 \leq \frac{z}{\rho_b P_1 - (\rho_n + P_2\rho_b) z}\right] \end{aligned} \tag{21}$$

Since, the channel considered here is the Rician fading channel, hence h is a Rician random variable containing large scale fading parameter $\sqrt{\beta_k}$ and ψ_k, for simplicity $\sqrt{\beta_k}$ is dropped from analysis, as it can be put back in final result. Further, h^2 has non-central χ^2 distribution, CDF of which is given in [43] as

$$F_x(x) = 1 - Q_1\left(\sqrt{2K}, \sqrt{2(K+1)x}\right) \tag{22}$$

where K is Rician factor of the channel and $Q_1(a,b)$ is Marcum Q-Function of order 1, which is given by $Q_M(a,b) = \int_b^\infty x\left(\frac{x}{a}\right)^{M-1} \times e^{-\frac{x^2+a^2}{2}} I_{M-1}(ax)\ dx$, and $I_M(x)$ is modified Bessel function of order M. Using equations (21) and (22) one can formulate that

$$\begin{aligned} F_{\gamma_1^{uav}}(z) &= 1 - Q_1\left(\sqrt{2K_{ub}}, \sqrt{2(K_{ub}+1)\frac{z}{\rho_b P_1 - (\rho_n + P_2\rho_b) z}}\right) \\ F_{\gamma_1^{uav}}(z) &= 1 - Q_1\left(\sqrt{2K_{ub}}, \sqrt{2(K_{ub}+1)\frac{z}{c_1 - c_2 z}}\right) \end{aligned} \tag{23}$$

where, $c_1 = \rho_b P_1$, $c_2 = \rho_n + P_2\rho_b$, and K_{ub} is Rician factor of link between base station and UAV. Since, it is difficult to evaluate equations (20) with (23). Hence, a series-based expansion of Marcum Q function as given in [45], is used to simplify the integral

$$Q_M(\alpha,\beta) = 1 - e^{-\frac{\alpha^2}{2}} \sum_{j=0}^{\infty} \frac{1}{j!}\left(\frac{\alpha^2}{2}\right)^j \frac{\mathbb{G}\left(M+j, \frac{\beta^2}{2}\right)}{\Gamma(M+j)} \tag{24}$$

where, $\mathbb{G}(a,b)$, represents the incomplete lower Gamma function given as $\mathbb{G}(s,x) = \int_0^x t^{s-1}e^{-t}dt$, and $\Gamma(x)$ is Gamma function. Using equations (23) and (24), we have

$$F_{\gamma_1^{uav}}(z) = e^{-K_{ub}} \sum_{j=0}^{\infty} \frac{1}{j!}(K_{ub})^j \frac{\mathbb{G}\left(1+j, \frac{(K_{ub}+1)z}{c_1 - c_2 z}\right)}{\Gamma(1+j)} \tag{25}$$

So finally using equations (16), (20), and (25), average BLER of FGU1 and FGU2 can be derived as

$$\mathbb{E}\left(\epsilon^{u_1}\right) = v_1^{uav} e^{-K_{ub}} \sum_{j=0}^{\infty} \frac{1}{j!}(K_{ub})^j \frac{I_1}{\Gamma(1+j)} + v_1^{uav} e^{-K_{u1}} \sum_{j=0}^{\infty} \frac{1}{j!}(K_{u1})^j \frac{I_2}{\Gamma(1+j)} \tag{26a}$$

where I_1 and I_2 are given by

$$\begin{aligned} I_1 = B^{uav} c_1 \Bigg[&\left(\frac{1}{c_2 B^{uav}} - \frac{\phi_1^{uav}}{c_1 B^{uav}}\right) \mathbb{G}\left(1+j, \frac{B^{uav}\phi_1^{uav}}{c_1 - c_2\phi_1^{uav}}\right) - \left(\frac{1}{c_2 B^{uav}} - \frac{\Delta_1^{uav}}{c_1 B^{uav}}\right) \\ &\mathbb{G}\left(1+j, \frac{B^{uav}\Delta_1^{uav}}{c_1 - c_2\Delta_1^{uav}}\right) + \frac{e^{\frac{B^{uav}}{c_2}}}{c_2^{j+2}} \sum_{m=0}^{j} \binom{j}{m} \left(-B^{uav}\right)^{j-m} c_2^m \\ &\left(\mathbb{G}\left(m, \frac{B^{uav} c_1}{c_2(c_1 - c_2\Delta_1^{uav})}\right) - \mathbb{G}\left(m, \frac{B^{uav} c_1}{c_2(c_1 - c_2\phi_1^{uav})}\right)\right)\Bigg] \end{aligned} \tag{26b}$$

$$I_2 = B^{u1} g_1 \Bigg[\left(\frac{1}{g_2 B^{u1}} - \frac{\phi_1^{u1}}{g_1 B^{u1}} \right) \mathbb{G}\left(1+j, \frac{B^{u1}\phi_1^{u1}}{g_1 - g_2\phi_1^{u1}}\right) - \left(\frac{1}{g_2 B^{u1}} - \frac{\Delta_1^{u1}}{g_1 B^{u1}} \right) \mathbb{G}\left(1+j, \frac{B^{u1}\Delta_1^{u1}}{g_1 - g_2\Delta_1^{u1}}\right) + \frac{e^{\frac{B^{u1}}{g_2}}}{g_2^{j+2}} \sum_{m=0}^{j} \binom{j}{m} \left(-B^{u1}\right)^{j-m} g_2^m \left(\mathbb{G}\left(m, \frac{B^{u1} g_1}{g_2(g_1 - g_2\Delta_1^{u1})}\right) - \mathbb{G}\left(m, \frac{B^{u1} g_1}{g_2(g_1 - g_2\phi_1^{u1})}\right)\right)\Bigg] \tag{26c}$$

where, $B^{uav} = K_{ub} + 1$, $B^{u1} = K_{u1} + 1$, $c_1 = \rho_b P_1$, $c_2 = \rho_n + P_2\rho_b$, $g_1 = \rho_{uav} a_1$, and $g_2 = \rho_{uav} a_2$. K_{u1} and K_{u2} denote Rician factor of the links from UAV to FGU1 and UAV to FGU2, respectively. The detailed proof of equation (26) is given in appendix A.

Similarly, using the same approach as in the derivation of equation (26), one can proceed to find average BLER of FGU2, as given by

$$\mathbb{E}\left(\epsilon^{u2}\right) = v_1^{uav} e^{-K_{ub}} \sum_{j=0}^{\infty} \frac{1}{j!} (K_{ub})^j \frac{I_1}{\Gamma(1+j)} + v_2^{uav} e^{-K_{ub}} \sum_{j=0}^{\infty} \frac{1}{j!} (K_{ub})^j \frac{I_3}{\Gamma(1+j)} + v_1^{u1} e^{-K_{u2}} \sum_{j=0}^{\infty} \frac{1}{j!} (K_{u2})^j \frac{I_2'}{\Gamma(1+j)} + v_2^{u2} e^{-K_{u2}} \sum_{j=0}^{\infty} \frac{1}{j!} (K_{u2})^j \frac{I_4}{\Gamma(1+j)} \tag{27a}$$

here I_2' is equivalent to I_2 if B^{u1} is replaced by $B^{u2} = 1 + K_{u2}$. Also,

$$I_3 = B^{uav} d_1 \Bigg[\left(\frac{1}{d_2 B^{uav}} - \frac{\phi_1^{uav}}{d_1 B^{uav}} \right) \mathbb{G}\left(1+j, \frac{B^{uav}\phi_1^{uav}}{d_1 - d_2\phi_1^{uav}}\right) - \left(\frac{1}{d_2 B^{uav}} - \frac{\Delta_1^{uav}}{d_1 B^{uav}} \right) \mathbb{G}\left(1+j, \frac{B^{uav}\Delta_1^{uav}}{d_1 - d_2\Delta_1^{uav}}\right) + \frac{e^{\frac{B^{uav}}{d_2}}}{d_2^{j+2}} \sum_{m=0}^{j} \binom{j}{m} \left(-B^{uav}\right)^{j-m} d_2^m \left(\mathbb{G}\left(m, \frac{B^{uav} d_1}{d_2(d_1 - d_2\Delta_1^{uav})}\right) - \mathbb{G}\left(m, \frac{B^{uav} d_1}{d_2(d_1 - d_2\phi_1^{uav})}\right)\right)\Bigg] \tag{27b}$$

and

$$I_4 = \left| \Gamma\left(1+j, \frac{(B^{u2})}{g_2} z\right) - \frac{d_1}{(B^{u2})} \Gamma\left(2+j, \frac{(B^{u2})}{g_2} z\right) \right|_{\phi_2^{u2}}^{\Delta_2^{u2}} \tag{27c}$$

where, $d_1 = P_2\rho_b$, $d_2 = \rho_n$, and $\Gamma(a, b)$ is the upper incomplete Gamma function. The notation $|f(z)|_a^b$ implies $f(b) - f(a)$. It is to be noted that, large scale fading term was dropped from analysis after equation (21). Hence, to incorporate distance parameter back into the equations B^{uav}, B^{u1}, and B^{u2} has to be divided with there respective large scale fading parameter $\sqrt{\beta_k}$. Where, k represent link between BS to UAV and UAV to users i.e. $k \epsilon \{b, 1, 2\}$ while, α is taken as 2 for the presented work.

It may be noted that the average BLER derived in equation (26) and (27) can facilitate the computation of position of UAV, power allocation factors, and blocklength through BLER minimization to ensure the reliable communication with the knowledge of partial CSI only. In absence of these equations, the equation (13) and (15) will have to be used for the optimization. However, this will require the knowledge of instantaneous CSI and computation of instantaneous BLERs repeatedly over a long period of time to obtain the average BLER, which will significantly increase the complexity and latency.

In order to further decrease the complexity of equation (26) and (27), high SNR asymptotic BLER expressions of these equations are presented in the next section

3. High SNR approximation of average BLER for FGUs and problem formulation

As the exact average BLER results derived in Section 2 are complex in nature, which can make optimization more complex, hence a high SNR asymptotic approximation of the average BLER is derived in this section, which is applicable when transmission SNR (ρ_b) is high at BS for a given ρ_{uav} and vice versa. These results are later used in optimization to reduce the complexity.

In order to further decrease the complexity of equations derived in section 2.3, high SNR asymptotic BLER expressions of these equations are presented in the next section.

3.1. High SNR approximation of average BLER for FGUs

As given by equation (25), the CDF function is expressed as

$$F_{\gamma_1^{uav}}(z) = e^{-K_{ub}} \sum_{j=0}^{\infty} \frac{1}{j!} (K_{ub})^j \frac{\mathbb{G}\left(1+j, \frac{(K_{ub}+1)z}{c_1 - c_2 z}\right)}{\Gamma(1+j)}$$

From equation (8.354.1) of [46], the incomplete lower gamma function can be written as sum of series given as:

$$\mathbb{G}(a,t)=\sum_{n}^{\infty}\frac{(-1)^n t^{a+n}}{n!\,(a+n)} \tag{28}$$

If $t \ll 1$, then this series can be approximated by rejecting higher order terms of t to yield

$$\mathbb{G}(a,t)\approx\frac{t^a}{a} \tag{29}$$

Therefore, for FGU1 $z<\Delta_1^{u1}$, we have

$$\begin{aligned}\mathbb{E}\left(\epsilon^{u_1}\right)&=v_1^{uav}\int_{\phi_1^{uav}}^{\Delta_1^{uav}}F_{\gamma_1^{uav}}(z)dz+v_1^{u1}\int_{\phi_1^{u1}}^{\Delta_1^{u1}}F_{\gamma_1^{u1}}(z)dz\\&=e^{-K_{ub}}v_1^{uav}\sum_{j=0}^{\infty}\frac{1}{j!}(K_{ub})^j\frac{\int_{\phi_1^{uav}}^{\Delta_1^{uav}}\mathbb{G}\left(1+j,\frac{(K_{ub}+1)z}{c_1-c_2z}\right)}{\Gamma(1+j)}dz\\&\quad+e^{-K_{u1}}v_1^{u1}\sum_{j=0}^{\infty}\frac{1}{j!}(K_{u1})^j\frac{\int_{\phi_1^{u1}}^{\Delta_1^{u1}}\mathbb{G}\left(1+j,\frac{(K_{u1}+1)z}{g_1-g_2z}\right)}{\Gamma(1+j)}dz\end{aligned}$$

By using equation (28), $\mathbb{E}\left(\epsilon^{u_1}\right)$ can be written as:

$$\begin{aligned}\mathbb{E}\left(\epsilon^{u_1}\right)&=e^{-K_{ub}}v_1^{uav}\sum_{j=0}^{\infty}\frac{1}{j!}(K_{ub})^j\frac{\int_{\phi_1^{uav}}^{\Delta_1^{uav}}\left(\frac{(K_{ub}+1)z}{c_1-c_2z}\right)^{j+1}}{\Gamma(1+j)(j+1)}dz\\&\quad+e^{-K_{u1}}v_1^{u1}\sum_{j=0}^{\infty}\frac{1}{j!}(K_{u1})^j\frac{\int_{\phi_1^{u1}}^{\Delta_1^{u1}}\left(\frac{(K_{u1}+1)z}{g_1-g_2z}\right)^{(}j+1)}{\Gamma(1+j)(j+1)}dz\end{aligned}$$

By equation (3.194.1) of [46] the above expression can be further simplified as:

$$\begin{aligned}\mathbb{E}\left(\epsilon^{u_1}\right)&=e^{-K_{ub}}v_1^{uav}\sum_{j=0}^{\infty}\frac{1}{j!}(K_{ub})^j\frac{I_1^{\infty}}{\Gamma(1+j)(j+1)}\\&\quad+e^{-K_{u1}}v_1^{u1}\sum_{j=0}^{\infty}\frac{1}{j!}(K_{u1})^j\frac{I_2^{\infty}}{\Gamma(1+j)(j+1)}\end{aligned} \tag{30a}$$

where

$$I_1^{\infty}=\frac{\left|B^{uav}z^{j+2}\left(\frac{B^{uav}}{c_1}\right)^j{}_2F_1\left(1+j,2+j;\left(3+j;\frac{c_2}{c_1}z\right)\right)\right|_{\phi_1^{uav}}^{\Delta_1^{uav}}}{c_1(j+2)} \tag{30b}$$

and

$$I_2^{\infty}=\frac{\left|B^{u1}z^{j+2}\left(\frac{B^{u1}}{g_1}\right)^j{}_2F_1\left(1+j,2+j;\left(3+j;\frac{g_2}{g_1}z\right)\right)\right|_{\phi_1^{u1}}^{\Delta_1^{u1}}}{g_1(j+2)} \tag{30c}$$

where ${}_2F_1$ is hypergeometric function and the superscript '∞' indicates the high SNR approximation. Proceeding similarly for FGU2, a high SNR approximation can be derived as:

$$\begin{aligned}\mathbb{E}\left(\epsilon^{u_2}\right)&=e^{-K_{ub}}v_1^{uav}\sum_{j=0}^{\infty}\frac{1}{j!}(K_{ub})^j\frac{I_1^{\infty}}{\Gamma(1+j)(j+1)}\\&\quad+e^{-K_{u2}}v_1^{u2}\sum_{j=0}^{\infty}\frac{1}{j!}(K_{u2})^j\frac{I_2'^{\infty}}{\Gamma(1+j)(j+1)}\\&\quad+e^{-K_{ub}}v_2^{uav}\sum_{j=0}^{\infty}\frac{1}{j!}(K_{ub})^j\frac{I_3^{\infty}}{\Gamma(1+j)(j+1)}\\&\quad+e^{-K_{u2}}v_2^{u2}\sum_{j=0}^{\infty}\frac{1}{j!}(K_{u2})^j\frac{I_4^{\infty}}{\Gamma(1+j)(j+1)}\end{aligned} \tag{31a}$$

where

$$I_3^{\infty} = \frac{\left| B^{uav} z^{j+2} \left(\frac{B^{uav}}{d_1}\right)^j {}_2F_1\left(1+j, 2+j; \left(3+j; \frac{d_2}{d_1} z\right)\right)\right|_{\phi_2^{uav}}^{\Delta_2^{uav}}}{d_1(j+2)} \tag{31b}$$

and

$$I_4^{\infty} = \left| \frac{\left(\frac{K_{u2}+1}{d1}\right)^{1+j} z^{2+j}}{2+j} \right|_{\phi_2^{u2}}^{\Delta_2^{u2}} \tag{31c}$$

Here $I_2'^{\infty}$ and I_2^{∞} have the same relationship as explained before for I_2 and I_2'. It may be observed from equation (30) and (31) that the average BLER depends upon Rician factors through exponential and the summation terms present in the equation. As the Rician factor K is varied the exponential term dominates over the summation term. Therefore, as Rician factor K is increased, the average BLER deceases. Similarly, it is also seen that the average BLER decreases with increase in transmission SNRs, ρ_b and ρ_{uav}. Since, the complexity of high SNR approximation expressions is less than that of equations (26) and (27), therefore equations (30) and (31) are used for optimization in the next section.

3.2. Problem formulation

The requirements of 5G system are stringent in nature, such as, high reliability and low latency. Therefore, an objective function is formed with latency and reliability constraints. Objective function is optimized with respect to three parameters namely; power allocation, blocklength, and position of UAV, leading to an optimal solution. Here, average BLER of the users is minimized with respect to all the above-mentioned parameters. The problem is stated as:

$$\min_{(m_1, m_2, P_1, P_2, a_1, a_2, x_u av, Y_u av)} \quad \mathbb{E}(\epsilon^{u1}), \ \mathbb{E}(\epsilon^{u2}) \tag{32a}$$

$$s.t.\ \epsilon^{u1} < \epsilon^{max} \tag{32b}$$

$$\epsilon^{u2} < \epsilon^{max} \tag{32c}$$

$$m_1 + m_2 \leq M \tag{32d}$$

$$m_1, m_2 \ \epsilon \ \mathbb{Z} \tag{32e}$$

$$P_1 + P_2 \leq 1 \tag{32f}$$

$$a_1 + a_2 \leq 1 \tag{32g}$$

$$m_1, m_2 \geq 100 \tag{32h}$$

In the above problem, the constraints (32b) and (32c) are to keep the reliability criterion in check, while the constraint (32d) satisfies the latency criterion. Value of ϵ^{max} is kept 10^{-4}, while $\mathbb{M}$ is 800. Constraint (32e) ensures the value of blocklength of two phases as integer. Constraint (32f) and (32g) allows the power allocation to be less than or equal to 1, while (32h) keeps minimum values of m_1, m_2 greater than or equal to 100 for the expression of error in equation (6) to be valid.

Above mentioned objective function is highly non-linear in nature which makes it non trivial and time consuming to optimize. Therefore, to reduce the computational complexity of equation (32), the high SNR approximations given by equation (30) and (31) are used in place of exact expressions. The problem presented in equation (32) is non-convex problem. Hence, it cannot be solved using conventional convex optimizations techniques. Therefore, an exhaustive search algorithm is used to find the minimum of the objective function defined in equation (32).

The complexity of $\mathbb{E}(\epsilon^{u1})$, $\mathbb{E}(\epsilon^{u2})$ in equation (32a) depends upon the hypergeometric function ${}_2F_1(a, b, c, z)$ which has the reduced complexity of $O(M(n)log(n)^2)$ where $M(n)$ is the complexity of n-bits integer multiplication [47]. The complexity $M(n)$ depends upon the method employed for integer multiplication. For example, if the simple recursive method is used for integer multiplication $M(n) = O(n^{log(3)})$. The complexity of an exhaustive search algorithm having P number of search parameters each having s steps is $O(s^p)$. Therefore, the optimization of the proposed function of equation (32) involves the complexity of $O(s^5 M(n)log(n)^2)$ as there are five search parameters. However, the higher complexity of the proposed method should not be a limitation as the optimization process takes place at the BS. To further reduce the complexity fast exhaustive search can be employed in case of exhaustive search. Deep neural networks are known to be universal function approximators and have been recently employed for optimization of complex functions. Therefore, the optimization of proposed problem can also be investigated using deep neural network which is left as future work.

4. Simulation results

In this section, we discuss the findings of the proposed work. For simplicity, high SNR approximations stated in section 3.2, are used for the simulations. The decoding error probability requirement is set as given in Table-I. The large scale as well as small scale fading is taken into consideration. Two users are randomly generated in 200 x 200 m^2 area. The UAV flies at the height of 100 meters. The channel power gain β, can be calculated using the model given in equation (7) of [25] as:

Table 1
Parameters used for simulation.

PARAMETERS	VALUES
Height of UAV	100 meters
Upper limit of summation	20
Rician K-factor, UAV-BS, UAV-FGU1,2	5, 3, 1
Number of bits to FGU1,2; N_1, N_2	50, 50
Number of channel user in phase 1,2; m_1, m_2	200, 200
Power allocations (when required fixed) P_1, P_2, a_1, a_2.	0.8, 0.2, 0.8, 0.2
Reference channel power gain β at 1 m.	-50 dB

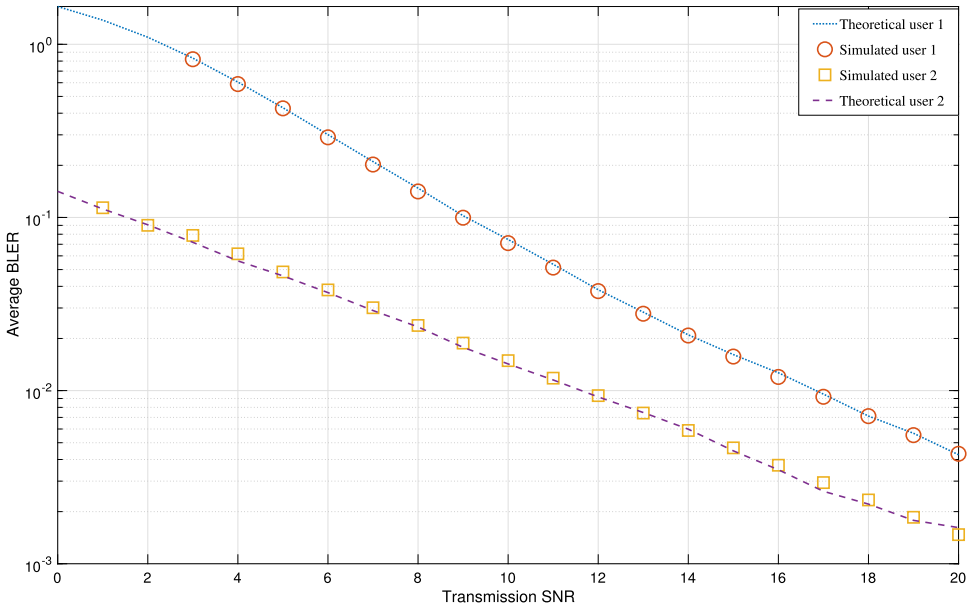


Fig. 2. Comparison of theoretical and simulated results of BLER for FGU1 and FGU2.

$$\beta = -\left(20\log\left(\frac{4\pi f_c D}{c}\right) + \eta_{LOS}\right) \tag{33}$$

where c is velocity of light, η_{LOS} are additional losses, approximately equal to 3 dB. D is the distance from transmitter and f_c is frequency of operation, while negative sign shows the losses. The β_0 in equation (1) can be obtained using equation (33) using $D = 1$ meter and operating frequency as 5 GHz. Thus, channel gain at 1 meter, β_0 is approximated as -50 dB. The upper limit of the summation is set as 20 for trade-off between accuracy and complexity. Table 1 also shows the other parameters required for the simulation. These parameters are set as shown in Table 1 until and unless specified differently. The simulation results are shown in Fig. 2 and also compared with results computed from equations (26) and (27) derived in Section 1. As one can see both theoretical computation and simulation concur each other, which validates the approximations used to derive the results and the accuracy of expression derived for average BLER of FGU1, and FGU2. Fig. 3 shows the result of high SNR expressions derived in Section 2, along with the simulated results. Here also, there is good agreement between the theoretical results and the simulated results for both the users. For FGU1, results match at around 10 dB while for the FGU2 the overlapping starts at around 14 dB. Hence, this again validates the approximation used to obtain the expression under high SNR result for average BLER.

Fig. 4 shows the comparison of BLER between NOMA, OMA and suboptimal NOMA where position of UAV has been fixed while the other two parameters namely blocklength and power are optimized. This clearly shows the importance of joint optimization of all the three factors to minimize the average BLER. In the present simulations of OMA we consider that the blocklength is equally shared between the two users. Another interesting fact one can see from this figure is, the effect of Rician factor on BLER. As pointed out in this Figure, for Rician factor of 5, 3, and 1, FGU2 has higher BLER compared to that of FGU1. This is because when none of the channel is inherently strong, as higher power is allocated to FGU1, it becomes the stronger user, hence has lower BLER. While on the other hand, when Rician factor is 10, 5, 1 i.e. the difference in channel conditions are clearly, large then FGU2, though allocated less power, still acts as stronger user Therefore for stronger statistical channel FGU2 has lower BLER.

Fig. 5, examines the effect of blocklength on the average BLER. As $\mathbb{M}$ is varied from 300 to 5000, BLER for OMA and NOMA both decrease monotonically. As conjectured, larger blocklength leads to high reliability performance in both NOMA and OMA. The difference in error performance of NOMA and OMA is due to the fact that while in NOMA both users can use whole blocklength for transmission, in OMA both users have to share the same in each transmission. Again the Rician factors play similar role as in Fig. 4.

In Fig. 6, variation of position of UAV is shown as SNR increases, i.e. the power from base station increases. It can be seen that as the SNR goes from 15 dB to 25 dB, the UAV moves closer to the users, to achieve better channel between FGUs and UAV, thus reduces the BLER at receiving end. While for smaller power, it requires UAV to fly close to BS to maintain an efficient channel so that BLER at UAV does not increase. Hence with higher SNR, the UAV can provide coverage to larger and more remote areas. All the results discussed in this section

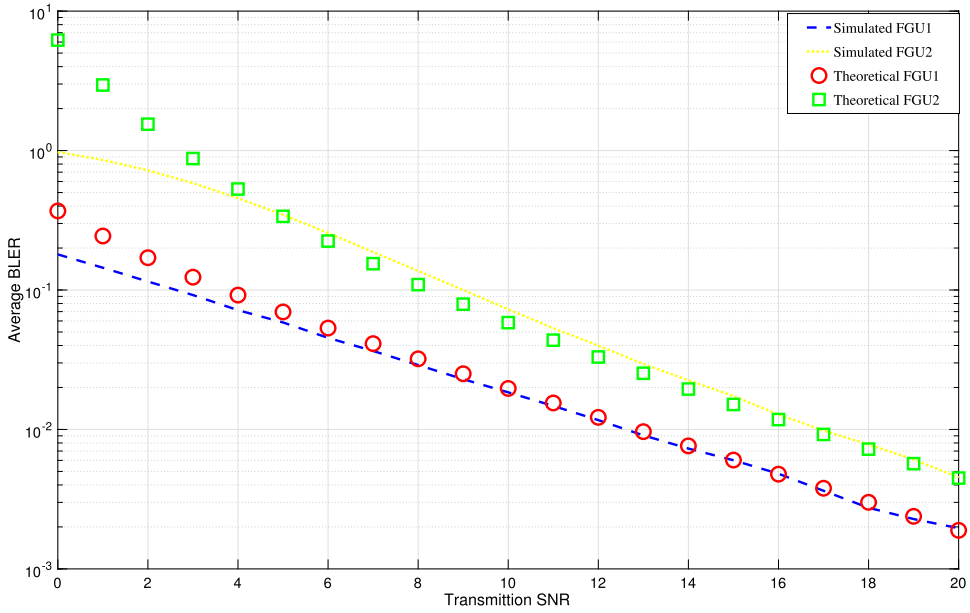


Fig. 3. Comparison of theoretical and simulated BLER for FGU1 and FGU2 under high SNR.

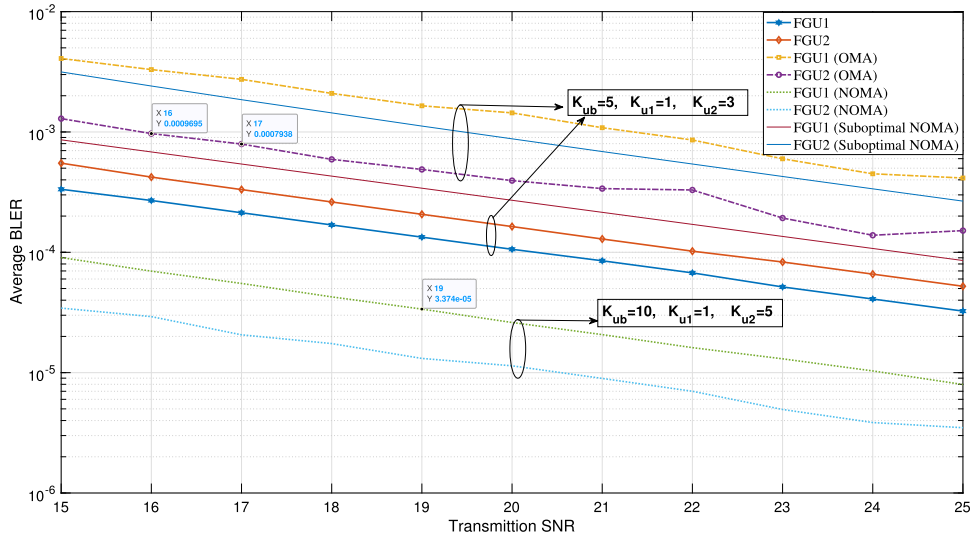


Fig. 4. Comparison of OMA, NOMA and Suboptimal NOMA.

show the effectiveness of the work done in this paper and exploit the potential of NOMA with FBL in UAV based system. Fig. 7 shows the radius of coverage of UAV for a given SNR. Since, FGU2 is a weaker user, hence FGU2 must get coverage to perform efficiently under the reliability constraint as set in equation (32). Therefore, Maximum radius of coverage of UAV under fixed SNR can be evaluated by equation (27), such that, BLER remains less than 10^{-4}. Evaluating the radius of coverage analytically through equation (27) is intractable. Hence, a simulation based study is performed to investigate the coverage provided by the UAV at different SNR. Fig. 7 shows the variation of the radius of coverage with transmission SNR, while the transmission SNR from the base station is fixed at 5, 10, 15 dB. Other parameters are kept same as given in Table 1. It is also observed that a significant change in radius of coverage for different transmission SNR of BS is observed when the transmission SNR of UAV is greater than 10 dB this is because at lower SNR the BLER between UAV and FGUs dominates and significantly higher coverage is not obtained even when the BS transmission SNR is increased.

5. Conclusion

This paper, investigates the performance of NOMA with Finite blocklength for UAV system, where both BS to UAV and UAV to FGU links are served in NOMA fashion. This improves the latency and the reliability performance of the system. This is due to the fact that in OMA either the time slot used for transmission will be doubled or the blocklength will be halved while serving its users. The former will increase the latency while latter will lower the reliability. To investigate the performance of the NOMA-based system presented in this paper, a closed-form expression for average BLER is derived for two far ground users. In addition, high SNR approximation expression for average BLER for two FGU's has also been derived. The accuracy of both, exact as well as high SNR expressions, has been established by

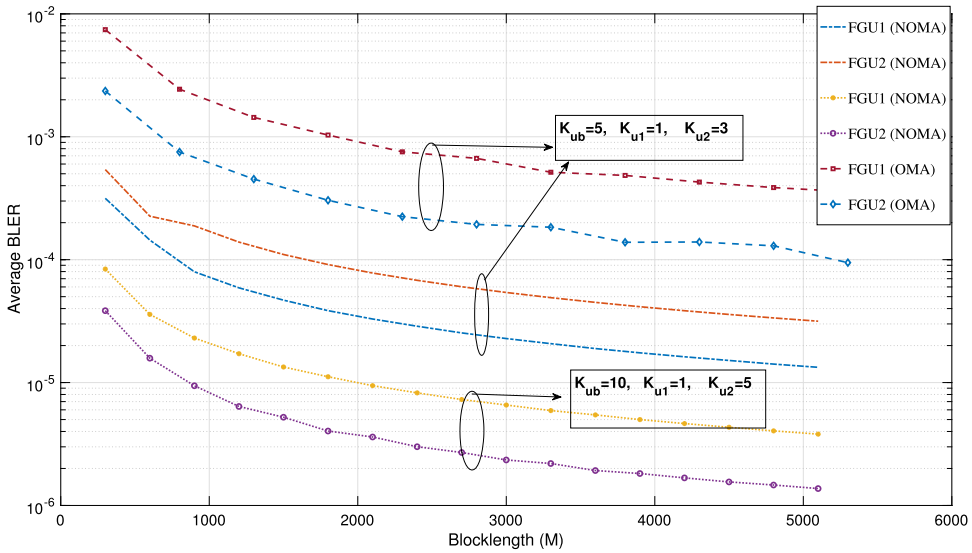


Fig. 5. Effect of blocklength ($\mathbb{M}$) on Average BLER.

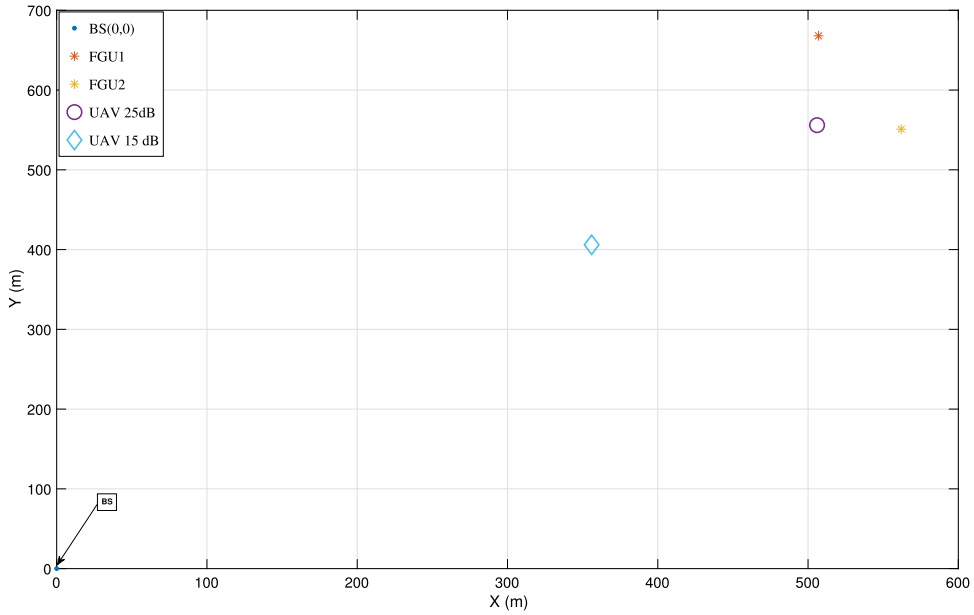


Fig. 6. Location of UAV with different SNR.

Monte Carlo simulations and coverage provided by the proposed method is investigated. The expression obtained for BLER in this paper facilitate the joint optimization of the blocklength, power allocation, and the position of UAV. It also allows us to evaluate the performance of the system for any given values of these parameters. It can be concluded from the results that all three parameters are needed to be optimized to achieve lower average BLER. The comparison between OMA and NOMA has been done, establishing that NOMA can meet low average BLER constraints. Based on asymptotic behaviour of BLER at high SNR condition, simulations prove that NOMA requires lower blocklength than OMA to achieve the same reliabilities, when they are in the order of $10^{-4} - 10^{-5}$.

CRediT authorship contribution statement

All persons who meet authorship criteria are listed as authors, and all authors certify that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript. Furthermore, each author certifies that this material or similar material has not been and will not be submitted to or published in any other publication before its appearance in the Digital Signal Processing journal.

Category 1
Conception and design of study: Shardul Thapliyal.
Analysis: Shardul Thapliyal.

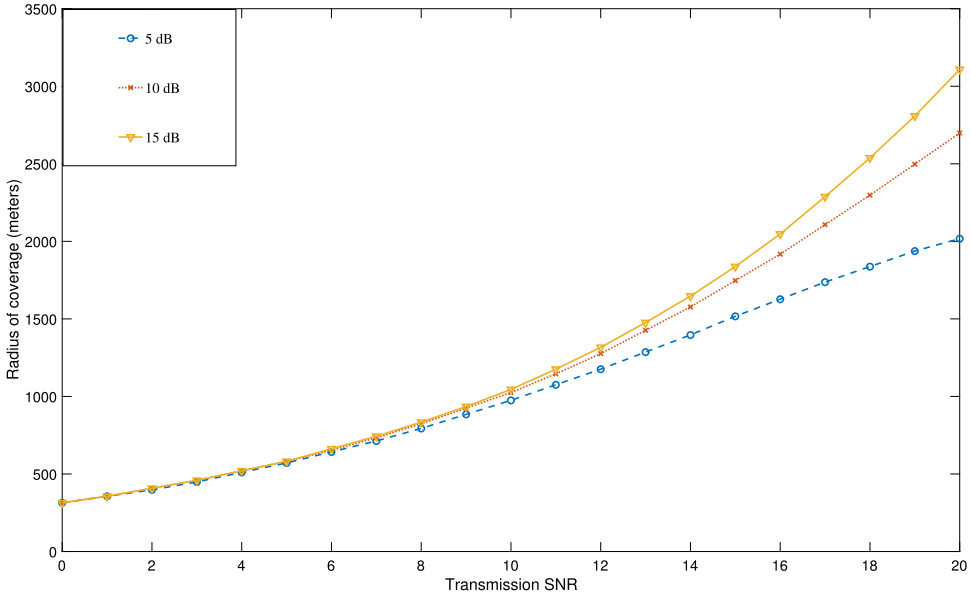


Fig. 7. Variation of radius of coverage with transmission SNR.

Category 2
Drafting the manuscript: Shardul Thapliyal, Rajoo Pandey, Chhagan Charan.
Revising the manuscript critically for important intellectual content: Rajoo Pandey, Chhagan Charan.

Category 3
Approval of the version of the manuscript to be published: Rajoo Pandey, Chhagan Charan, Shardul Thapliyal.

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.

Appendix A. Proof of equation (26)

Equation (26) can be proved as follows:
From equation (25)

$$F_{\gamma_1^{uav}}(z) = e^{-K_{ub}} \sum_{j=0}^{\infty} \frac{1}{j!} (K_{ub})^j \frac{\mathbb{G}\left(1+j, \frac{(K_{ub}+1)z}{c_1-c_2 z}\right)}{\Gamma(1+j)} \tag{A.1}$$

From Eq. (16)

$$\mathbb{E}\left(\epsilon^{u1}\right) = \mathbb{E}\left(\epsilon r_1^{uav}\right) + \mathbb{E}\left(\epsilon r_1^{u1}\right) \tag{A.2}$$

Using equation (A.1) and equation (20), equation (A.2) can be expressed as:

$$\begin{aligned}\mathbb{E}\left(\epsilon^{u1}\right) = v_1^{uav} \int_{\phi_1^{uav}}^{\Delta_1^{uav}} e^{-K_{ub}} \sum_{j=0}^{\infty} \frac{1}{j!} (K_{ub})^j \frac{\mathbb{G}\left(1+j, \frac{(K_{ub}+1)z}{c_1-c_2 z}\right)}{\Gamma(1+j)} dz \\ + v_1^{u1} \int_{\phi_1^{u1}}^{\Delta_1^{u1}} e^{-K_{u1}} \sum_{j=0}^{\infty} \frac{1}{j!} (K_{ub})^j \frac{\mathbb{G}\left(1+j, \frac{(K_{u1}+1)z}{g_1-g_2 z}\right)}{\Gamma(1+j)}\end{aligned} \tag{A.3}$$

The two terms of equation (A.3) can be simplified separately. The first term indicating average BLER of message 1 at UAV $\mathbb{E}\left(\epsilon r_1^{uav}\right)$ is simplified as follows:

$$\mathbb{E}\left(\epsilon r_1^{uav}\right) = v_1^{uav} \int_{\phi_1^{uav}}^{\Delta_1^{uav}} e^{-K_{ub}} \sum_{j=0}^{\infty} \frac{1}{j!} (K_{ub})^j \frac{\mathbb{G}\left(1+j, \frac{(K_{ub}+1)z}{c_1-c_2 z}\right)}{\Gamma(1+j)} dz \tag{A.4}$$

$$= v_1^{uav} e^{-K_{ub}} \sum_{j=0}^{\infty} \frac{1}{j!} (K_{ub})^j \frac{\int_{\phi_1^{uav}}^{\Delta_1^{uav}} \mathbb{G}\left(1+j, \frac{(K_{ub}+1)z}{c_1-c_2 z}\right) dz}{\Gamma(1+j)} \tag{A.5}$$

Let the numerator term involving the integral in equation (A.5) be denoted by I_1. Then

$$I_1 = \int_{\phi_1^{uav}}^{\Delta_1^{uav}} \mathbb{G}\left(1+j, \frac{(K_{ub}+1)z}{c_1-c_2 z}\right) dz \tag{A.6}$$

Now substituting $t = \frac{(K_{ub}+1)z}{c_1-c_2 z}$, new limits to integral can be calculated as:

$$l_1 = \frac{B^{uav}\phi_1^{uav}}{c_1 - c_2\phi_1^{uav}} \quad \text{and} \quad l_2 = \frac{B^{uav}\Delta_1^{uav}}{c_1 - c_2\Delta_1^{uav}}$$

where $B^{uav} = K_{ub} + 1$. Now the equation (A.6) can be expressed as:

$$I_1 = B^{uav} c_1 \int_{l_1}^{l_2} \frac{\mathbb{G}(1+j,t)}{(B^{uav}+c_2 t)^2} dt \tag{A.7}$$

Applying integral by parts on equation (A.7)

$$I_1 = \frac{B^{uav} c_1}{c_2} \left\{ \frac{\mathbb{G}(1+j,l_1)}{B^{uav}+c_2 l_1} - \frac{\mathbb{G}(1+j,l_2)}{B^{uav}+c_2 l_2} + \int_{l_1}^{l_2} \frac{(t)^j e^{-t}}{B^{uav}+c_2 t} dt \right\} \tag{A.8}$$

Substituting, $B_{uav} + c_2 t = x$, and then applying binomial expansion

$$I_1 = \frac{B^{uav} c_1}{c_2} \left\{ \frac{\mathbb{G}(1+j,l_1)}{B^{uav}+c_2 l_1} - \frac{\mathbb{G}(1+j,l_2)}{B^{uav}+c_2 l_2} + \frac{e^{\frac{B^{uav}}{c_2}}}{c_2^{1+j}} \left[\sum_{m=0}^{j} \binom{j}{m} (-B^{uav})^{j-m} c_2^m \left(\int_{B^{uav}+c_2 l_1}^{B^{uav}+c_2 l_2} e^{-\frac{x}{c_2}} x^{m-1} dx \right) \right] \right\} \tag{A.9}$$

From equation (3.381) of [46] integral in equation (A.9) can be solved to simplify the above equation as:

$$I_1 = \frac{B^{uav} c_1}{c_2} \left\{ \frac{\mathbb{G}(1+j,l_1)}{B^{uav}+c_2 l_1} - \frac{\mathbb{G}(1+j,l_2)}{B^{uav}+c_2 l_2} + \frac{e^{\frac{B^{uav}}{c_2}}}{c_2^{1+j}} \left[\sum_{m=0}^{j} \binom{j}{m} (-B^{uav})^{j-m} c_2^m \left(\Gamma\left(m, l_1 + \frac{B^{uav}}{c_2}\right) - \Gamma\left(m, l_2 + \frac{B^{uav}}{c_2}\right) \right) \right] \right\} \tag{A.10}$$

Where $\Gamma(a,b)$ is upper incomplete Gamma function, putting back the values of l_1 and l_2 in equation (A.10), and converting upper gamma function back to lower gamma function

$$\begin{aligned} I_1 = B^{uav} c_1 \Bigg[& \left(\frac{1}{c_2 B^{uav}} - \frac{\phi_1^{uav}}{c_1 B^{uav}} \right) \mathbb{G}\left(1+j, \frac{B^{uav}\phi_1^{uav}}{c_1 - c_2\phi_1^{uav}}\right) - \left(\frac{1}{c_2 B^{uav}} - \frac{\Delta_1^{uav}}{c_1 B^{uav}} \right) \\ & \mathbb{G}\left(1+j, \frac{B^{uav}\Delta_1^{uav}}{c_1 - c_2\Delta_1^{uav}}\right) + \frac{e^{\frac{B^{uav}}{c_2}}}{c_2^{j+2}} \sum_{m=0}^{j} \binom{j}{m} (-B^{uav})^{j-m} c_2^m \\ & \left(\mathbb{G}\left(m, \frac{B^{uav} c_1}{c_2(c_1 - c_2\Delta_1^{uav})}\right) - \mathbb{G}\left(m, \frac{B^{uav} c_1}{c_2(c_1 - c_2\phi_1^{uav})}\right) \right) \Bigg] \end{aligned} \tag{A.11}$$

Substituting I_1 back to equation (A.5) and deriving similarly for second term of equation (A.3) proof of equation (26) is complete.

References

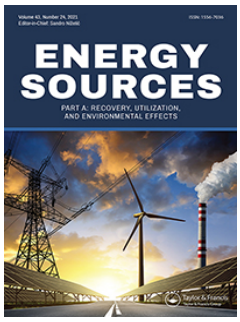
[1] K. Yang, N. Yang, N. Ye, M. Jia, Z. Gao, R. Fan, Non-orthogonal multiple access: achieving sustainable future radio access, IEEE Commun. Mag. 57 (2) (2019) 116–121.
[2] D. Wan, M. Wen, F. Ji, H. Yu, F. Chen, Non-orthogonal multiple access for cooperative communications: challenges, opportunities, and trends, IEEE Wirel. Commun. 25 (2) (2018) 109–117.
[3] E.G. Larsson, O. Edfors, F. Tufvesson, T.L. Marzetta, Massive MIMO for next generation wireless systems, IEEE Commun. Mag. 52 (2) (2014) 186–195.
[4] X. Li, X. Yang, L. Li, J. Jin, N. Zhao, C. Zhang, Performance analysis of distributed MIMO with ZF receivers over semi-correlated $\mathcal{K}$ fading channels, IEEE Access 5 (2017) 9291–9303.
[5] Z. Gao, L. Dai, D. Mi, Z. Wang, M.A. Imran, M.Z. Shakir, MmWave massive-MIMO-based wireless backhaul for the 5G ultra-dense network, IEEE Wirel. Commun. 22 (5) (2015) 13–21.
[6] S. Alemaishat, O.A. Saraereh, I. Khan, S.H. Affes, X. Li, J.W. Lee, An efficient precoding scheme for millimetre-wave massive MIMO systems, Electronics 8 (Aug 2019).
[7] C.H. Liu, L.C. Wang, Optimal cell load and throughput in green small cell networks with generalized cell association, IEEE J. Sel. Areas Commun. 34 (5) (2016) 1058–1072.
[8] X. Li, J. Li, L. Li, L. Du, J. Jin, D. Zhang, Performance analysis of cooperative small cell systems under correlated Rician/Gamma fading channel, IET Signal Process. (2018) 12.
[9] M. Jia, X. Gu, Q. Guo, W. Xiang, N. Zhang, Broadband hybrid satellite-terrestrial communication systems based on cognitive radio toward 5G, IEEE Wirel. Commun. 23 (6) (2016) 96–106.
[10] X. Tang, K. An, K. Guo, Y. Huang, S. Wang, Outage analysis of non-orthogonal multiple access-based integrated satellite-terrestrial relay networks with hardware impairments, IEEE Access 7 (2019) 141258–141267.
[11] F. Mohammed, A. Idries, N. Mohamed, J. Al-Jaroodi, I. Jawhar, UAVs for smart cities: opportunities and challenges, in: 2014 International Conference on Unmanned Aircraft Systems (ICUAS), 2014, pp. 267–273.

[12] T. Pobkrut, T. Eamsa-ard, T. Kerdcharoen, Sensor drone for aerial odor mapping for agriculture and security services, in: 2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2016, pp. 1–5.
[13] P.B. Quater, F. Grimaccia, S. Leva, M. Mussetta, M. Aghaei, Light unmanned aerial vehicles (UAVs) for cooperative inspection of PV plants, IEEE J. Photovolt. 4 (4) (2014) 1107–1113.
[14] C.J. Katila, A. Di Gianni, C. Buratti, R. Verdone, Routing protocols for video surveillance drones in IEEE 802.11s wireless mesh networks, in: 2017 European Conference on Networks and Communications (EuCNC), 2017, pp. 1–5.
[15] X. Zhang, L. Duan, Fast deployment of UAV networks for optimal wireless coverage, IEEE Trans. Mob. Comput. 18 (3) (2019) 588–601.
[16] Z. Ullah, F. Al-Turjman, L. Mostarda, Cognition in UAV-aided 5G and beyond communications: a survey, IEEE Trans. Cogn. Commun. Network. 6 (3) (2020) 872–891.
[17] Fadi Al-Turjman, M.F. Enver Ever, Drones in IoT Enabled Spaces, CRC Press, Taylor and Francis Group, 2019.
[18] M. Mozaffari, W. Saad, M. Bennis, Y.H. Nam, M. Debbah, A tutorial on UAVs for wireless networks: applications, challenges, and open problems, IEEE Commun. Surv. Tutor. 21 (3) (2019) 2334–2360.
[19] Luis Duque, et al., Synthesis of Unmanned Aerial Vehicle Applications for Infrastructures, ASCE Library, 2018.
[20] Z. Khosravi, M. Gerasimenko, S. Andreev, Y. Koucheryavy, Performance evaluation of UAV-assisted mmWave operation in mobility-enabled urban deployments, in: 2018 41st International Conference on Telecommunications and Signal Processing (TSP), 2018, pp. 1–5.
[21] S.M.R. Islam, M. Zeng, O.A. Dobre, K.S. Kwak, Nonorthogonal Multiple Access (NOMA): How It Meets 5G and Beyond, Wiley, 2019.
[22] H. Lin, F. Gao, S. Jin, G.Y. Li, A new view of multi-user hybrid massive MIMO: non-orthogonal angle division multiple access, IEEE J. Sel. Areas Commun. 35 (10) (2017) 2268–2280.
[23] L. Liu, S. Zhang, R. Zhang, Multi-beam UAV communication in cellular uplink: cooperative interference cancellation and sum-rate maximization, IEEE Trans. Wirel. Commun. 18 (10) (2019) 4679–4691.
[24] Z. Lin, M. Lin, J.B. Wang, T. de Cola, J. Wang, Joint beamforming and power allocation for satellite-terrestrial integrated networks with non-orthogonal multiple access, IEEE J. Sel. Top. Signal Process. 13 (3) (2019) 657–670.
[25] X. Liu, J. Wang, N. Zhao, Y. Chen, S. Zhang, Z. Ding, et al., Placement and power allocation for NOMA-UAV networks, IEEE Wirel. Commun. Lett. 8 (3) (2019) 965–968.
[26] X. Jiang, Z. Wu, Z. Yin, Z. Yang, Zhao N. Power, Consumption minimization of UAV relay in NOMA networks, IEEE Wirel. Commun. Lett. 9 (5) (2020) 666–670.
[27] Y. Polyanskiy, H.V. Poor, S. Verdu, Channel coding rate in the finite blocklength regime, IEEE Trans. Inf. Theory 56 (5) (2010) 2307–2359.
[28] X. Sun, S. Yan, N. Yang, Z. Ding, C. Shen, Z. Zhong, Short-packet downlink transmission with non-orthogonal multiple access, IEEE Trans. Wirel. Commun. 17 (7) (2018) 4550–4564.
[29] Y. Yu, H. Chen, Y. Li, Z. Ding, B. Vucetic, On the performance of non-orthogonal multiple access in short-packet communications, IEEE Commun. Lett. 22 (3) (2018) 590–593.
[30] M. Amjad, L. Musavian, Performance analysis of NOMA for ultra-reliable and low-latency communications, in: 2018 IEEE Globecom Workshops (GC Wkshps), 2018, pp. 1–5.
[31] S. Xu, T.H. Chang, S.C. Lin, C. Shen, G. Zhu, Energy-efficient packet scheduling with finite blocklength codes: convexity analysis and efficient algorithms, IEEE Trans. Wirel. Commun. 15 (8) (2016) 5527–5540.
[32] C. She, C. Yang, T.Q.S. Quek, Cross-layer optimization for ultra-reliable and low-latency radio access networks, IEEE Trans. Wirel. Commun. 17 (1) (2018) 127–141.
[33] Y. Hu, J. Gross, A. Schmeink, On the capacity of relaying with finite blocklength, IEEE Trans. Veh. Technol. 65 (3) (2016) 1790–1794.
[34] C. Pan, H. Ren, Y. Deng, M. Elkashlan, A. Nallanathan, Joint blocklength and location optimization for URLLC-enabled UAV relay systems, IEEE Commun. Lett. 23 (3) (2019) 498–501.
[35] H. Ren, C. Pan, Y. Deng, M. Elkashlan, A. Nallanathan, Joint power and blocklength optimization for URLLC in a factory automation scenario, IEEE Trans. Wirel. Commun. 19 (3) (2020) 1786–1801.
[36] Y. Hu, G. Sun, G. Zhang, M.C. Gursoy, A. Schmeink, Optimal resource allocation in ground wireless networks supporting unmanned aerial vehicle transmissions, IEEE Trans. Veh. Technol. 69 (8) (2020) 8972–8984.
[37] C. She, C. Liu, T.Q.S. Quek, C. Yang, Y. Li, Ultra-reliable and low-latency communications in unmanned aerial vehicle communication systems, IEEE Trans. Commun. 67 (5) (2019) 3768–3781.
[38] M. Amjad, L. Musavian, S. Aïssa, Noma versus oma in finite blocklength regime: link-layer rate performance, IEEE Trans. Veh. Technol. 69 (12) (2020) 16253–16257.
[39] Y. Xu, C. Shen, T.H. Chang, S.C. Lin, Y. Zhao, G. Zhu, Transmission energy minimization for heterogeneous low-latency NOMA downlink, IEEE Trans. Wirel. Commun. 19 (2) (2019) 1054–1069.
[40] S. Schiessl, M. Skoglund, J. Gross, NOMA in the uplink: delay analysis with imperfect CSI and finite-length coding, IEEE Trans. Wirel. Commun. 19 (6) (2020) 3879–3893.
[41] I. Budhiraja, S. Tyagi, S. Tanwar, N. Kumar, J.J. Rodrigues, Tactile Internet for smart communities in 5G: an insight for NOMA-based solutions, IEEE Trans. Ind. Inform. 15 (5) (2019) 3104–3112.
[42] S.K. Zaidi, S.F. Hasan, X. Gui, Outage analysis of ground-aerial NOMA with distinct instantaneous channel gain ranking, IEEE Trans. Veh. Technol. 68 (11) (2019) 10775–10790.
[43] C. Zhan, Y. Zeng, R. Zhang, Energy-efficient data collection in UAV enabled wireless sensor network, IEEE Wirel. Commun. Lett. 7 (3) (2018) 328–331.
[44] B. Makki, T. Svensson, M. Zorzi, Finite block-length analysis of the incremental redundancy HARQ, IEEE Wirel. Commun. Lett. 3 (5) (2014) 529–532.
[45] A. Annamalai, C. Tellambura, J. Matyjas, A new twist on the generalized marcum Q-function QM(a, b) with fractional-order M and its applications, in: 2009 6th IEEE Consumer Communications and Networking Conference, 2009, pp. 1–5.
[46] D. Zwillinger, Table of Integrals, Series, and Products, Academic, New York, NY, USA, 2014.
[47] H. Cheng, B. Gergel, E. Kim, E. Zima, Space-efficient evaluation of hypergeometric series, SIGSAM Bull. 39 (2) (2005 Jun) 41–52.

Shardul Thapliyal is currently pursuing Ph.D from National Institute of Technology, Kurukshetra, India. He received his B.Tech (Bachelors of Technology) in electronics and communication from Graphic Era University, Dehradun in 2012 and M.Tech (Masters of technology) degree in digital signal processing from G.B. Pant engineering college (GBPEC), Uttarakhand, India in 2015. His research interest includes Wireless communication, Cognitive Radios, Non Orthogonal Multiple Access, Short packet communication.

Rajoo Pandey is presently working as a Professor in the Department of Electronics and Communication Engineering, National Institute of Technology Kurukshetra, India. He received his Bachelor of Engineering (B.E.) degree in 1989 from Government Engineering College (GEC) Jabalpur, and Master of Technology (M.Tech) from NIT-Kurukshetra (formerly REC Kurukshetra). He received his Ph.D. degree from Indian Institute of Technology, Roorkee in 2002. His research interests include signal processing, communication systems, Digital Signal Processing, Image processing and cognitive radios etc.

Chhagan Charan is working as an Assistant Professor in the Department of Electronics and Communication Engineering, National Institute of Technology, Kurukshetra, India. He received his B. E. (Bachelor of Engineering) degree in Electronics and Communication Engineering from Govt. Engineering College, Bikaner (University of Rajasthan, Jaipur) in 2009 and M. Tech. Degree (Master of Technology) in Microwave and Optical Communication Engineering from Delhi Technological University (DTU), Delhi in 2011. He completed his PhD degree in Electronics and Communication Engineering at National Institute of Technology, Kurukshetra in 2018. His Research interests include Cognitive Radio, Digital communication, Microwave engg., Wireless sensor networks, and Next generation of communication systems.



Energy Sources, Part A: Recovery, Utilization, and Environmental Effects

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ueso20

Numerical simulation of cascaded cyclone separator for nanosize aerosol

Sonawane Chandrakant R., Atul Kulkarni, Rohan V. Sawant, Anand Kumar Pandey, Hitesh Panchal, Kishor Kumar Sadasivuni & Anil Kumar

To cite this article: Sonawane Chandrakant R., Atul Kulkarni, Rohan V. Sawant, Anand Kumar Pandey, Hitesh Panchal, Kishor Kumar Sadasivuni & Anil Kumar (2021): Numerical simulation of cascaded cyclone separator for nanosize aerosol, Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, DOI: 10.1080/15567036.2021.1974127

To link to this article: https://doi.org/10.1080/15567036.2021.1974127

Published online: 15 Oct 2021.

Submit your article to this journal

Article views: 107

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=ueso20

ENERGY SOURCES, PART A: RECOVERY, UTILIZATION, AND ENVIRONMENTAL EFFECTS
https://doi.org/10.1080/15567036.2021.1974127

Taylor & Francis
Taylor & Francis Group

Check for updates

Numerical simulation of cascaded cyclone separator for nanosize aerosol

Sonawane Chandrakant R.[a], Atul Kulkarni[b], Rohan V. Sawant[a], Anand Kumar Pandey[a], Hitesh Panchal[c], Kishor Kumar Sadasivuni[d], and Anil Kumar[e]

[a]Symbiosis Institute of Technology, Symbiosis International Deemed University, Pune, India; [b]Symbiosis Centre for Nanoscience and Nanotechnology, Symbiosis International Deemed University, Pune, India; [c]Mechanical Engineering Department , Government Engineering College, Patan, India; [d]Centre for Advanced Materials, Qatar University, Doha, Qatar; [e]Mechanical Engineering Department, Delhi Technological University, New Delhi, India

ABSTRACT

Among the conventionally used sampling methods, the cyclone-based air samplers are considered to be the most effective and simplest method. However, the collection efficiency of these commercially available air samplers is found to be limited. It reaches a critically low value for nanosized aerosol particle separation. This paper presents a computational fluid dynamics-based numerical simulation of a cascaded cyclone separator planned to be used for the separation of nanosize airborne viruses. The turbulent incompressible flow model with a volume of a fluid formulation is employed during the simulation. The discrete particle method is used to model and track the nanosized aerosol particle trajectory. The fluid properties of the airborne viruses are assigned to the aerosol particles during the simulation to obtain realistic flow results. The Stairmand high-efficiency cyclone geometry is redesigned by the numerical experiment method to obtain high collection efficiency for nanosize aerosol particles, and a new set of geometrical ratios has been proposed. Then the effect of inlet flow rate on the collection efficiency is presented. The fluid flow dynamics are observed, and the detailed distribution of particles collected in various portions of the cascaded cyclone is presented. The present cascaded cyclone could separate nanosized aerosol particles with an average efficiency of more than 93%.

ARTICLE HISTORY
Received 11 May 2021
Revised 19 August 2021
Accepted 20 August 2021

KEYWORDS
Cyclone separator; aerosol separation; multiphase flow; solid-gas separation; incompressible flow; nanosized; turbulent modeling

Introduction

Several applications in biotechnology demand the collection and separation of various biological particles like bacteria, enzymes, viruses, etc. present in the air, These colloidal suspended particles in the air are known as aerosols. Several techniques like screens, filters, centrifuge, etc. are available for collecting and separating these aerosols. However, the cyclone separation method is one of the most effective separating devices popularly used for separating all types of particles from the main flowing fluid (Zhao 2004). Cyclone has a simple structure with no moving parts, resulting in longer and convenient operations with low operational and maintenance costs. A typical cyclone consists of conical bottom connected to the cylindrical top. The fluid enters tangentially through an inlet connected to the top cylindrical portion. Cyclone has two outlets, the first one located at the bottom side of the conical portion known as the underflow outlet. The second outlet is located at the top of the

CONTACT Sonawane Chandrakant R. crsonawane@gmail.com Symbiosis Institute of Technology, Symbiosis International Deemed University, Pune, India; Panchal Hitesh Engineerhitesh2000@gmail.com Government Engineering College, Patan, India

© 2021 Taylor & Francis Group, LLC

cylindrical portion, usually known as a vortex finder, as it helps develop a vortex inside the cyclone. Depending upon application requirements, the necessary output can be obtained through one of the cyclone outlets.

The separation process inside the cyclone separator is mainly driven by the centrifugal forces developed by the vortex generated inside the cyclone, which throws the heavy particles toward the cyclone's wall. The cyclone separators are used for applications starting from vacuum cleaning purposes to a wide range of industrial applications, such as the mining industry, power industry, etc. Cyclone separators are extensively studied, both experimentally and numerically, by several authors. Several models/techniques are available to predict and evaluate cyclone performance. In general, these models can be categorized as (i) model-based on theoretical and semi-empirical formulations (Barth 1956); (ii) statistical models (Casal and Martinez-Benet 1983), and (iii) models based on numerical or computational fluid dynamics simulations.

Early work on cyclones includes an experimental investigation (Avci and Karagoz 2001) where various factors like particle size, flow rate, cyclone size, and gas properties affecting the cyclone collection efficiency have been explored and presented. The performance of cyclone separators based on significant nondimensional numbers like Reynolds number and the Stokes number has also been presented. The experimental investigation as well as the statistical approach (Casal and Martinez-Benet 1983) related to the effect of pressure drop, solid inlet loading on cyclone performance had also been reported.

In this paper, we have presented detailed literature reviews on numerical as well as computational fluid dynamics modeling. Cullivan et al. (2004) were among the first researcher who simulated the hydro cyclones using a transient three-dimensional computational fluid dynamics solver. The air core dynamics developed inside the cyclone separator were visualized and explained using CFD simulations. Gimbun et al. (2005) investigated various parameters like inlet flow velocity, inlet temperature, and pressure affecting the cyclone collection efficiency using commercial CFD Fluent 6.0 software. They also compared CFD results with various empirical models and concluded that CFD results are in good arrangement with Li and Wang model along with experimental results. Bhaskar et al. (2007) simulated turbulent flow inside the cyclone using different turbulence models and compared them with the experimental results. Turbulence modeling employed in the simulation helped them to predict realistic flow scenarios occurring inside the cyclone by reducing the error between experimental and simulated results. They also observed that an increase in the inlet pressure develops higher axial as well as tangential velocities along with an increase in the mass flow rate which ultimately helps the cyclone separator to improve its performance by reducing cyclone cut size diameter. Khairy and Chris (2010) used the Muschelknautz method of modeling (MM) and CFD modeling to investigate cyclone performance. Their CFD analysis confirmed the strong dependence of cyclone performance on the inlet and an outlet (vortex finder) cyclone dimension. Selami (2014) numerically investigated pressure variations inside the cyclone due to changes in vortex finder diameter as well as the due to the height of various cyclone parts. They observed that pressure drop across cyclone increases with an increase in vortex finder diameter. In contrast, pressure drop decreases with an increase in the height of cyclone.

Bogodage and Leung (2015) numerically simulated the performance of cyclones having different outlet geometries. They analyzed the particle trajectories and their variation with various outlet sections. Their CFD simulation with added down-comers geometry inside the cyclone shows a significant improvement in the cyclone collection efficiency. Houben et al. (2016) used a central vortex stabilizing rod inside the cyclone and conducted the CFD simulation using Ansys Fluent 13 software. They intended to develop a cyclone separator having high efficiency by suppressing the central vortex core precession. The movement of particles inside the vortex core was tracked to understand the flow characteristics inside the cyclone. Chengming et al. (2016) simulated the flow dynamics inside the cyclone separator for separating different particles sizes and observed that heavy particles were accumulated near-wall whereas the lightest particles were located at the center of the cyclone. They also evaluated various forces acting on the particles and observed that the particles

inside the cyclone are mainly affected by the drag force and centrifugal forces. Due to the drag force, the particles are pulled inside the core and they are balanced by the outward acting centrifugal force pushing particles outside the vortex in radial directions. They recommended that the other forces like pressure force, Saffman lift force, etc., were much smaller in magnitude than drag forces hence can be neglected.

Balestrin et al. (2017) investigated cyclone separators using the Eulerian–Lagrangian-based CFD modeling approach to capture external particles with having a size below 5 μm. Based on CFD simulation they recommended that cyclone collections efficiency can be improved by promoting the secondary vortex flow inside the cyclone. Giwoon et al. (2017) simulated a wet cyclone separator for virus sampling purposes using the Lagrangian method with discrete phase modeling (DPM). They also carried out CFD simulations for tracking the various virus sizes and observed that cyclone separator efficiency drastically reduces with the reduction in virus size. Paweł et al. (2017) simulated the gas and gas-particle flowing inside the cyclone separator using commercial CFD code. Based on the CFD simulation, they observed that solid particles form a distinct strand near-wall region and they can reach the hopper within two turns.

Recently, Nakhaei et al. (2020) had reviewed CFD studies available for cyclone separators operating at different solid loadings and elevated temperature conditions. They have also summarized the existing simulation studies of cyclones at ambient temperatures. Nassaj and Toghraie (2019) simulated a cyclone separator with guided channels and evaluated pressure drop and cyclone separation efficiency changes. The presence of a guided channel helps to create four inlets into the cyclone and develop a uniform flow field inside the cyclone, causing an increase in tangential velocity and an increase in centrifugal forces inside the cyclone separator. Qiang et al. (2019) investigated the cyclone separator having vortex finder having cuts located at 45°, 135°, and 225° opposite to the tangential inlet separator. They observed that the cyclone separator having a notched vertex finder performs better than the original cyclone separator hence improving the overall cyclone performance. Behrang, Shirvani, and Hashemabadi (2019) investigated a cyclone separator consisting of multiple helical channels separating fine particles from gas streams. Their cyclone separator consists of two major parts, the contracting gas section, and multiple helical channels build inside the cyclone. The kinetic energy and centrifugal forces help to remove the external particles even at low velocities (Kabeel et al. 2020). Park and Go (2020) used machine learning and CFD simulation to investigate the characteristics of cyclone separators. Their model reveals that drag forces and centrifugal forces contribute equally with 50% separation efficiency, hence indicating that the critical diameter was one of the essential dependent parameters for the cyclone separator. Khoi and Yong (2020) investigated a gas–solid flow inside a novel cyclone separator having four inlets. They used a Reynolds stress model-based CFD simulation along with a DPM to track the motion of the particles inside the cyclone separator. Liu et al. (2021) numerically studied the vertical separator used at shale gas gathering and transportation stations. They investigated the separator erosion due to the different gas velocities, the mass flow rate of sand particles, and sand particle size.

The above literature review reveals that CFD simulations are popularly used to predict and evaluate single cyclone separator characteristics. However, it has also been seen that, as suspended particle size reduces, the cyclone separation efficiency reduces drastically. Chuen-Jinn et al. (2007) were one of the few studies available, which had investigated an axial flow cyclone separator for removing nanosized particles. Hence, in this paper, we have numerically investigated the cascaded cyclone separator for separating aerosol having a diameter in the range of micrometer to the nanometer. This nanosized aerosol particle represents the airborne virus planned to be sampled out for bioscience applications (Sonawane, Panchal, Sadasivuni, 2021; Mandal and Sonawane, 2014) The fluid properties of the representative virus are used for aerosol during the CFD simulation to obtain realistic flow conditions. The turbulent incompressible flow inside the cascaded cyclone is simulated using a volume of fluid formulation (Sonawane, More, Pandey, 2021; Mandal et al., 2011) The commercial Ansys-Fluent 19.2 is used to carry out CFD simulation with the k–ε turbulent model (Panchal and Patel 2018; Panchal and Shah, 2013) which is found suitable for modeling the aerosol. The discrete particle model is used

to track the particle trajectory inside the cascaded cyclone separator. The objective here is to develop a high-efficiency cyclone separator that is capable to separate nanosized particles hence the Stairmand cyclone is redesigned and cascaded in a series configuration to obtain the high particle separation efficiency. Firstly, the CFD results were validated for the single Stairmand cyclone case and compared with analytical model results for validations. Then, CFD simulations are carried out on the newly proposed cascaded cyclone for various inlet flow rates to predict and optimized the cyclone separation efficiency. The CFD simulations are also carried out to understand the fluid flow characteristics and visualize the particle trajectories along with evaluation of fluid and particle dynamics happening inside the cascaded cyclone.

Cascaded cyclone separator design

Figure 1 shows the different views of the three-dimensional CAD model of a cascaded cyclone separator designed for separating nanosized aerosol particles. The two cyclones are connected in series. The outlet of the first cyclone is connected with the inlet of the second cyclone. Each cyclone has only one inlet, having a circular cross-section, arranged such that the flow enters inside both the cyclone in tangential to the cyclone axis. The vortex finder is placed in axial directions such that it helps to create a vortex inside the cyclones. Detailed dimensions of the modified cascaded cyclone, which is redesigned from the Stairmand cyclone geometry, and necessary cyclone geometrical ratios are also mentioned in Table 1.

Mathematical model used for CFD analysis

The flow inside the cascaded cyclone separator is found to be complex and turbulent due to the vortex-like motion of the fluid. The two vortexes are found to be formed; the first outer vortex is created near the cyclone wall forcing heavy particles toward the bottom of the cyclone. The second one, rotating in opposite directions, is created at the center of the cyclone forcing flow toward the vortex finder, i.e., cyclone exit. Hence simulating the flow inside the cyclone needs a good turbulent model with incompressible flow modeling. The suspended aerosol particles also need to be tracked properly to evaluate the flow particle dynamics. The aerosol particle properties, which represent the airborne viruses, need to be accurately prescribed to obtain realistic flow simulations.

The aerosol particle separation process development consists of two major components: the cascaded cyclone and fluidics system (not discussed in this paper). Figure 1 show that the cascaded cyclone is a component made using two air cyclone samplers and connected with the elbow. This cyclone consists of a circular cross-section inlet connected to the main body in tangential directions. Note that the vortex finder of the first cyclone is connected through the elbow to the next cyclone such that fluid enters tangentially to the second cyclone. The fluidics system, not discussed here, supplies the pressurized air containing aerosol mixed with the water vapors at required flow rates. The aerosol concentration, as well as the size of the aerosol particles in the sampled air, can be controlled during the CFD simulation. The water vapor from the fluidics system helps to separate nanosized aerosol as water vapor forms the nuclei at these suspended external particles and gets stick to the cyclone wall.

In this paper, a cascaded cyclone separator is redesigned using the design of numerical experiment (DOnE) approach to separate aerosol representing the airborne viruses. The commercially available Ansys Fluent 19.2 (ANSYS Inc.) is used for cascaded cyclone computational fluid dynamics simulations (Sonawane et al., 2019; Sonawane et al., 2021). The incompressible flow simulations are solved using the SIMPLE-C solver (Patankar 1980), which solved the Navier-Stokes equations by pressure-velocity linked iterative procedure. Upwind second-order accurate discretization in space is used to evaluate convective fluxes. As mentioned, the flow inside the cyclone is turbulent and dominated by the vortex-like motion; hence we have adopted the k–ε (RNS) turbulence model (Wilcox 1998) with

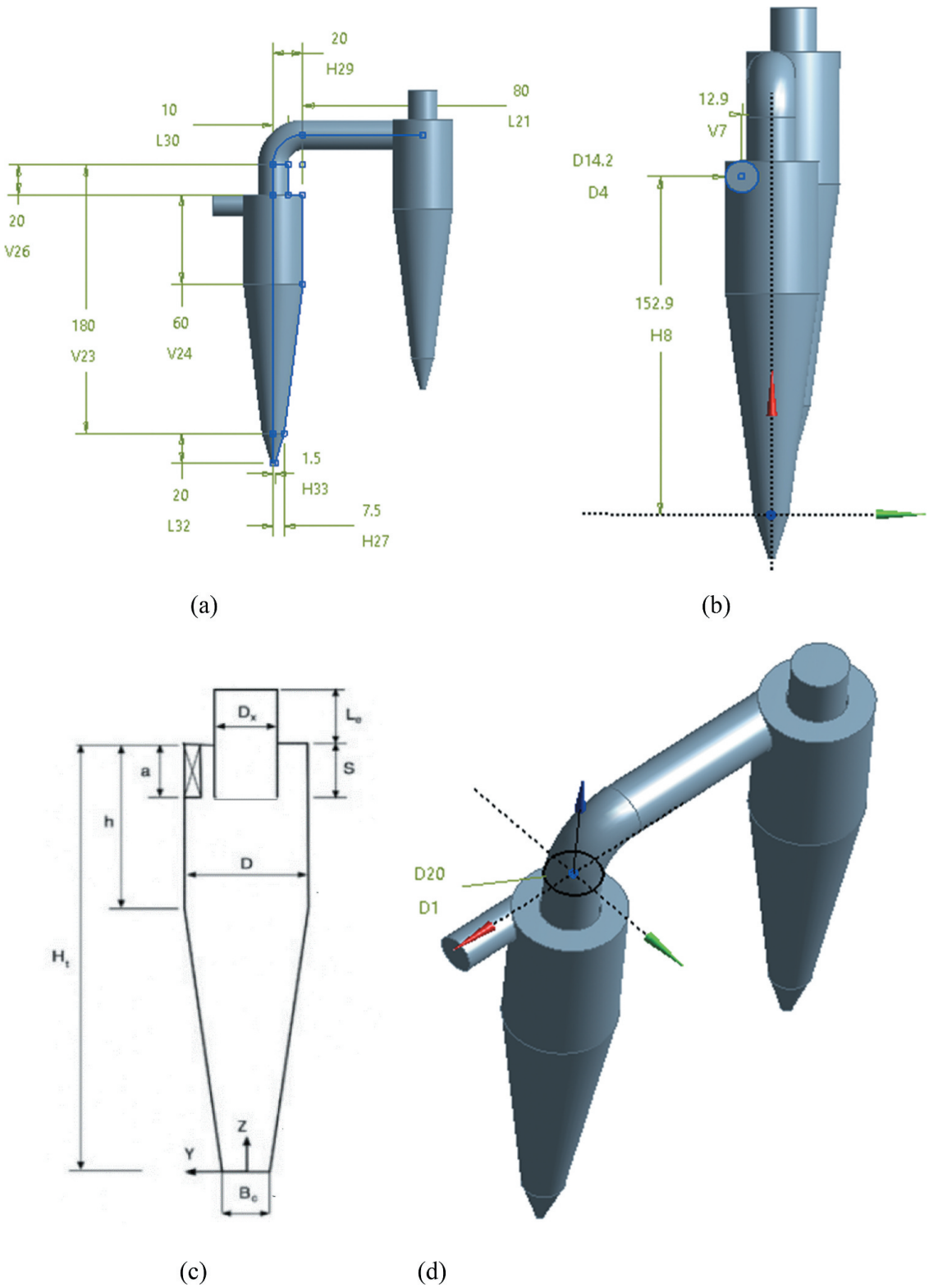


Figure 1. 3D CAD model of cascaded cyclone separator (a) front view (b) side view (c) details of geometrical ratios for a cyclone (d) 3-D View.

enhanced wall treatment with curvature corrections. To track the aerosol particle trajectory, the Lagrangian method-based DPM is employed. As multiphase fluid (with a combination of air with suspended nanosized aerosol particles and water vapor produced inside the fluidic system) is supplied at the inlet of the cascaded cyclone, the volume of fluid (vof) modeling is adopted here to take care of multiphase modeling.

Table 1. Details of geometrical dimensions of cascaded cyclone.

Item description	Dimension in (mm)	
	Cyclone-1	Cyclone-2
Diameter upper cylindrical portion "D"	40	40
Height of upper cylindrical portion "h"	60	60
Diameter of inlet "a"	14.2	14.2
Height of lower conical portion "H_t – h"	180	180
Diameter of bottom outlet "B_c"	4	4
Thickness of cyclone "t"	2	2
Diameter of outlet pipe "D_x"	20	20
Length of outlet pipe inside the cyclone "S"	40	40
Length of outlet pipe outside the cyclone "L_o"	40	40

The Navier–Stokes equation used to evaluate the flow dynamics are the continuity equation

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x_i}(\rho u_i) = 0 \tag{1}$$

where ρ is the density, u_i (u, v, and w) are the velocities in x, y, and z directions, t is time.

The momentum equations are

$$\frac{\partial \rho u_i}{\partial t} + \frac{\partial u_i u_j}{\partial x_j} = -\frac{\partial p}{\partial x_i} + \frac{\partial \tau_{ij}}{\partial x_j} + \rho f \tag{2}$$

and energy equation is

$$\frac{\partial}{\partial t}\left[\rho\left(e + \frac{u_i u_j}{2}\right)\right] + \frac{\partial}{\partial x_j}\left\{u_j\left[\rho\left(e + \frac{u_i u_j}{2}\right) + p\right]\right\} = -\frac{\partial}{\partial x_i}\left(K\frac{\partial T}{\partial x_i} + u_j \tau_{ij}\right) \tag{3}$$

where p is pressure, e is the internal energy, and τ_{ij} are the deviatoric stress tensor calculated with total viscosity, i.e., μ is the laminar viscosity caused by the fluid, and μ_T is the turbulent viscosity caused by the turbulent flow, K is the thermal conductivity of the fluid. The turbulent viscosity, μ_T, is evaluated using the k–ε (RNS) turbulence model (Wilcox 1998) as

$$\frac{\partial(\rho k)}{\partial t} + \frac{\partial(\rho k u_i)}{\partial x_i} = \frac{\partial}{\partial x_i}\left[\left(\mu + \frac{\mu_T}{\sigma_k}\right)\frac{\partial k}{\partial x_i}\right] + G_k + G_b - \rho\varepsilon - Y_M + S_k \tag{4}$$

$$\frac{\partial(\rho \varepsilon)}{\partial t} + \frac{\partial(\rho \varepsilon u_i)}{\partial x_i} = \frac{\partial}{\partial x_i}\left[\left(\mu + \frac{\mu_T}{\sigma_\varepsilon}\right)\frac{\partial \varepsilon}{\partial x_i}\right] + \rho C_1 S\varepsilon - \rho C_2 \frac{\varepsilon^2}{k + \sqrt{\gamma\varepsilon}} \tag{5}$$

The model constant can be given as (Wilcox 1998)

$$C_{1\varepsilon} = 1.44; C_{2\varepsilon} = 1.9; C_\mu = 0.09; \sigma_k = 1.0; \sigma_\varepsilon = 1.2$$

The equations for DPM for tracking particle trajectory can be written in terms of the particle force balance equation as

$$\frac{du_i^p}{dt} = F_d\left(u_i - u_i^p\right) + \frac{g_i\left(\rho_p - \rho\right)}{\rho_p} + \frac{F_i}{\rho_p} \tag{6}$$

where,

$F_d\left(u_i - u_i^p\right)$= drag force function of relative velocity.

$\frac{g_i\left(\rho_p - \rho\right)}{\rho_p}$ = Gravity force(7)

$\frac{F_i}{\rho_p}$= additional force, if acting in the flow (may be added as per flow conditions)

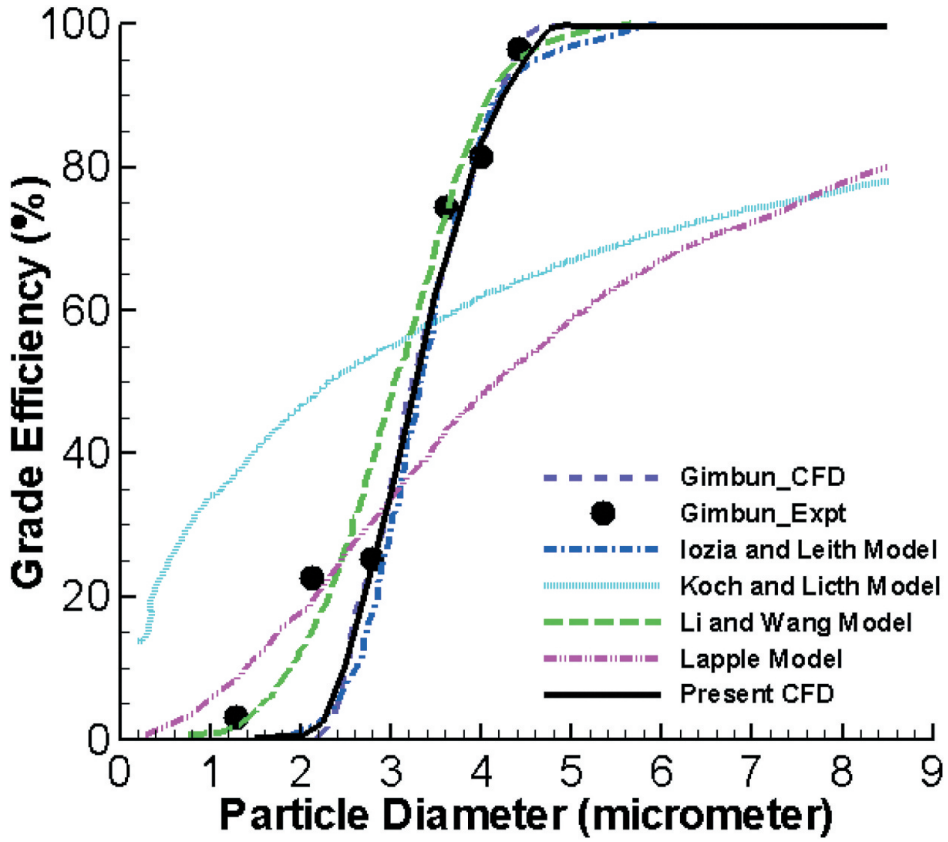


Figure 2. Comparison of collection efficiency of present CFD results produced for Stairmand high-efficiency cyclone with Gimbun et al. (Gimbun et al. 2005).

Based on the above governing equations, two nondimensional numbers can be identified to design the cyclone separators are Reynolds number, Re and Stokes number, Stk defined as

$$Re = \frac{\rho_a u D}{\mu} \tag{8}$$

and

$$Stk = \frac{C_c \rho_p D_p^2}{18 \mu D} \tag{9}$$

where ρ_a is the air density, u is the inlet average velocity of air, D is the inlet characteristics diameter, μ is the air laminar viscosity, C_c is the slip correction factor, ρ_p is the density of the particle, D_p is aerodynamic particle diameter.

Results and discussion

Firstly, a validation test on Stairmand cyclone geometry is carried out before simulating the flow characteristics inside the cascaded cyclone separator. This will provide the validation of the model setup adopted used in commercially Ansys-Fluent software. For validation purposes, we have created a 3 D model of a single cyclone separator with similar geometries, as Gimbun et al (2005), which is a basic geometry for Stairmand high-efficiency cyclone. We have simulated results for similar conditions mentioned by Gimbun et al. ie. inlet velocity of 5–20 m/s, suspended particle size ranges from 1–10 μm having a similar density as mentioned in (Gimbun et al. 2005). Figure 2 shows the comparison of grade efficiency obtained for the single cyclone having an inlet diameter of D = 0.305 and for inlet velocity of 15 m/s, as mentioned in (Gimbun et al. 2005). It can be seen that present CFD results

matched very well with CFD and experimental results produced by Gimbun et al. (Gimbun et al. 2005) along with separation efficiency obtained by various analytical models, (Avci and Karagoz 2001; Barth 1956; Zhao 2004). From Figure 2, it can also be seen that the current CFD setup used in Ansys Fluent software for simulating cyclone separators is producing results which are matching closely with Li and Wang model.

Now, to study the cascaded cyclone separator, we have used the same Ansys-Fluent setup, as validated in the above section. The incompressible turbulent multiphase flow inside the cascaded cyclone separator is simulated using the k–ε turbulence model. The volume of flow formulation is used to take care of multiphase flow inside the cyclone. The suspended particles are tracked using the DPM. Figure 2 shows the detailed three-dimensional model of the cascaded cyclone separator along with all necessary dimensions mentioned in Table 1. Figure 3 shows the detailed view of the meshing generated using ICEM CFD software for cascaded cyclone separator geometry. Four different meshes: mesh-1: 96356, mesh-2: 16458, mesh-3: 260686, and mesh-4: 35768 mixed elements are tested for mesh convergence study. The variation in the maximum velocity attended at the various section of the cyclone separator as well as the outlet velocity and pressure drop across the cascaded cyclone is observed during the mesh convergence study. As unstructured mesh with mixed elements is utilized here hence the mesh refinement is carried out only at necessary locations instead of increasing the mesh across all the cyclone domains. Based on the mesh convergence study, the final unstructured mesh used for the current computations contains 260686 mixed elements with 52591 nodes (i.e. mesh-3 is selected). Note that we have created the boundary layer meshes near all the walls of the cyclone to satisfy the no-slip condition properly.

As per boundary conditions are concerned, no-slip boundary conditions are mentioned at all cascaded cyclone walls with "stick" boundary conditions specified for suspended particles. The "stick" boundary condition assumption can be justified as inlet flowing fluid is a mixed multiphase system

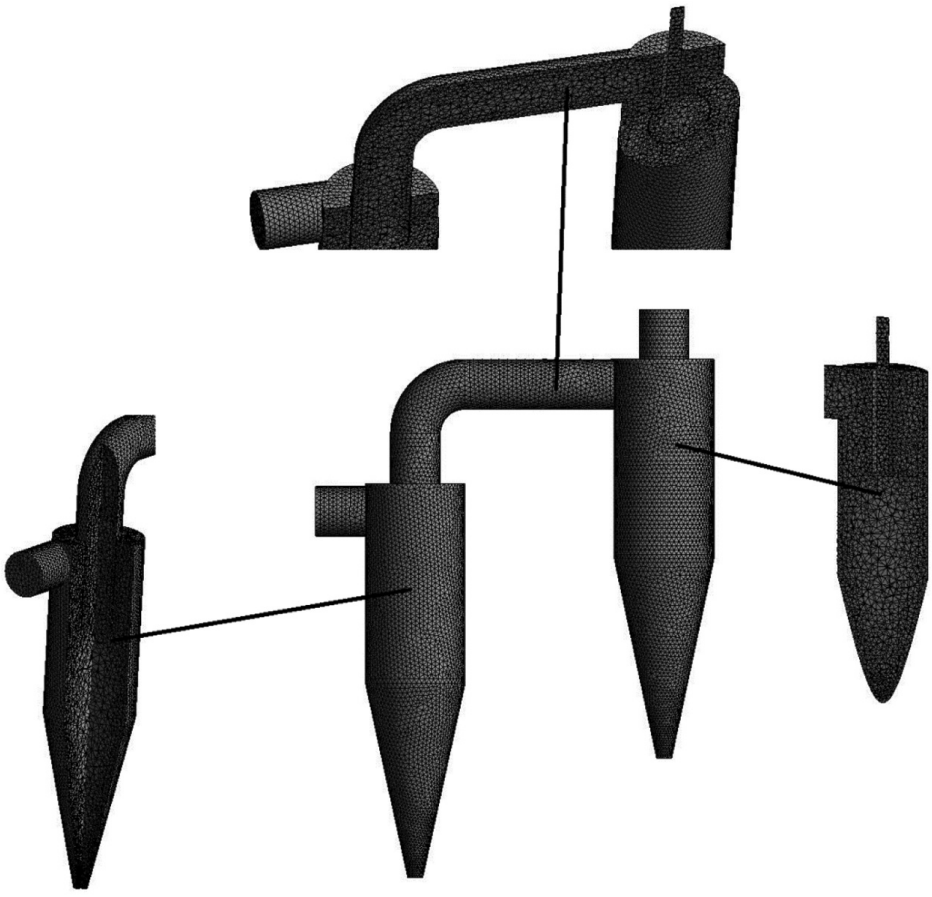


Figure 3. Unstructured mesh generated for cascaded cyclone separator.

Table 2. Details of various airborne viruses [30].

Virus name	Family	Size (nm)	Particle motion
Influenza virus	Orthomyxoviridae	80–120	Rosin
Newcastle disease virus	Paramyxoviridae	150	Rammler Distribution,
Vesicular stomatitis virus	Rhabdoviridae	60 × 200	Weibull
Human coronavirus 229E	Coronaviridae	120–160	Distribution,
Rotavirus	Reoviridae	100	
Rhinotracheitis virus	Herpesviridae	200	

containing air, water vapor, and suspended particles. The water vapor helps to form nucelli at the suspended particles and gets stick to the cyclone wall instead of rebounding from the wall as and when they reach the cyclone walls. The water vapor generated in the fluidics system also makes the suspended particle heavier and larger in diameter compared to the original ones. Both the bottom outlet of the cascaded cyclone as well as an outlet of the second cyclone is open to the atmospheric pressure, hence pressure far-field condition is applied here. At the inlet, for mentioned cross-section, the inlet velocities corresponding to the prescribed flow rate (100–900 lpm) are supplied as velocity inlet boundary conditions. All three species, i.e., air, suspended particles representing the virus here, and water vapors generated in the fluidics system, are assumed to have the same inlet velocity. The volume fraction of suspended particles as well as the size of suspended particles is controlled through DPM inputs which are supplied as an injection.

In this paper, we attempted to capture and separate airborne viruses, i.e., aerosol to be used for bioscience applications. Hence, the suspended particle properties are considered similar to various virus properties, as mentioned by Verreault et al. 2008 . Although we have not explicitly used virus details, however, sufficient data (virus size, density, and particle motion) is considered to mimic the wide range of viruses motion in the air. Table 2 shows the details of few airborne virus properties that can be followed within the scope of the present numerical study.

The particle movement is assumed to follow Weibull distribution, for which the Rosin Rammler model is used for distributing the mass flow among the different sizes of the particle as

$$Y_d = 1 - e^{\left(-\frac{d}{\underline{d}_0}\right)^n} \tag{10}$$

where Y_{d} is the cumulative fraction of the particles by weight having a diameter less than d, $\underline{d}_0$ is characteristics particle size considered, and n is the size distribution constant.

Here, we have used suspended particles, which represent the airborne viruses, having a diameter in the range of 90 nm to 50 μm with a constant average density of 1.2 kg/m^3 [30]. We have injected around 8900 numbers of suspended particles through the spray injection function from the inlet, which will be following the Rosin Rammler distribution as per equation (10). It is to be noted that, in the actual scenario, the concentration of the viruses is in ppm, hence to capture a sufficient number of viruses, a large amount of air should be passed through a cyclone separator. It can be done in two ways – either by running the separation process for several hours or by providing contagious air having sufficient virus particles for a short period. Both represent possible cases of actual and lab testing scenarios, respectively. In the end, to obtain a steady solution during each CFD simulation, the residuals for all the flow variables are ensured to converged less than 10^{-6} levels.

As mentioned in the literature review and referring to Figure 2, it can be seen that the efficiency of a single Stairmand cyclone reduces with a reduction in the suspended particle size. As per Figure 2, the separation efficiency reduces below 50% for the particle size of 4 μm and below. Hence, the same Stairmand geometry needs to be re-develop or re-designed to get a new geometry based on the design

of the numerical experiment strategy. We have modified this Stairmand high-efficiency cyclone geometry by performing several numerical CFD simulations experiments. The final design is presented in Figure 1, and CFD simulation results for the same geometry are discussed here.

The cascaded cyclone performance is now studied for the various inlet flow rate of air. Note that inlet flow rate is the sum of a flow rate of air, a flow rate of suspended particles, and the flow rate of added water vapors. The inlet flow rates vary from 100 lpm to 900 lpm. The aerosol particles trajectory, the pressure drop across both the cyclones are studied. The fluid flow dynamics are also studied by comparing the velocities inside the cascaded cyclone separator. In all the numerical simulations, the number of particles collected and the range of diameter collected in both the cyclones are studied.

Figures 4–8 shows the particle trajectory tracked and pressure contours obtained for an inlet flow rate of 100 lpm, 300 lpm, 500 lpm, 700 lpm, and 900 lpm, respectively. From Figures 4(a)–8(a), it can be seen that, as the flow rate increases, more and more particles are collected in the first cyclone. For the 100 lpm Figure 5(a) it can be seen that more number of particles circulating in the second cyclone, indicating that more number of particles escaped from the first cyclone to the second cyclone. With an increase in the flow rate, the number of particles escaping from the first cyclone reduces hence more and more particles get separated in the cascaded cyclone. Figure 8(a) shows the particles that escape from the first cyclone, get collected in the second cyclone effectively increasing the cyclone performance. This qualitatively indicates that the cyclone separation efficiency increases with an increase in the inlet flow rate for the current configurations.

Figure 4(b)–8(b) shows the pressure contours obtained for the cascaded cyclone separator. It can be seen that the pressure drop for 100 lpm is just 120 Pa which increases to 9412 Pa for the 900 lpm. The increased pressure drop is not affecting the cyclone separation efficiency. However, it decides the pumping power required to operate the cyclone. Hence the work required for the operation of cascaded cyclone separator increases almost linearly with an increase in flow rate from 100 lpm to 900 lpm.

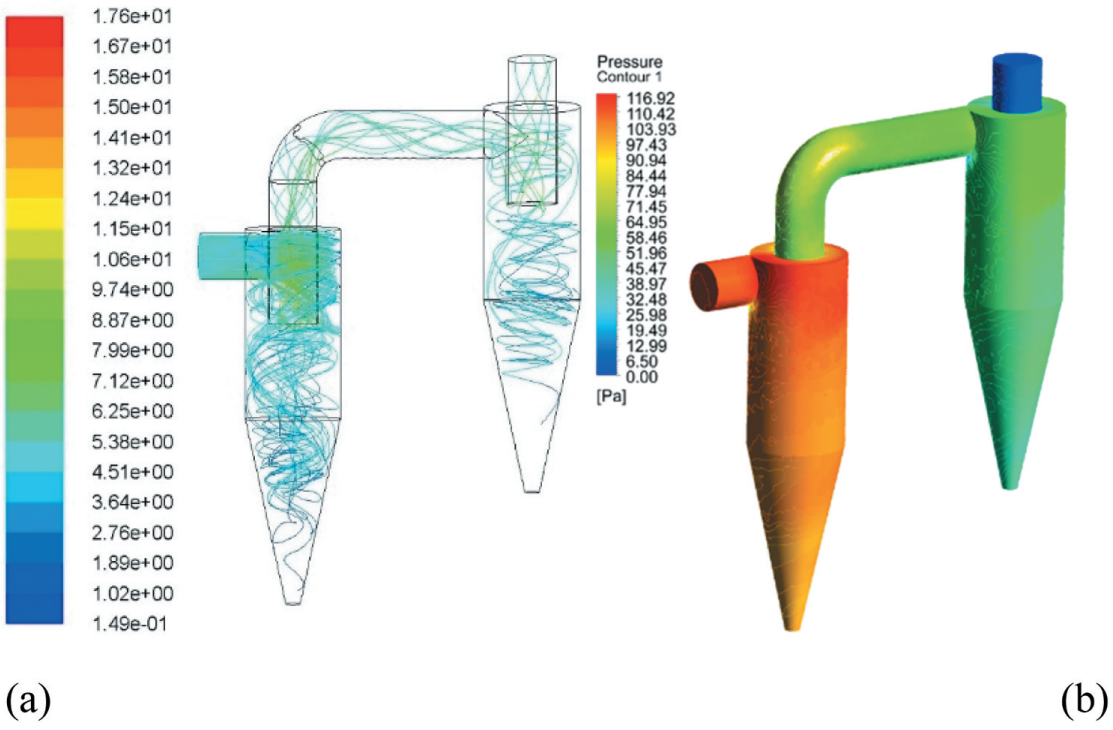


Figure 4. Cascaded cyclone CFD results for (a) particle trajectory and (b) pressure contours obtained at 100 lpm.

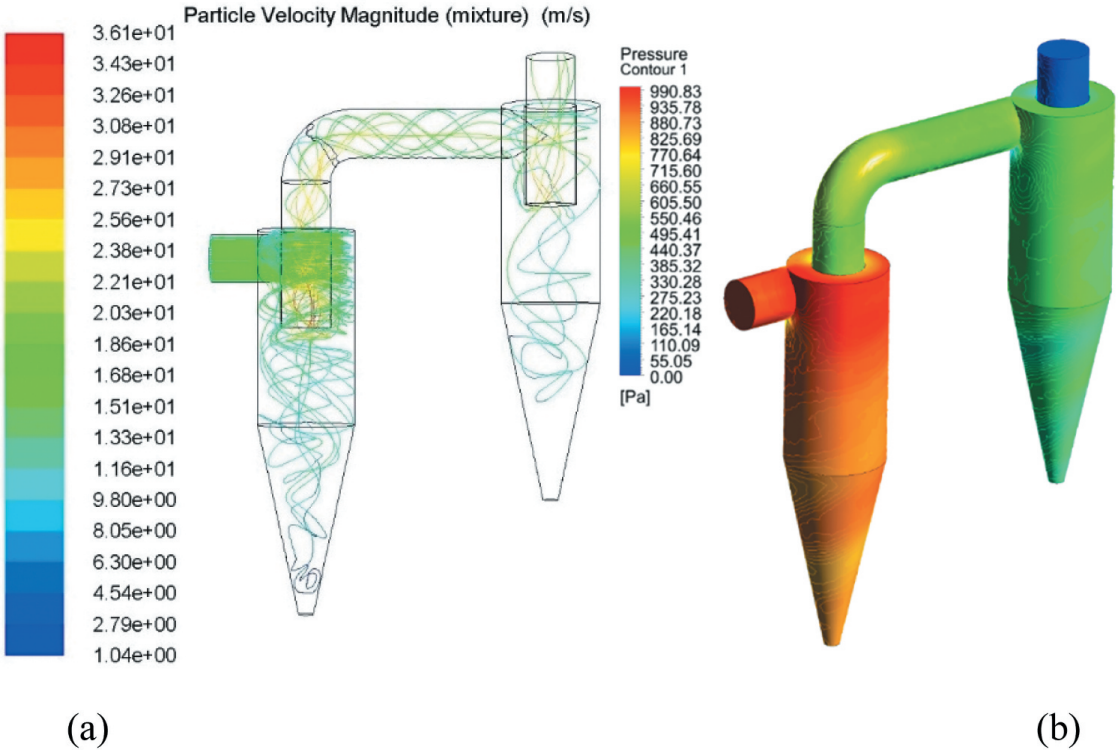


Figure 5. Cascaded cyclone CFD results for (a) particle trajectory and (b) pressure contours obtained at 300 lpm.

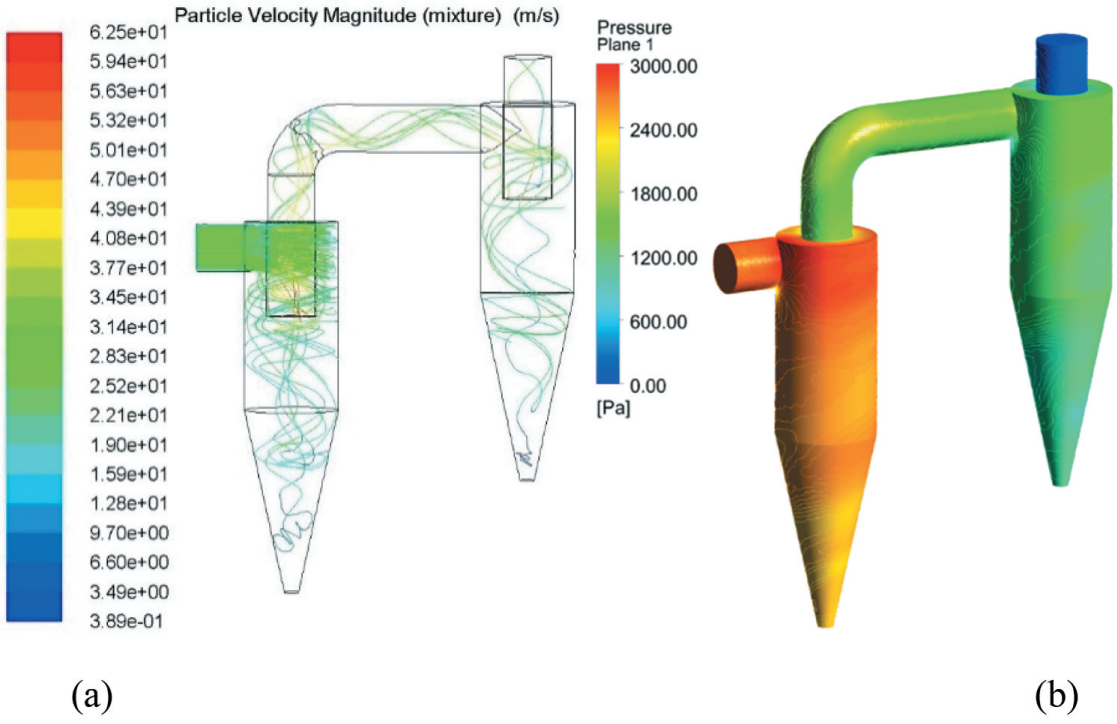


Figure 6. Cascaded cyclone CFD results for (a) particle trajectory and (b) pressure contours obtained at 500 lpm.

From Figures 4–8, it can also be seen that the inlet fluid enters tangentially in both the cyclone, and by shape and design, it forms a strong swirling vortex kind of motion inside cyclones. Two kinds of forces mainly dominate the vortex motion. First, the strong centrifugal force formed due to vortex kind of fluid motion and balancing drag forces.

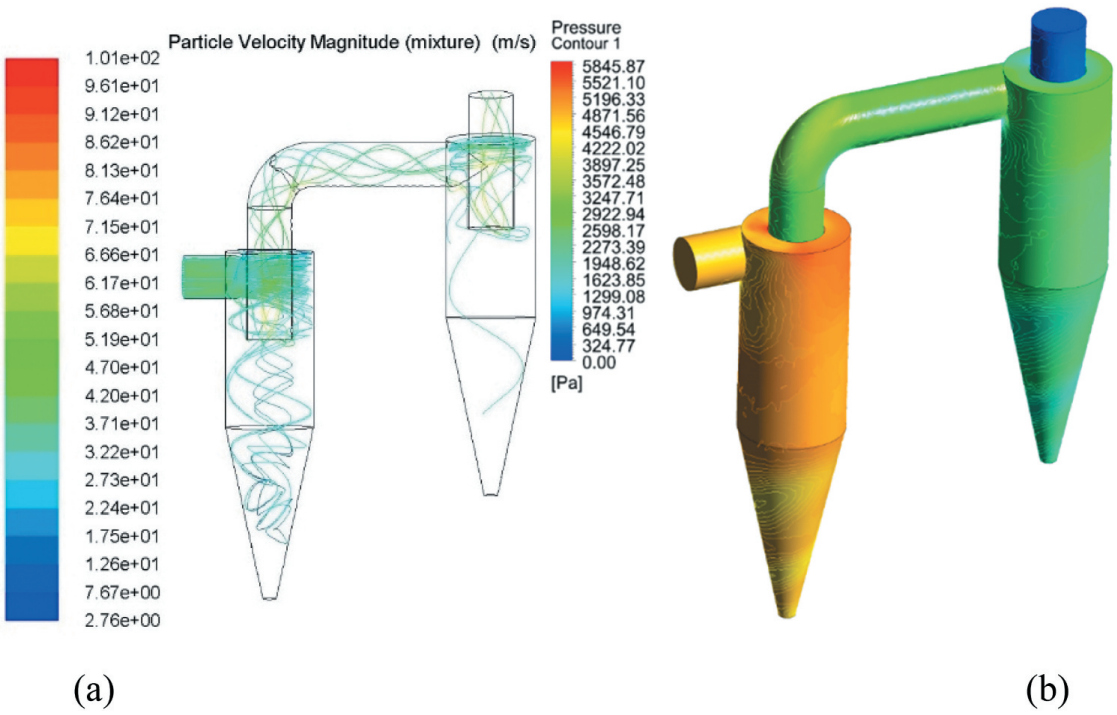


Figure 7. Cascaded cyclone CFD results for (a) particle trajectory and (b) pressure contours obtained at 700 lpm.

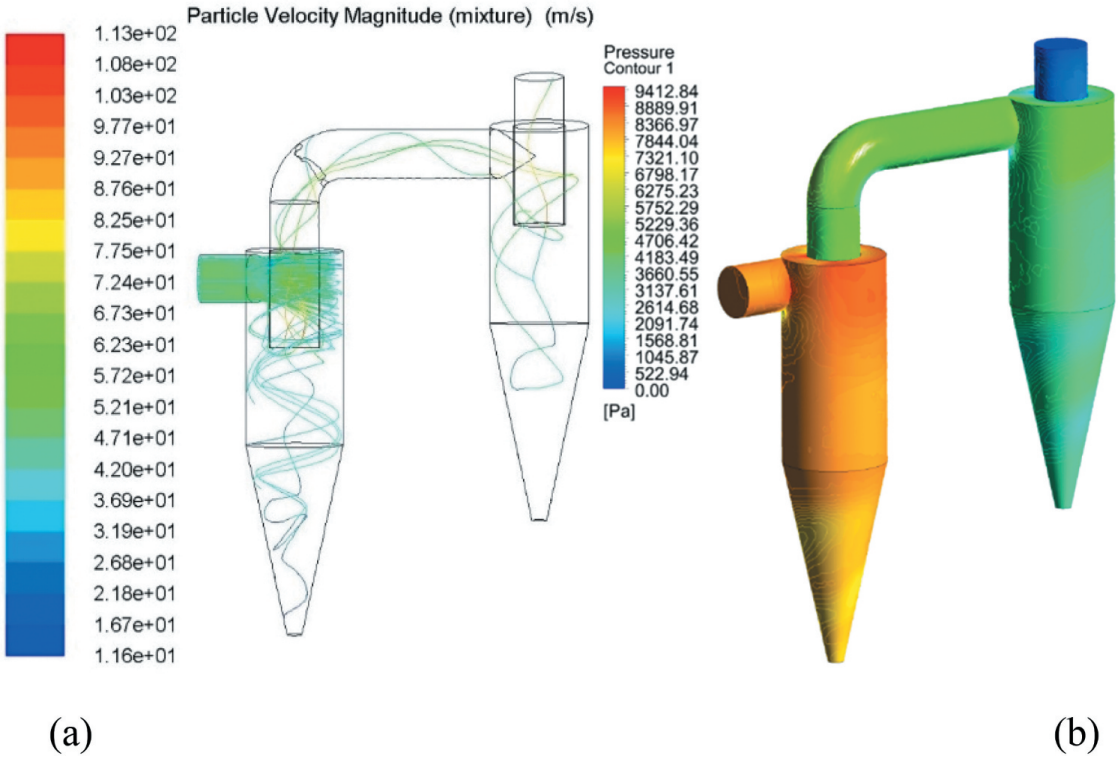


Figure 8. Cascaded cyclone CFD results for (a) particle trajectory and (b) pressure contours obtained at 900 lpm.

To study the flow dynamics, which also reveals the intensity of the forces acting inside the cascaded cyclone separator, we have plotted the total magnitude of the fluid velocity (which represents the kinetic energy) at various cyclone sections. Figure 9 shows the schematic of the cascaded cyclone separator indicating four different cross-sections: A–A, B–B, C–C, and D–D on both the cyclones where velocity is extracted and plotted. Section A–A is located near the bottom exit of the cyclones, B–

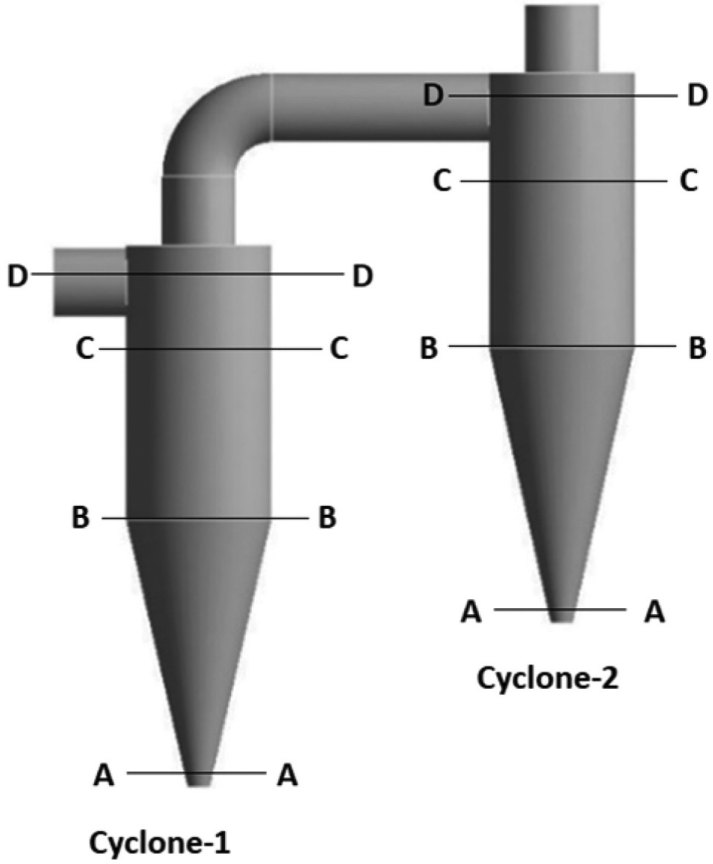


Figure 9. Schematic information of various cross-sections of the cascaded cyclone where flow dynamics are studied.

B is the cross-section where the cylindrical part of the cyclone connects to the conical parts, C–C is the cross-section where starting point/line of vortex finder (not visible here, but please refer Figure 1) and D–D cross-section corresponding to the level centerline of the inlet, however, this section is perpendicular to the inlet.

Figure 10 shows the variation of total velocity plotted at cross-section D–D, i.e., at the inlet of both the cyclone of the separator for various inlet flow rates. It can be seen that the fluid enters with the high inlet tangential velocity, which helps to form a vortex inside the cyclone. Please note the central part of the plot, which shows the outlet velocity through the vortex finder. From the figure, it can be seen that due to core vortex dynamics, the outlet velocity from the vortex finder is larger in magnitude than the fluid inlet velocity. It can also be observed that the difference in the magnitude of outlet velocity and inlet velocity grows with an increase in the flow rates. Interestingly, the magnitude of the maximum velocity reached in both the cyclone is similar, indicating that serial coupling is effective for removing more and more particles with smaller and smaller diameters.

Figure 11 shows the variation of velocity plotted at section C–C, located just above the vortex finder inlet. For a low flow rate of 100 lpm – 500 lpm, the velocity distribution is not symmetric. The effect of an outer vortex and central core vortex can distinctly be visible in cyclone-1. As the flow rate increases, the increased kinetic energy of fluid brings more uniformity in the velocity magnitude for cyclone-1. However, the reverse scenario is visible in cyclone-2 as compared to cyclone-1. As cyclone-2 outlet is exposed to atmospheric pressure, the central vortex core rushes to the outlet more quickly and attends a maximum velocity in cyclone-2, which is much larger than cyclone-1 case.

Figure 12 shows the variation of velocity plotted at sections B–B; this section is located almost in the middle of the cyclone height. The velocity distribution is uniform and near to the symmetric one. This indicates that a vortex develops from the inlet and attends uniform velocity distribution somewhere near this location. The velocity magnitudes are high and contribution maximum for the particle separation. The

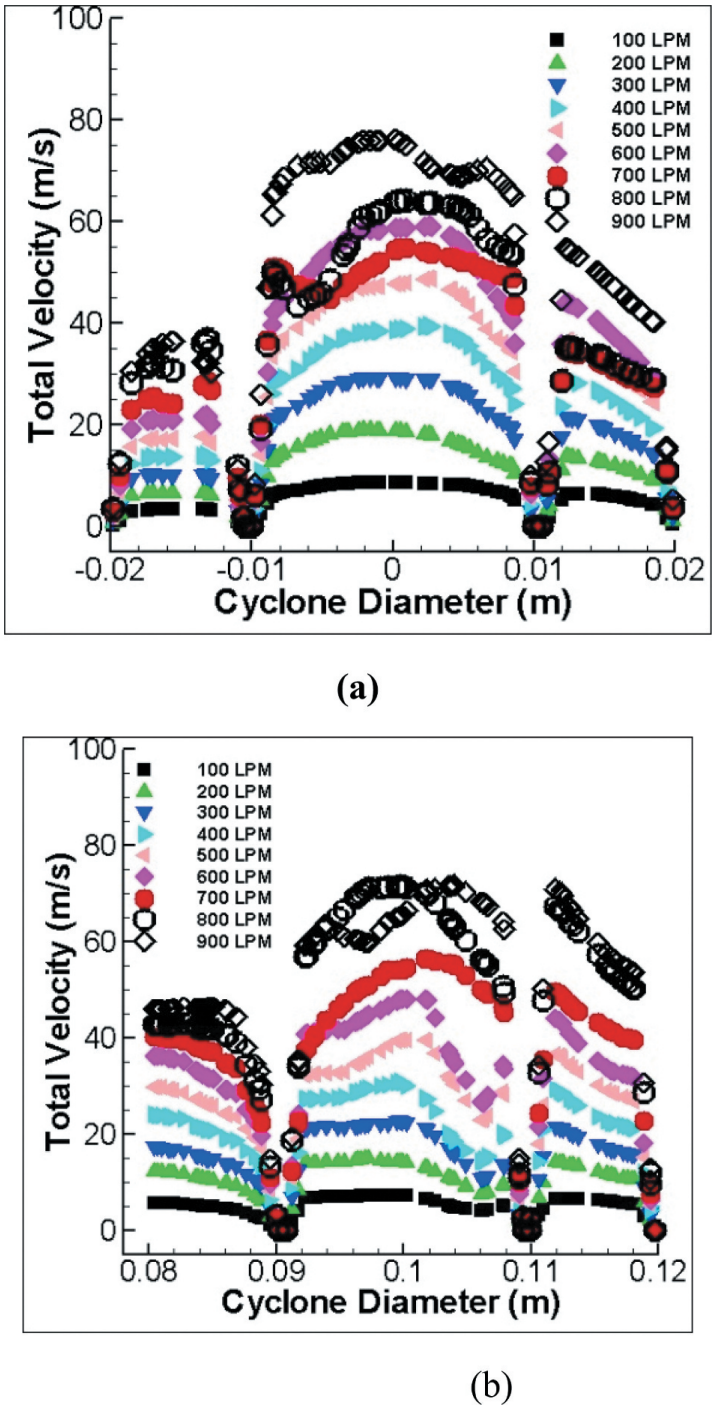
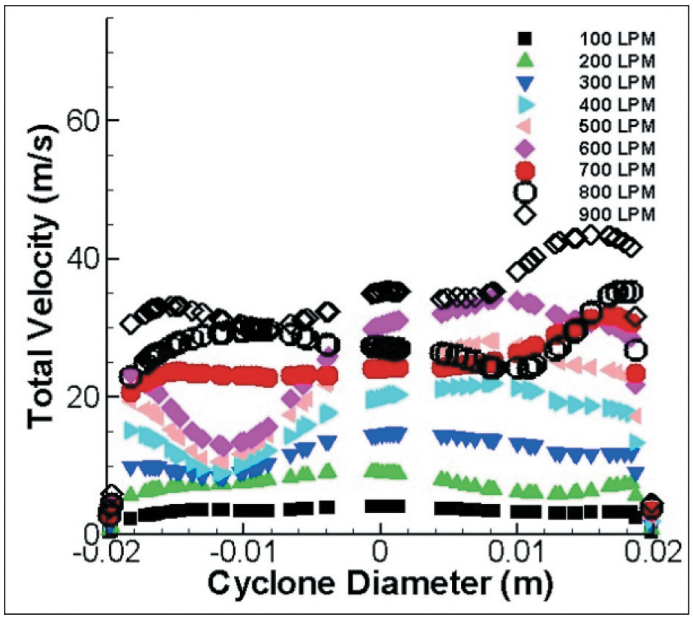


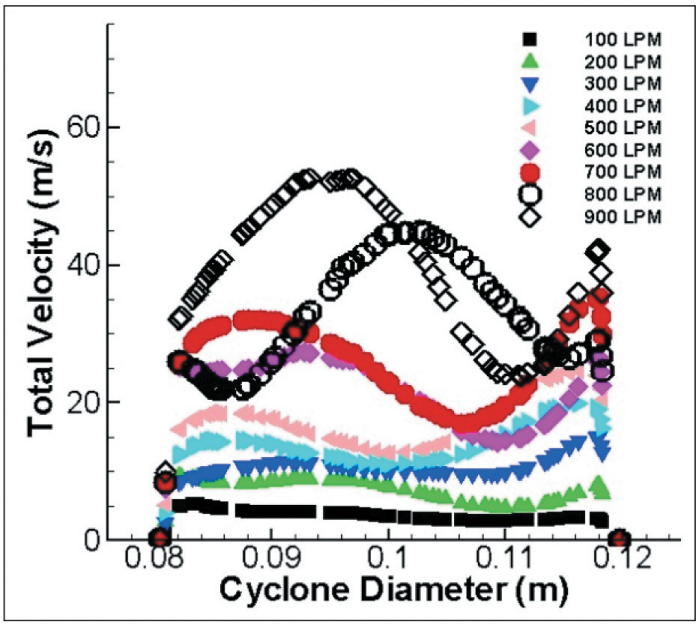
Figure 10. Total velocity plotted at cross-section D–D (a) Cyclone-1 (b) Cyclone-2.

centrifugal forces will also be high in magnitude by this uniform velocity. Figure 13 shows the variation of velocity magnitude plotted at A–A sections located near the bottom exit of cyclones. The outer vortex converges here, resulting in the reduction of the velocity of an outer vortex and forming an inner core vortex from this point upwards. From all these Figures 12 and 13, the total tangential velocity is varying from its inlet value and may attain maximum. These plots are essential for designing and modifying the shape of cyclone separators to obtain high separation efficiency for nanosize aerosol particles.

To study the particle separation characteristics of the cascaded cyclone separator, we have divided the cascaded cyclones into three parts bottom (below section A–A), middle (between section A–A and B–B), and upper (between section B–B and C–C) and studied the particle collected at the wall in these cyclones. Table 3 shows the maximum particles trapped/separated in various zones created on a cascaded cyclone



(a)

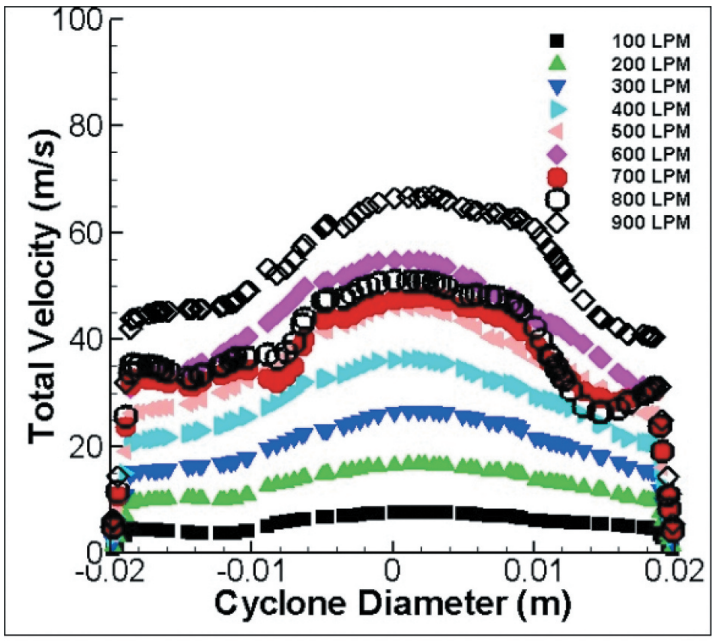


(b)

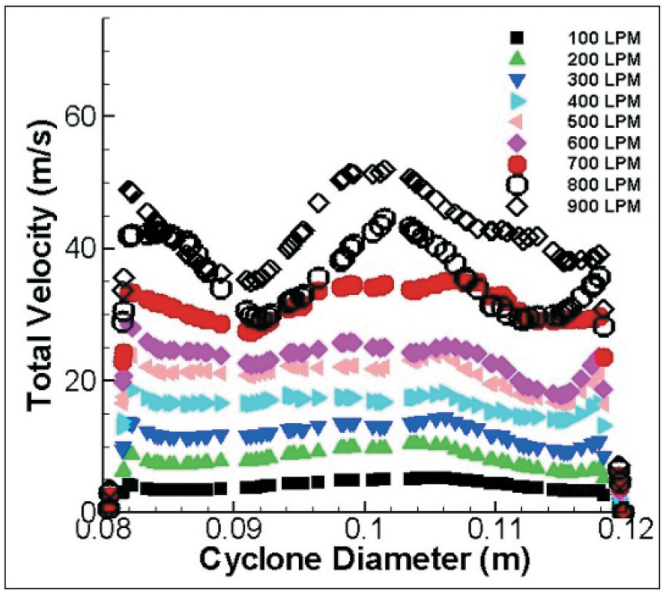
Figure 11. Total velocity plotted at cross-section C–C (a) Cyclone-1 (b) Cyclone-2.

separator for a 100 lpm inlet flow rate. From Table 3, it is clear that maximum particles (around 92%) are trapped and separated in middle1 (zone 2) and upper1 (zone 3) sections. When the remaining particles enter into cyclone-2, the same procedure is going on in the middle and upper part, trapping more particles. Hence it is confirmed that the centrifugal force helps in trapping more particles. Table 4 shows the range of particle diameter captured/separated in various sections for 100 lpm flow rate. From Table 4, it can be seen that from an outlet of the cyclone-2, only 74 particles of size less than 100 nm escaped, indicating a very high separation efficiency of the present designed cascaded cyclone separator.

Figure 14 shows the distribution of particles escaped at the outlet of the cascaded cyclone separator for different flow rates. The figure shows that the particle size of the particle escaping from the outlet is around 100 nm, indicating that all larger particles (than 100 nm) are captured and separated using the present cascaded cyclone separator. The figure shows that as the flow rate increases from 100 to 900



(a)

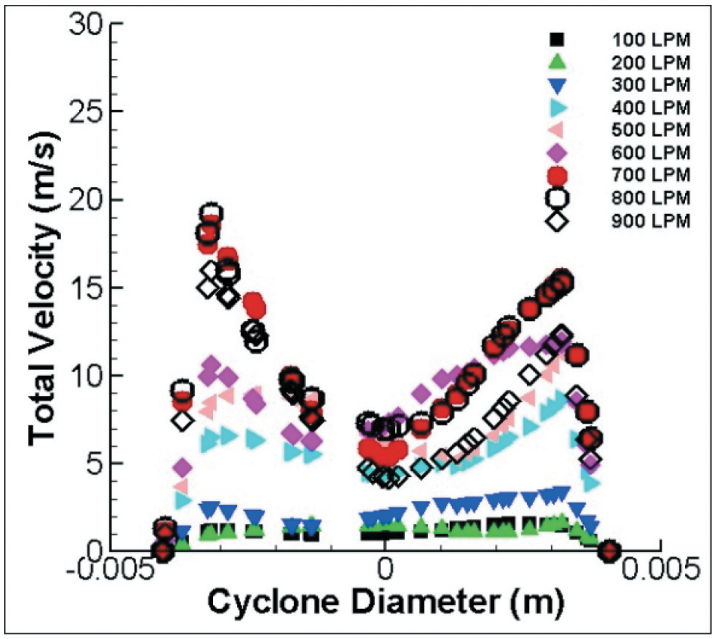


(b)

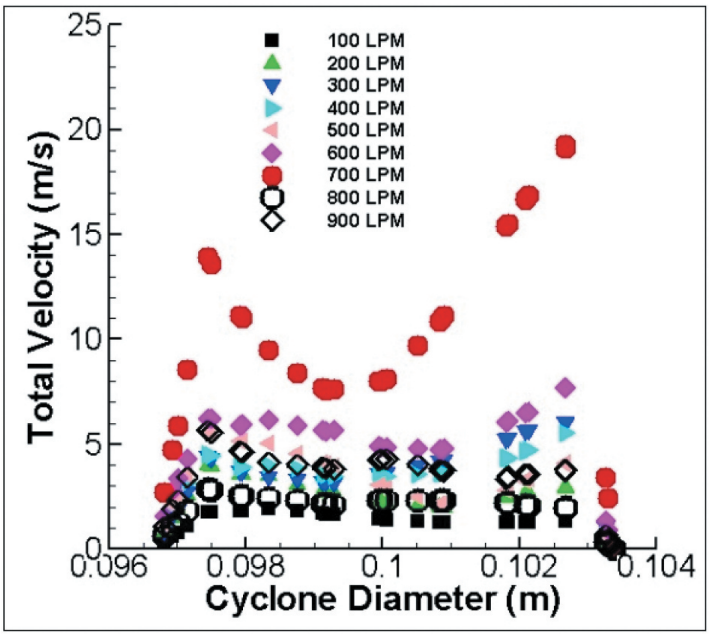
Figure 12. Total velocity plotted at cross-section B–B (a) Cyclone-1 (b) Cyclone-2.

Table 3. Particle distribution for 100 lpm.

Part of cascaded cyclone	Zone identity	Number of particles collected at respective zones	Percentage of particle collected/ separated
Bottom-1	Zone 1	169	1.89
Middle-1	Zone 2	6023	67.40
Upper-1	Zone 3	2208	24.71
Elbow	Zone 4	95	1.06
Bottom-2	Zone 5	48	0.54
Middle-2	Zone 6	91	1.02
Upper-2	Zone 7	55	0.62
Escaped from cascaded cyclone	Zone 8	247	2.76
Total particle injected		8936	100



(a)



(b)

Figure 13. Total velocity plotted at cross-section A–A (a) Cyclone-1 (b) Cyclone-2.

Table 4. Size-wise separation of particles for 100 lpm.

Particle size	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 6	Zone 7	Zone 8
less than 100 nano	9	25	18	12	8	22	8	74
100 nano to 5000 nano	56	182	99	59	38	65	42	168
5000 nano to 10000 nano	91	655	106	24	2	4	5	5
More than 10000 nano	13	5161	1985	0	0	0	0	0
Total Particles	169	6023	2208	95	48	91	55	247

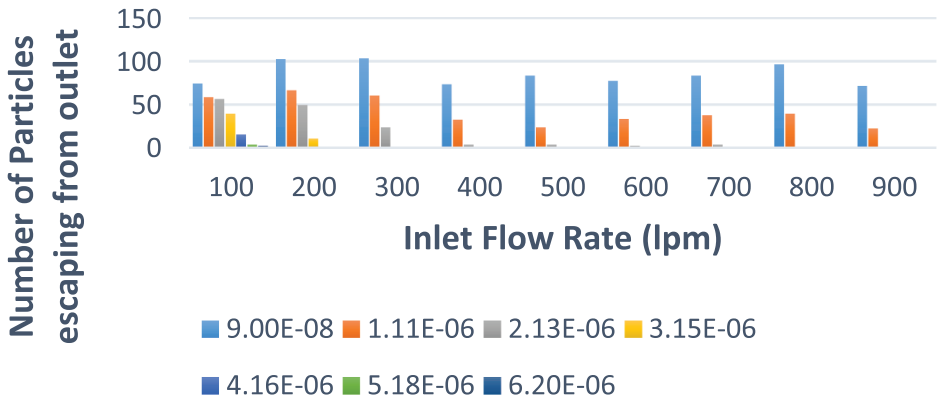


Figure 14. Comparison of particle (number and size) escaping from the outlet of cascaded cyclone separator for various inlet flow rates.

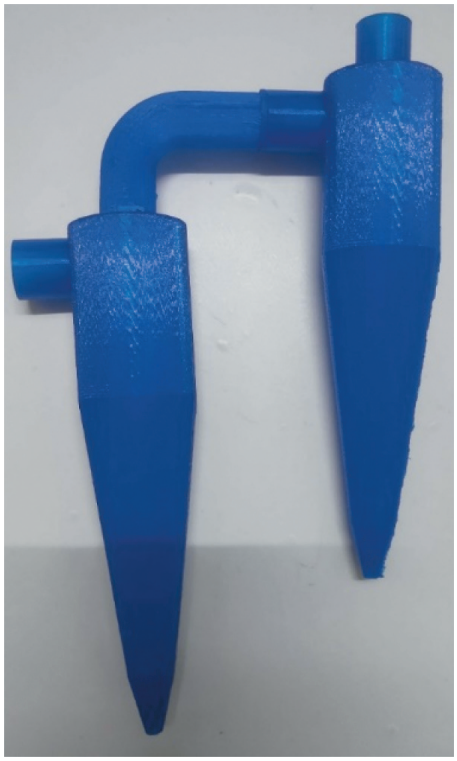


Figure 15. 3D printed cascaded cyclone separator.

lpm, the distribution of escaped particle sizes is reduced drastically. Hence quantity, as well as the size of the particle escaping from the cyclone separator, could greatly be improved by using higher mass flow rates for the present cascaded cyclone separator.

Conclusion

In this paper, computational fluid dynamics-based numerical simulation of a cascaded cyclone separator to be used for separating aerosol representing airborne viruses is presented. The cascaded cyclone separator is redesigned from the Stairmand single cyclone geometry using the design of numerical experiment (DOnE) method. The performance of the cascaded cyclone separator is then numerically evaluated for various flow rates (100–900 lpm). Based on the numerical simulation conducted, it is concluded that the high separation efficiency of the cascaded cyclone could be obtained by properly designing and selecting the appropriate geometrical ratio for the cyclone geometry. For the current cascaded cyclone configuration, due to the high inlet

flow velocity of the air and the large centrifugal force generated, cyclone-1 of the cascaded cyclone is capable to separate the majority of the heavier and larger size aerosol particles. These larger sizes and heavier particles are trapped mostly in the middle and upper part of the cyclone-1. The average size of aerosol particles escaping from cyclone-1 has a diameter smaller than 5 μm. Hence, cyclone-2 will be forced to separate lighter and smaller aerosol particles escaped from cyclone-1. So, cascaded cyclone separator efficiency is mostly contributed from cyclone-1 however cyclone-2 is contributing to improving it further especially for nanosize aerosol particles. The outlet of the cascaded cyclone rejects mostly lighter particles having a size below 90 nm for all mass flow rates. From the numerical analysis, it can also be concluded that increasing the inlet flow rate increases the cyclone separation efficiency, especially for cascaded configuration. The efficiency for all flows rate is found to be more than 93% (for an inlet flow rate of 100 lpm) and can reach a maximum value of 97% for 900 lpm. Hence the current designed cascaded cyclone could effectively be used for separating aerosol – airborne viruses and bacteria having nanometer sizes. Now the next task in near future is to develop an experimental setup of cascaded cyclone separator designed in the current work, to start with we have fabricated the cascaded cyclone (refer to Figure 15) using 3 D printing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Dr. Sonawane Chandrakant R. is currently working as an Associate Professor at the Mechanical Engineering Department, Symbiosis Institute of Technology, Symbiosis International Deemed University, Pune, Maharashtra, India. He received his PhD from the Aerospace Engineering Department, Indian Institute of Technology Bombay, Mumbai, India, in 2013. He has 18 years of teaching experience at the undergraduate as well as postgraduate levels. He published more than 60 publications as the main author and co-author in international journals, conferences, and books. He is also working on several funded projects from the Science and Engineering Research Board, Department of Science and Technology, and Government of India.

Dr. Atul Kulkarni received his PhD degree at the University of Pune, India in 2005. He is currently associated with Symbiosis Centre for Nanoscience and Nanotechnology (SCNN) of Symbiosis International (Deemed University), India, in the capacity of Research Professor and Head. His research interests are to develop multidisciplinary systems and sensors.

Rohan V. Sawant is a full-time research scholar at the Mechanical Engineering Department, Symbiosis Institute of Technology, Symbiosis International Deemed University, Pune, Maharashtra, India. He completed his M.Tech CAD-M from Symbiosis Institute of Technology, Pune in the year 2018. He has a keen interest in the computational fluid dynamics area.

Dr Anand Kumar Pandey is working as a professor and head at the Mechanical Engineering Department, Symbiosis Institute of Technology, Symbiosis International Deemed University, Pune, Maharashtra, India. He received his PhD from the College of Engineering Pune in 2013. He has 25 years of research experience. He has a keen research interest in military equipment design and military vehicle performance enhancements, and he is associated with the defense research field for the last 20 years. His current research includes experimental as well as numerical simulation on internal combustion engines, alternative fuels, and very large power (1000 HP) engines. He is also working in the field of renewal and sustainable energy.

Dr. Hitesh Panchal obtained his Ph.D. in Solar Thermal Engineering in 2015 from KSV University Gandhinagar. He is currently working as Assistant Professor in Mechanical Engineering Department in Government Engineering College Patan, Gujarat. Dr. Panchal has published 100 plus Research papers in reputed International Journal publications like Taylor and Francis, Springer, Elsevier, ASME, etc., and received 4000 plus Google scholar citations with H index 41 and I10 index 99.Dr. Panchal obtained 28 prestigious awards includes Tech GURU Award, Pedagogical Innovation Award, SSIP Prashansha Award, Preeminent Researcher Award, Best Ph.D. Thesis Award, Young Scientist Award, Bright Researchers Award, and many more. Dr. Panchal has completed 2 Research projects from GUJCOST and SYST, DST and currently working on Research project under STEM grant, Education Department, Government of Gujarat. He filed

26 Full patents and 12 Industrial designs, and among them, two full patents and four industrial designs had granted.His areas of interest includes "Solar Desalination, I C Engine emission analysis, Automobile engineering etc. Recently Dr. Panchal is selected in the Top 2 % Indian scientist in Energy Category published by Stanford University.

Dr. Kishor Kumar Sadasivuni is currently working as a Research Assistant Professor and the group leader of SmartNanoSolutions at Center for Advanced Materials, Qatar University. Dr. Kishor's research has its roots from the analytical chemistry he mastered during his Master's degree in Andhra University, India. He received his Ph.D. in Materials Science and Engineering from the University of South Brittany at Lorient, France in 2012 under the supervision of Professor Yves Grohens and Professor Sabu Thomas. He has enormous knowledge and valuable experience in his field owing to over 12 years of active research experience of 12 years. Dr. Kishor has more than 250 research articles published in international peer-reviewed journals with a total citation of 5300 and h-index 39. He is also the author of 20 book chapters and the editor of 8 books. Dr. Kishor is an expert in nanomaterials engineering, especially sensing and flexible electronics. He has a keen interest in the development of different devices based on smart materials by engineering IoT. Some of Dr. Kishor's inventions are protected by 2 US patents and 2 Indian patents and presently 5 patent disclosures have been submitted by him. His strong academic background and passion for research is remarkable and reflected in his fast-paced career growth. Dr. Kishor's enthusiastic and sincere approach towards research works are reflected in the forms of various grants and awards offered to him. He is a group leader and is presently the lead principal investigator for 11 research projects, covering the NPRP, UREP, IRCC grants of Qatar National Research Fund and the Qatar University. He is also involved in other three research grants as the principal investigator, summarizing a total grant amount of 3M$. Dr. Kishor is a team player and has collaborated actively with researchers (more than 450 co-authors as evident from the Scopus data) in several disciplines of computer science, biomedical sciences, industrial engineering, and electrical engineering from all over the world (USA, France, South Korea, Oman, Spain, Italy, Australia, Malaysia). Dr. Kishor's achievements have been recognized by several awards such as Tyre & Rubber Industry Leadership Acknowledgement Awards (TRILA); Young Research Scholar of the Year 2017. He has been included in the world's top 2 % scientists according to a list compiled by Stanford University in the year 2019 and for this he was recently honored by the QU.

Dr Anil kumar is currently working as Assocaite Professor, Mechanical Engineering Department, Delhi Technological University, New Delhi India. He has more than 15 years of teaching and Research institute. He has published more than 100+ research papers and obtained more than 4000+ Google Scholar citations.

ORCID

Sonawane Chandrakant R. http://orcid.org/0000-0002-3408-5060
Atul Kulkarni http://orcid.org/0000-0003-1696-4982
Rohan V. Sawant http://orcid.org/0000-0003-2692-4204
Hitesh Panchal http://orcid.org/0000-0002-3787-9712

References

Avci, A., and I. Karagoz. 2001. Theoretical investigation of pressure losses in cyclone separators. *International Communications in Heat and Mass Transfer* 28 (1):107–17. doi:10.1016/S0735-1933(01)00218-4.

Balestrin, E., R. K. Decker, D. Noriler, J. C. S. C. Bastos, and H. F. Meier. 2017. An alternative for the collection of small particles in cyclones: Experimental analysis and CFD modeling. *Separation and Purification Technology* 184:54–65. doi:10.1016/j.seppur.2017.04.023.

Barth, W. 1956. Design and layout of the cyclone separator on the basis of new investigations. *Brennstow-Wäerme-Kraft (BWK)* 8 (4):1–9.

Behrang, M., M. Shirvani, and S. H. Hashemabadi. 2019. Multi-Helical-Channel dust separator: CFD simulation and experiment. *Chemical Engineering Research & Design* 146:1–10. doi:10.1016/j.cherd.2019.03.025.

Bhaskar, K. U., Y. Rama Murthy, M. Ravi Raju, S. Tiwari, J. K. Srivastava, and N. Ramakrishnan. 2007. CFD simulation and experimental validation studies on hydrocyclone. *Minerals Engineering* 20:60–71. doi:10.1016/j.mineng.2006.04.012.

Bogodage, S. G., and A. Y. T. Leung. 2015. CFD simulation of cyclone separators to reduce air pollution. *Powder Technology* 286:488–506. doi:10.1016/j.powtec.2015.08.023.

Casal, J., and J. M. Martinez-Benet. 1983. A better way to calculate cyclone pressure drop. *Chemical Engineering* 90 (2):99–100.

Chengming, S., P. Binbin, J. Mengting, B. Wang, X. Delong, and C. Yanxin. 2016. Numerical analysis of forces exerted on particles in cyclone separators. *Powder Technology* 294:437–48. doi:10.1016/j.powtec.2016.02.052.

Chuen-Jinn, T., C. Sheng-Chieh, P. Rafal, and M. Arkadiusz. 2007. Study of an axial flow cyclone to remove nanoparticles in vacuum. *Environmental Science & Technology* 41-5:1689–95. doi:10.1021/es060518o.

Cullivan, J. C., R. A. Williams, T. Dyakowski, and C. R. Cross. 2004. New understanding of a hydrocyclone flow field and separation mechanism from computational fluid dynamics. *Minerals Engineering* 17 (5):651–60. doi:10.1016/j.mineng.2004.04.009.

Gimbun, J., T. G. Chuah, S. Y. Choong Thomas, and A. Fakhru'l-Razi. 2005. A CFD Study on the prediction of cyclone collection efficiency. *International Journal for Computational Methods in Engineering Science and Mechanics* 6 (3):161–68. doi:10.1080/15502280590923649.

Giwoon, S., A. Chisung, K. Atul, S. W. Gyu, and K. Taesung. 2017. Highly efficient in-line wet cyclone air sampler for airborne virus detection. *Journal of Mechanical Science and Technology* 31 (9):4363–69. doi:10.1007/s12206-017-0835-4.

Houben, J. J. H., C. Weiss, E. Brunnmair, and S. Pirker. 2016. CFD simulations of pressure drop and velocity field in a cyclone separator with central vortex stabilization rod. *Journal of Applied Fluid Mechanics* 9 (1):487–99. doi:10.18869/acadpub.jafm.68.224.23934.

Kabeel, A. E., M. Elkelawy, H. A. Mohamad, A. M. Elbanna, H. Panchal, M. Suresh, and M. Israr. 2020. The influences of loading ratios and conveying velocity on gas-solid two phase flow characteristics: A comprehensive experimental CFD-DEM study. *International Journal of Ambient Energy*. doi:10.1080/01430750.2020.1758777.

Khairy, E., and L. Chris. 2010. Optimization of the cyclone separator geometry for minimum pressure drop using mathematical models and CFD simulations. *Chemical Engineering Science* 65 (22):6048–58. doi:10.1016/j.ces.2010.08.042.

Khoi, L. D., and Y. J. Yong. 2020. Numerical investigation on the performance and flow pattern of two novel innovative designs of four-inlet cyclone separator, *Chemical Engineering & Processing: Process Intensification* 150:107867. doi:10.1016/j.cep.2020.107867.

Liu, E., D. Tian, W. Li, J. Chen, and Q. Chen. 2021. Study on erosion behavior and separation efficiency of a shale gas vertical separator. *Energy & Fuels* 35 (5):3878–86. doi:10.1021/acs.energyfuels.0c03979.

Mandal, J. C., and C. R. Sonawane. 2014. Simulation of moderator flow and temperature inside calandria of CANDU reactor using artificial compressibility method. *Heat Transfer Engineering* 35 (14–15):1254–66. doi:10.1080/01457632.2013.876802.

Mandal, J. C., C. R. Sonawane, A. R. Iyer, and S. J. GosaviInamdar. 2011. Incompressible flow computations over moving boundary using a novel upwind method. *Computers & Fluids* 46 (1):348–52. doi:10.1016/j.compfluid.2010.08.011.

Nakhaei, M., B. Lu, Y. Tian, W. Wang, K. Dam-Johansen, and H. Wu. 2020. CFD modeling of gas–solid cyclone separators at ambient and elevated temperatures. *Processes* 8:228. doi:10.3390/pr8020228.

Nassaj, O. R., and D. Toghraie. 2019. Masoud afrand, effects of multi inlet guide channels on the performance of a cyclone separator. *Powder Technology* 356:353–72. doi:10.1016/j.powtec.2019.08.038.

Panchal, H. N., and N. Patel. 2018. ANSYS CFD and experimental comparison of various parameters of a solar still. *International Journal of Ambient Energy* 39 (6):551–57. doi:10.1080/01430750.2017.1318785.

Panchal, H. N., and P. K. Shah. 2013. Modeling and verification of hemispherical solar still using ANSYS CFD. *International Journal of Energy and Environment* 3 (4):427–40.

Park, D., and J. S. Go. 2020. Design of cyclone separator critical diameter model based on machine learning and CFD. *Processes* 8:1521. doi:10.3390/pr8111521.

Patankar, S. V. 1980. *Numerical heat transfer and fluid flow*. Taylor & Francis. ISBN 978-0-89116-522-4.Boca Raton. https://doi.org/10.1201/9781482234213

Paweł, K., K. Adam, A. Białecki Ryszard, and P. Adamczyk Wojciech. 2017. Numerical simulation of a dense solid particle flow inside a cyclone separator using the hybrid Euler–Lagrange approach. *Particuology* 31:170–80. doi:10.1016/j.partic.2016.09.003.

Qiang, L., W. Qinggong, X. Weiwei, Z. Zilin, and Z. Konghao. 2019. Experimental and computational analysis of a cyclone separator with a novel vortex finder. *Power Technology* S0032-5910(19):30896–4. doi:10.1016/j.powtec.2019.10.073.

Selami, D. 2014. A practical model for estimating pressure drop in cyclone separators: An experimental study. *Powder Technology* 268:329–38. doi:10.1016/j.powtec.2014.08.024.

Sonawane, C. R., J. C. Mandal, and S. P. Roa. 2019. High-resolution incompressible flow computations over unstructured mesh using SDWLS gradients. *Journal of the Institution of Engineers (India): Series C (Scopus)* 100 (1):83–96. doi:10.1007/s40032-017-0390-x).

Sonawane, C. R., Y. B. More, and A. Pandey. 2021. Numerical simulation of unsteady channel flow with a moving indentation using solution dependent weighted least squares based gradients calculations over unstructured mesh. *Heat Transfer Engineering Journal* 2021. doi:10.1080/01457632.2021.1874661.

Sonawane, C. R., H. Panchal, and K. K. Sadasivuni. 2021. Numerical simulation of flow-through heat exchanger having helical flow passage using high order accurate solution dependent weighted least square based gradient calculations. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*. doi:10.1080/15567036.2021.1900457.

Verreault, D., S. Moineau, and C. Duchaine. 2008. Methods for sampling of airborne viruses. *Microbiology and Molecular Biology Reviews : MMBR* 72 (3):413–44. doi:10.1128/MMBR.00002-08.

Wilcox, D. C. 1998. *Turbulence modeling for CFD*, 2nd ed., Vol. 1998, 174. Anaheim: DCW Industries.

Zhao, B. 2004. A theoretical approach to pressure drop across cyclone separators. *Chemical Engineering Technology* 27:1105–08. doi:10.1002/ceat.200402089.

Mediterr. J. Math. (2021) 18:260
https://doi.org/10.1007/s00009-021-01895-2
1660-5446/21/060001-15
published online October 25, 2021
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2021

Mediterranean Journal of Mathematics



On Certain Exact Differential Subordinations Involving Convex Dominants

S. Sivaprasad Kumar and Shagun Banga

Abstract. Let h be a non-vanishing convex univalent function and p be an analytic function in $\mathbb{D}$. We consider the differential subordination $\psi_i(p(z), zp'(z)) \prec h(z)$ with the admissible functions $\psi_1 := (\beta p(z)+\gamma)^{-\alpha}\left(\frac{(\beta p(z)+\gamma)}{\beta(1-\alpha)} + zp'(z)\right)$ and $\psi_2 := \frac{1}{\sqrt{\gamma\beta}} \arctan\left(\sqrt{\frac{\beta}{\gamma}} p^{1-\alpha}(z)\right) + \left(\frac{1-\alpha}{\beta p^{2(1-\alpha)}(z)+\gamma}\right)\frac{zp'(z)}{p^{\alpha}(z)}$. The objective of this paper is to find the dominants, preferably the best dominant (say q) of the solution of the above differential subordination satisfying $\psi_i(q(z), nzq'(z)) = h(z)$. Furthermore, we show that $\psi_i(q(z), zq'(z)) = h(z)$ is an exact differential equation and q is a convex univalent function in $\mathbb{D}$. In addition, we estimate the sharp lower bound of $\operatorname{Re} p$ for different choices of h and derive a univalence criterion for functions in $\mathcal{H}$ (class of analytic normalized functions) as an application to our results.

Mathematics Subject Classification. Primary 30C45; Secondary 30C50.

Keywords. Differential subordination, exact differential equation, convex functions, best dominant.

1. Introduction

Let $\mathcal{H}[a, n]$ denote the class of analytic functions f defined on the open unit disk $\mathbb{D} = \{z : |z| < 1\}$, of the form $f(z) = a + a_n z^n + a_{n+1} z^{n+1} + \cdots$, where n is a positive integer and a,$a_i \in \mathbb{C}$ for all $i \geq n$ with $a_n \neq 0$. Let $\mathcal{H} := \{f \in \mathcal{H}[0, 1] : f'(0) = 1\}$ and $\mathcal{S}$ be the subclass of $\mathcal{H}$ consisting of the univalent functions. The Carathéodory class $\mathcal{P}$ consists of the analytic functions p defined on $\mathbb{D}$ of the form $p(z) = 1 + p_1 z + p_2 z^2 + \cdots$, having positive real part in $\mathbb{D}$. Consider two analytic functions f and g in $\mathbb{D}$, we say $f(z)$ is subordinate to $g(z)$, denoted by $f(z) \prec g(z)$, if $f(z) = g(\omega(z))$,

The work presented here is supported by a Research Fellowship from the Department of Science and Technology, New Delhi.

where ω is a Schwarz function, such that $\omega(0) = 0$ and $|\omega(z)| < 1$. If g is univalent in $\mathbb{D}$, then $f(z) \prec g(z)$ if and only if $f(0) = g(0)$ and $f(\mathbb{D}) \subseteq g(\mathbb{D})$. Ma and Minda [6] unified many subclasses of starlike and convex functions, as follows:

$$\mathcal{S}^*(\phi) = \left\{ f \in \mathcal{S} : \frac{zf'(z)}{f(z)} \prec \phi(z) \right\} \text{ and } \mathcal{C}(\phi) = \left\{ f \in \mathcal{S} : 1 + \frac{zf''(z)}{f'(z)} \prec \phi(z) \right\},$$

where ϕ is an analytic univalent function having positive real part in $\mathbb{D}$ and $\phi'(0) > 0$, such that $\phi(\mathbb{D})$ is symmetric with respect to the real axis and starlike with respect to $\phi(0) = 1$. For $\phi(z) = (1 + z)/(1 - z)$, the above classes, respectively, reduce to $\mathcal{S}^*$, the class of starlike univalent functions and $\mathcal{C}$, the class of convex univalent functions. A function $f \in \mathcal{H}$ is starlike if and only if it satisfies

$$\operatorname{Re}\left(\frac{zf'(z)}{f(z)}\right) > 0, \quad z \in \mathbb{D}.$$

This characterization for starlikeness of functions in $\mathcal{H}$ is also known as Nevanlinna's criterion. For functions in $\mathcal{H}$, the convex characterization [2] and the univalence criterion [10, Lemma 1.1., p.10769] yield the characterization for convex univalent functions in $\mathcal{H}$ as follows:

Conclusion 1.1. The function $f \in \mathcal{H}$ is convex univalent in $\mathbb{D}$ if and only if

$$\operatorname{Re}\left(1 + \frac{zf''(z)}{f'(z)}\right) > 0, \quad z \in \mathbb{D}.$$

Geometrically, it is also known that a function $f(z)$ is convex (starlike) if it maps $\mathbb{D}$ onto a convex (starlike) domain. The characterization for convex functions $f \in \mathcal{H}$ given in Conclusion 1.1 can be extended to convex functions in $\mathcal{H}[a, 1]$. For this, let us consider an analytic function $\tilde{f}(z) = a + a_1 z + a_2 z^2 + \cdots$ $(a_1 \neq 0)$. Now, we assume $f(z) = \frac{\tilde{f}(z)-a}{a_1}$, then $f \in \mathcal{H}$, and we have

$$\operatorname{Re}\left(1 + \frac{z\tilde{f}''(z)}{\tilde{f}'(z)}\right) = \operatorname{Re}\left(1 + \frac{zf''(z)}{f'(z)}\right). \tag{1.1}$$

Using Eq. (1.1), Conclusion 1.1, and the geometric definition of convex functions, we conclude the following:

Conclusion 1.2. The function $\tilde{f}(z) = a + a_1 z + a_2 z^2 + \cdots \in \mathcal{H}[a, 1]$ is convex univalent in $\mathbb{D}$ if and only if it satisfies

$$\operatorname{Re}\left(1 + \frac{z\tilde{f}''(z)}{\tilde{f}'(z)}\right) > 0, \quad z \in \mathbb{D}.$$

The theory of differential subordinations is extensively studied in [8]. For details regarding dominants and best dominants of the differential subordinations, see [8, pp. 16]. For some more work in this direction, one may refer [1,3,5]. The set of analytic and univalent functions $\tilde{q}$ on $\overline{\mathbb{D}}\backslash E(\tilde{q})$, where

$$E(\tilde{q}) = \{\zeta \in \partial\mathbb{D} : \lim_{z\to\zeta} \tilde{q}(z) = \infty\},$$

such that $\tilde{q}'(\zeta) \neq 0$ for $\zeta \in \partial\mathbb{D}\backslash E(\tilde{q})$ is denoted by $\widetilde{Q}$. The theory of admissible functions paved a way in finding the dominants of the solutions of differential subordination altogether with a distinctive approach. The admissible function is defined as follows:

Definition 1.3. [8] Let Ω be a set in $\mathbb{C}$, $\tilde{q} \in \widetilde{Q}$ and n be a positive integer. The class of admissible functions $\Psi_n[\Omega, \tilde{q}]$ consists of the functions $\psi : \mathbb{C}^2 \times \mathbb{D} \to \mathbb{C}$, which satisfy the admissibility condition

$$\psi(r, s; z) \notin \Omega, \tag{1.2}$$

whenever $r = \tilde{q}(\zeta)$, $s = m\zeta\tilde{q}'(\zeta)$, $z \in \mathbb{D}$, $\zeta \in \partial\mathbb{D}\backslash E(\tilde{q})$ and $m \geq n$.

In particular, if Ω is a simply connected domain, $\Omega \neq \mathbb{C}$ and h conformally maps $\mathbb{D}$ onto Ω, we denote the class $\Psi_n[\Omega, \tilde{q}]$ by $\Psi_n[h, \tilde{q}]$. Recently, in [9], authors have studied the applications of admissibility conditions for various known classes of starlike functions. We consider a special case of the class $\Psi_n[\Omega, \tilde{q}]$ by taking $\Omega = \{w : \operatorname{Re} w > 0\}$, $\tilde{q}(\mathbb{D}) = \Omega$, $\tilde{q}(0) = 1$, $E(\tilde{q}) = \{1\}$ and $\tilde{q} \in \widetilde{Q}$, wherein the above admissibility condition given in (1.2) reduces to

$$\psi(\rho i, \sigma; z) \notin \Omega,$$

where ρ and $\sigma \in \mathbb{R}$, $\sigma \leq -(n/2)|1 - i\rho|^2$, $z \in \mathbb{D}$ and $n \geq 1$. Set $\Psi_n[\Omega, \tilde{q}] =: \Psi_n\{1\}$ for the above particular case.

Now, we recall that a first-order differential equation of the type

$$M(x, y)\mathrm{d}x + N(x, y)\mathrm{d}y = 0, \tag{1.3}$$

where $M(x, y)$ and $N(x, y)$ have continuous partial derivatives in some domain $D \subset \mathbb{R}^2$, is called an exact differential equation if it satisfies $\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x}$. The solution of such a differential equation (1.3) is given by

$$\int M(x, y)\mathrm{d}x + \int \tilde{N}(y)\mathrm{d}y = c, \tag{1.4}$$

where c is the integration constant and $\tilde{N}(y)$ is equal to that part of the expression $N(x, y)$, which is independent of x. This concept we carry to the complex plane by considering the following general first-order differential subordination implication:

Implication A. Let $\psi : \mathbb{C}^2 \times \mathbb{D} \to \mathbb{C}$ be an analytic function and h be univalent in $\mathbb{D}$. For some suitable a_0, if $p \in \mathcal{H}[a_0, n]$ satisfies

$$\psi(p(z), zp'(z); z) \prec h(z),$$

then $p \prec q$, where q be the best (a_0, n)-dominant satisfying the differential equation

$$\psi(q(z), nzq'(z)) = h(z).$$

In view of the Implication A, we furnish the following definition:

Definition 1.4. Let ψ, p, and q are as stated in the Implication A. A first-order differential subordination of the form

$$\psi(p(z), zp'(z); z) \prec h(z)$$

is said to be *exact differential subordination* if the following conditions hold:

(i) There exists $\widetilde{M}(z,p)$ and $N(z,p)$ satisfying $\frac{\partial \widetilde{M}(z,p)}{\partial p} = \frac{\partial N(z,p)}{\partial z}$, such that

$$\psi(p(z), zp'(z)) = \widetilde{M}(z,p) + N(z,p)\ p'(z).$$

(ii) When $n = 1$, the equation $\psi(q(z), nzq'(z)) = h(z)$ reduces to an exact differential equation of the form

$$M(z,q)\ \mathrm{d}z + N(z,q)\ \mathrm{d}q = 0,$$

where $M(z,q) := \widetilde{M}(z,q) - h(z)$ and $\frac{\partial M(z,q)}{\partial q} = \frac{\partial N(z,q)}{\partial z}$.

In the present investigation, we mainly focus on the following exact type of differential subordinations:

$$(\beta p(z) + \gamma)^{-\alpha} \left(\frac{(\beta p(z) + \gamma)}{\beta(1-\alpha)} + zp'(z) \right) \prec h(z), \tag{1.5}$$

and

$$\frac{1}{\sqrt{\gamma\beta}} \arctan\left(\sqrt{\frac{\beta}{\gamma}} p^{1-\alpha}(z) \right) + \left(\frac{1-\alpha}{\beta p^{2(1-\alpha)}(z) + \gamma} \right) \frac{zp'(z)}{p^{\alpha}(z)} \prec h(z), \tag{1.6}$$

where h is a non-vanishing convex univalent function in $\mathbb{D}$ with $h(0) = a$ and $p \in \mathcal{H}[a_0, n]$ for an appropriate a_0. We denote the expression on left side of (1.5) and (1.6) by $\psi_1(p(z), zp'(z))$ and $\psi_2(p(z), zp'(z))$, respectively. The speciality of the above two differential subordinations is that they generalize a result studied by Hallenbeck and Ruscheweyh [4]. In the sequel, we find the dominant and the best dominant q, of the solutions of the above differential subordinations, which satisfy

$$\psi_i(q(z), nzq'(z)) = h(z) \quad (i = 1, 2). \tag{1.7}$$

Now, we show that the above two differential subordinations are exact. Take $\widetilde{M}(z,p) := \frac{(\beta p+\gamma)^{1-\alpha}}{\beta(1-\alpha)}$ and $N(z,p) := \frac{z}{(\beta p+\gamma)^{\alpha}}$ in (1.5), then we have $\frac{\partial \widetilde{M}(z,p)}{\partial p} = \frac{\partial N(z,p)}{\partial z}$. Furthermore, for $n = 1$, Eq. (1.7) with $i = 1$, reduces to the following exact differential equation in z and q:

$$M(z,q)\ \mathrm{d}z + N(z,q)\ \mathrm{d}q = 0, \tag{1.8}$$

as $\frac{\partial M(z,q)}{\partial q} = \frac{\partial N(z,q)}{\partial z}$, where $M(z,q) = \frac{(\beta q+\gamma)^{1-\alpha}}{\beta(1-\alpha)} - h(z)$. From Definition 1.4, we get (1.5) is an exact differential subordination. Similarly, we can show that $\psi_2(p(z), zp'(z)) \prec h(z)$ is an exact differential subordination. Since (1.8) is an exact differential equation, we can obtain its solution analogous to (1.4) as follows:

$$M(z,q)\ \mathrm{d}z + \int \tilde{N}(z)\ \mathrm{d}q = c,$$

where $\tilde{N}(z)$ is equal to that part of the expression $N(z,q)$, which is independent of q. Upon simplification of the above equation, we get

$$q(z) = \frac{\left(\frac{\beta(1-\alpha)}{z} \int_0^z h(t)\mathrm{d}t \right)^{\frac{1}{1-\alpha}} - \gamma}{\beta}. \tag{1.9}$$

Interestingly, the above solution q of the exact differential equation (1.8) obtained analogous to (1.4) coincides with the best dominant of the associated differential subordination (1.5), which we illustrate in the subsequent section. Note that the same holds even in the case of differential subordination (1.6). Furthermore, we show that the dominant obtained in each of the cases is convex univalent using the concept of admissibility condition (1.2), which ultimately generalizes and improves some of the earlier known results. We also give some specific examples to justify our claims. In addition, we estimate the best dominant for different choices of h. Also, we find a criterion for univalence of functions in $\mathcal{H}$ as an application to our results. Furthermore, using this univalence criterion, we obtain a few more results. In what follows, let us presume convex to mean convex univalent.

2. Main Results

We need the following Hallenbeck and Ruscheweyh [4] result with $\gamma = 1$ to prove our main results.

Lemma 2.1. [4] *Let h be convex in $\mathbb{D}$ having $h(0) = a$. If $P \in \mathcal{H}[a, n]$ satisfies*

$$P(z) + zP'(z) \prec h(z),$$

then $P(z) \prec Q(z) \prec h(z)$, where

$$Q(z) = \frac{1}{nz^{1/n}} \int_0^z h(t)t^{(1/n)-1}\mathrm{d}t, \tag{2.1}$$

is convex and is the best $(a, n)-$dominant.

The following theorem deals in finding the best dominant of the exact type differential subordination:

Theorem 2.2. *Let $\beta(\neq 0)$, γ be complex numbers, $-1 \leq \alpha \leq 0$ and h be a non-vanishing convex function with $h(0) = a$. Let $a_0 = ((\beta(1-\alpha)a)^{1/(1-\alpha)} - \gamma)/\beta$ and $p \in \mathcal{H}[a_0, n]$ satisfies*

$$\frac{(\beta p(z) + \gamma)^{1-\alpha}}{\beta(1-\alpha)} + \frac{zp'(z)}{(\beta p(z) + \gamma)^{\alpha}} \prec h(z); \tag{2.2}$$

then, $p(z) \prec q(z) \prec H(z)$, where $H(z) := ((\beta(1-\alpha)h(z))^{\frac{1}{1-\alpha}} - \gamma)/\beta$ and

$$q(z) = \frac{\left(\frac{\beta(1-\alpha)}{nz^{1/n}} \int_0^z h(t)t^{(1/n)-1}\mathrm{d}t\right)^{\frac{1}{1-\alpha}} - \gamma}{\beta}. \tag{2.3}$$

Further $H(z)$ and $q(z)$ are convex in $\mathbb{D}$ and q is the best (a_0, n)-dominant.

Proof. We first prove that q and H are convex in $\mathbb{D}$. Since the function h here satisfies the conditions of h of Lemma 2.1, the expression (2.3) becomes

$$q(z) = \frac{(\beta(1-\alpha)Q(z))^{1/(1-\alpha)} - \gamma}{\beta}, \tag{2.4}$$

where Q is given by (2.1). Thus, q is well defined and analytic in $\mathbb{D}$ as $\beta \neq 0$. Let $P(z) = \frac{(\beta p(z)+\gamma)^{1-\alpha}}{\beta(1-\alpha)}$, then $P \in \mathcal{H}[a,n]$ and the subordination (2.2) becomes $P(z)+zP'(z) \prec h(z)$. Now, from Lemma 2.1, we have $P(z) \prec Q(z) \prec h(z)$, which implies $(\beta p(z)+\gamma)^{1-\alpha} \prec \beta(1-\alpha)Q(z) \prec \beta(1-\alpha)h(z)$. Since $h(z)$ does not vanish in $\mathbb{D}$, we have $q'(z) \neq 0$ and therefore $1+zq''(z)/q'(z) =: s(z)$ is analytic in $\mathbb{D}$. Logarithmic differentiation of (2.4) yields

$$1 + \frac{zQ''(z)}{Q'(z)} = 1 + \frac{zq''(z)}{q'(z)} + A(z) =: \tilde{\psi}(s(z); z), \tag{2.5}$$

where

$$A(z) = -\alpha\beta \frac{zq'(z)}{\beta q(z)+\gamma} = \frac{\alpha}{\alpha-1}\frac{zQ'(z)}{Q(z)},$$

and we can write (2.5) as $\tilde{\psi}(r) = r + A(z)$. By Lemma 2.1, the function Q is convex; thus, from Conclusion 1.2, we have $\operatorname{Re}\tilde{\psi}(s(z)) > 0$. Now, to prove q is convex, by Conclusion 1.2, it suffices to show $\operatorname{Re} s(z) > 0$. To accomplish this task, we use [8, Theorem 2.3i, p. 35], and so, we need to show that $\tilde{\psi} \in \Psi_n\{1\}$. Furthermore, using the admissibility condition for the same, it is equivalent to prove that $\operatorname{Re}(\tilde{\psi}(\rho i; z)) \not> 0$. From Eq. (2.5), we obtain $\operatorname{Re}(\tilde{\psi}(\rho i)) = \operatorname{Re}(\rho i + A(z)) = \operatorname{Re} A(z)$. Using the fact that $Q(z) \neq 0$ for all $z \in \mathbb{D}$, we get $zQ'(z)/Q(z)$ is analytic in $\mathbb{D}$ and since $P(z) \prec Q(z)$, we have $Q(0) = a(\neq 0)$, which implies

$$\left.\frac{zQ'(z)}{Q(z)}\right|_{z=0} = 0,$$

and also, we have $Q'(z) \neq 0$ for all z; therefore, $zQ'(z)/Q(z)$ lies on either side of the imaginary axis. So clearly, we have $\operatorname{Re} A(z) \not> 0$. Thus, q is convex in $\mathbb{D}$. Similarly, we can prove the function H is convex in $\mathbb{D}$, since $zh'(z)/h(z)$ $(h(0) = a)$ and $zQ'(z)/Q(z)$ behave alike. We now proceed to show $p(z) \prec q(z) \prec H(z)$ and q is the best (a_0, n)- dominant. Let $\psi_1(r,s) := \frac{(\beta r+\gamma)^{1-\alpha}}{\beta(1-\alpha)} + \frac{s}{(\beta r+\gamma)^{\alpha}}$, which corresponds to left-hand side of the subordination (2.2). First, we show $p \prec H$ by proving $\psi_1 \in \Psi_n[h,H]$, or equivalently

$$\psi := \psi_1(H(\zeta), m\zeta H'(\zeta)) = \frac{(\beta H(\zeta)+\gamma)^{1-\alpha}}{\beta(1-\alpha)} + \frac{m\zeta H'(\zeta)}{(\beta H(\zeta)+\gamma)^{\alpha}} \notin h(\mathbb{D}),$$

whenever $|\zeta| = 1$ and $m \geq n$. From hypothesis, replacing H by its expression in terms of h, in ψ, we obtain

$$\left|\arg\left(\frac{\psi - h(\zeta)}{\zeta h'(\zeta)}\right)\right| = |\arg m| < \pi/2.$$

As the function $h(\mathbb{D})$ is convex and $m \geq 1$, $h(\zeta) \in h(\partial\mathbb{D})$, and $\zeta h'(\zeta)$ is the outer normal to $h(\partial\mathbb{D})$ at $h(\zeta)$, we arrive at the conclusion that $\psi \notin h(\mathbb{D})$, and therefore, we obtain $p(z) \prec H(z)$. Now, we show $p \prec q$. A simple calculation shows that q, given by (2.3), satisfies the following differential equation:

$$\frac{(\beta q(z)+\gamma)^{1-\alpha}}{\beta(1-\alpha)} + \frac{nzq'(z)}{(\beta q(z)+\gamma)^{\alpha}} = \psi_1(q(z), nzq'(z)) = h(z). \tag{2.6}$$

Now, we apply [8, Theorem 2.3f, pp.32] to show q is the best dominant. Without loss of generality, we can assume h and q are analytic and univalent on $\overline{\mathbb{D}}$ and $q'(\zeta) \neq 0$ for $|\zeta| = 1$, which shows that $q \in \widetilde{Q}$. To complete the proof, we now show $\psi_1 \in \Psi_n[h, q]$. This is equivalent to show that

$$\acute{\psi} := \psi_1(q(\zeta), m\zeta q'(\zeta)) = \frac{(\beta q(\zeta) + \gamma)^{1-\alpha}}{\beta(1-\alpha)} + \frac{m\zeta q'(\zeta)}{(\beta q(\zeta) + \gamma)^{\alpha}} \notin h(\mathbb{D}),$$

whenever $|\zeta| = 1$ and $m \geq n$. Using Eqs. (2.3) and (2.6), we obtain

$$Q(\zeta) + \frac{m}{n}\left(h(\zeta) - Q(\zeta)\right) = \acute{\psi}.$$

Since we have $Q(z) \prec h(z)$ from Lemma 2.1. This clearly implies $Q(\zeta) \notin h(\mathbb{D})$ and also we know $m/n \geq 1$, it is easy to observe that $\acute{\psi} \notin h(\mathbb{D})$. This completes the proof. □

Let $n = 1$, and then, the best dominant q, given in (2.3) coincides with the solution (1.9) obtained by solving its associated exact differential equation. By taking $\beta = 1$ and $\gamma = 0$ in Theorem 2.2, we get the following result:

Corollary 2.3. *Let $-1 \leq \alpha \leq 0$, h be a non-vanishing convex function in $\mathbb{D}$ with $h(0) = a$ and $a_0 = ((1-\alpha)a)^{1/(1-\alpha)}$. If $p \in \mathcal{H}[a_0, n]$ satisfies*

$$\frac{p^{1-\alpha}(z)}{1-\alpha} + \frac{zp'(z)}{p^{\alpha}(z)} \prec h(z),$$

then $p(z) \prec q(z) \prec H(z)$, where $H(z) := ((1-\alpha)h(z))^{\frac{1}{1-\alpha}}$ and

$$q(z) = \left(\frac{(1-\alpha)}{nz^{1/n}} \int_0^z h(t)t^{(1/n)-1}\mathrm{d}t\right)^{\frac{1}{1-\alpha}}.$$

Furthermore, $H(z)$ and $q(z)$ are convex in $\mathbb{D}$ and q is the best (a_0, n)-dominant.

In Corollary 2.3, if we choose $\alpha = -1$ and h as any convex function with $h(0) = 1$ and $\operatorname{Re} h(z) > 0$, then we obtain the following result.

Corollary 2.4. *Let h be a convex function with $h(0) = 1$ and $\operatorname{Re} h(z) > 0$. If $p \in \mathcal{H}[1, n]$ satisfies the following:*

$$p^2(z) + 2zp(z)p'(z) \prec h(z),$$

then

$$p(z) \prec q(z) = \sqrt{Q(z)}, \text{ where } Q(z) = \frac{1}{nz^{1/n}} \int_0^z h(t)t^{(1/n)-1}\mathrm{d}t. \tag{2.7}$$

The function q is convex and is the best $(1, n)$-dominant.

Remark 2.5. The Corollary 2.4 improves the already known result of Miller and Mocanu [8, Theorem 3.1e., p. 77], by additionally establishing q, given in (2.7) is convex in $\mathbb{D}$ apart from being the best dominant. Moreover, the Corollary 2.3 generalizes the known result [8, Theorem 3.1e., p. 77] for all α $(-1 \leq \alpha \leq 0)$ and extends the same by proving the convexity of the function q, as well.

For requisite knowledge about the Hypergeometric functions, one may refer [8, p. 5–7]. We now obtain the following corollary for different choices of h in Theorem 2.2, when β is a real number.

Corollary 2.6. *Let $-1 \leq \alpha \leq 0$, $\beta > 0$ and γ be a complex number. If $p \in \mathcal{H}[(\beta(1-\alpha))^{1/(1-\alpha)} - \gamma)/\beta, n]$ and $\psi_1(p(z), zp'(z))$, given in (1.5) satisfies the following:*

(i) $\psi_1 \prec (1+Az)/(1+Bz)$, $-1 \leq B < A \leq 1$, *with* $\lambda_1 := {}_2F_1\left(1, 1/n, 1+1/n; B\right) - \frac{A}{n+1}\ {}_2F_1\left(1, 1+1/n, 2+1/n; B\right)$

(ii) $\psi_1 \prec e^{\mu z}$, $(|\mu| \leq 1)$, *with* $\lambda_2 := {}_1F_1\left(1/n, 1/n+1; -\mu\right)$

(iii) $\psi_1 \prec \sqrt{1+\kappa z}$, $\kappa \in [0,1]$, *with* $\lambda_3 := {}_2F_1\left(-1/2, 1/n, 1+1/n; \kappa\right)$

(iv) $-\frac{\rho_2\pi}{2} < \arg\psi_1 < \frac{\rho_1\pi}{2}$, *where* $\rho = \frac{\rho_1-\rho_2}{\rho_1+\rho_2}$, $\rho' = \frac{\rho_1+\rho_2}{2}$, $0 < \rho_1, \rho_2 \leq 1$, *and* $c = e^{\rho\pi i}$, *with* $\lambda_4 := \sum_{k=0}^{\infty}\left(\frac{\binom{\rho'}{k}(-c)^k\, {}_2F_1(\rho', 1/n+k, 1+1/n+k; -1)}{1+nk}\right)$;

then, for the above parts $(i) - (iv)$, respectively, we have $\mathrm{Re}\, p(z) > \zeta_i(\alpha, \beta, \gamma, \lambda_i)$, where $\zeta_i(\alpha, \beta, \gamma, \lambda_i) = \mathrm{Re}(((\beta(1-\alpha)\lambda_i)^{1/(1-\alpha)} - \gamma)/\beta)$ with i corresponding to its appropriate integral value between 1 and 4. The result is sharp.

Proof. Let $h(z) = (1+Az)/(1+Bz)$, $e^{\mu z}$, $\sqrt{1+\kappa z}$ and $((1+cz)/(1-z))^{\rho'}$, respectively, for the parts $(i) - (iv)$. Then, from Theorem 2.2, we have $\mathrm{Re}\, p(z) > \min_{|z|\leq 1} \mathrm{Re}\, q(z)$, where q is given by (2.4). Therefore, it suffices to find the minimum value of real part of Q in each of the parts, where Q is given by (2.1).

(i) We have $h(z) = (1+Az)/(1+Bz)$, then, from Eq. (2.1), we get

$$\begin{aligned} Q(z) &= \frac{1}{nz^{1/n}} \int_0^z \left(\frac{1+At}{1+Bt}\right) t^{(1/n)-1} \mathrm{d}t \\ &= \frac{1}{n} \int_0^1 (1+Btz)^{-1} t^{(1/n)-1} \mathrm{d}t + \frac{Az}{n} \int_0^1 (1+Btz)^{-1} t^{1/n} \mathrm{d}t \\ &= {}_2F_1\left(1, 1/n, 1+1/n; -Bz\right) + \frac{Az}{n+1}\ {}_2F_1\left(1, 1+1/n, 2+1/n; -Bz\right). \end{aligned}$$

Now, we have $\min_{|z|\leq 1} \mathrm{Re}\, Q(z) = Q(-1) = \lambda_1$, where

$$\lambda_1 := {}_2F_1\left(1, 1/n, 1+1/n; B\right) - \frac{A}{n+1}\ {}_2F_1\left(1, 1+1/n, 2+1/n; B\right).$$

Hence the result.

(ii) We have $h(z) = \exp(\mu z)$, then from Eq. (2.1) we obtain

$$\begin{aligned} Q(z) &= \frac{1}{nz^{1/n}} \int_0^z e^{\mu t} t^{1/n-1} \mathrm{d}t \\ &= \tfrac{1}{n} \int_0^1 e^{\mu t z} t^{1/n-1} \mathrm{d}t \\ &= {}_1F_1\left(1/n, 1/n+1; \mu z\right). \end{aligned}$$

Clearly, $\min_{|z|\leq 1} \mathrm{Re}\, Q(z) = Q(-1) = \lambda_2$, where

$$\lambda_2 := {}_1F_1\left(1/n, 1/n+1; -\mu\right).$$

This completes the proof for part (ii).

(iii) Let $h(z) = \sqrt{1+\kappa z}$, then from Eq. (2.1), we have

$$\begin{aligned} Q(z) &= \frac{1}{nz^{1/n}} \int_0^z \sqrt{1+\kappa t}\; t^{(1/n)-1} \mathrm{d}t \\ &= \frac{1}{n} \int_0^1 \sqrt{1+\kappa t z}\; t^{(1/n)-1} \mathrm{d}t \\ &= {}_2F_1\left(-1/2, 1/n, 1+1/n; -\kappa z\right). \end{aligned}$$

Thus, we have $\min_{|z|\le 1} \operatorname{Re} Q(z) = Q(-1) = \lambda_3$, where $\lambda_3 := {}_2F_1\left(-1/2, 1/n, 1+1/n; \kappa\right)$ and that completes the proof for part (iii).

(iv) We have $h(z) = \left(\frac{1+cz}{1-z}\right)^{\rho'}$, then from Eq. (2.1), we have

$$\begin{aligned} Q(z) &= \frac{1}{nz^{1/n}} \int_0^z \left(\frac{1+ct}{1-t}\right)^{\rho'} t^{(1/n)-1} \mathrm{d}t \\ &= \frac{1}{nz^{1/n}} \int_0^z \left(\sum_{k=0}^{\infty} \binom{\rho'}{k} c^k t^{(1/n+k-1)} (1-t)^{-\rho'} \right) \mathrm{d}t \\ &= \frac{1}{nz^{1/n}} \sum_{k=0}^{\infty} \left(\binom{\rho'}{k} c^k \int_0^1 (tz)^{(1/n+k-1)} (1-tz)^{-\rho'} z \mathrm{d}t \right) \\ &= \frac{1}{n} \sum_{k=0}^{\infty} \left(\binom{\rho'}{k} (cz)^k \int_0^1 t^{(1/n+k-1)} (1-tz)^{-\rho'} \mathrm{d}t \right) \\ &= \sum_{k=0}^{\infty} \left(\frac{\binom{\rho'}{k} (cz)^k {}_2F_1(\rho', 1/n+k, 1+1/n+k; z)}{1+nk} \right). \end{aligned}$$

Now, $\min_{|z|\le 1} \operatorname{Re} Q(z) = Q(-1) = \lambda_4$, where

$$\lambda_4 := \sum\nolimits_{k=0}^{\infty} \left(\frac{\binom{\rho'}{k} (-c)^k {}_2F_1(\rho', 1/n+k, 1+1/n+k; -1)}{1+nk} \right).$$

Hence the result. □

Remark 2.7. If A and B are replaced with each other in Corollary 2.6(i), then we have $\min_{|z|\le 1} \operatorname{Re} Q(z) = Q(1)$. Furthermore, by taking $A = 2a - 1$ $(a \in [0,1))$, $B = 1$, $\alpha = -1$, $\beta = 1$, and $\gamma = 0$, the above result reduces to the result of Miller and Mocanu [8, Corollary 3.1e.1, p. 79].

In the following theorem, we deal with another exact type differential subordination.

Theorem 2.8. *Let β, $\gamma(\neq 0)$ be complex numbers, $-1 \le \alpha \le 0$, h be a non-vanishing convex function in $\mathbb{D}$ with $h(0) = a$, $|\sqrt{\gamma\beta} h(z)| < \pi/2$ and $a_0 = (\sqrt{\gamma/\beta}\tan(\sqrt{\gamma\beta}a))^{1/(1-\alpha)}$. If $p \in \mathcal{H}[a_0, n]$ satisfies*

$$\frac{1}{\sqrt{\gamma\beta}} \arctan\left(\sqrt{\frac{\beta}{\gamma}} p^{1-\alpha}(z)\right) + \left(\frac{1-\alpha}{\beta p^{2(1-\alpha)}(z) + \gamma}\right) \frac{z p'(z)}{p^{\alpha}(z)} \prec h(z), \qquad (2.8)$$

then $p(z) \prec q(z) \prec H(z)$, *where* $H(z) = (\sqrt{\gamma/\beta}\tan(\sqrt{\gamma\beta}h(z)))^{1/(1-\alpha)}$ *and*

$$q(z) = \left(\sqrt{\frac{\gamma}{\beta}}\tan\left(\frac{\sqrt{\gamma\beta}}{nz^{1/n}}\int_0^z h(t)t^{(1/n)-1}\mathrm{d}t\right)\right)^{\frac{1}{1-\alpha}}. \tag{2.9}$$

Furthermore, $H(z)$ *and* $q(z)$ *are convex in* $\mathbb{D}$ *and* q *is the best* $(a_0, n)-$*dominant.*

Proof. We first show that q and H are convex in $\mathbb{D}$. As the function h here satisfies the condition of h of Lemma 2.1, thus q reduces to

$$q(z) = \left(\sqrt{\frac{\gamma}{\beta}}\tan\left(\sqrt{\gamma\beta}Q(z)\right)\right)^{\frac{1}{1-\alpha}}, \tag{2.10}$$

where Q is given by (2.1) and the function q is well defined and analytic in $\mathbb{D}$. Let $P(z) = (1/\sqrt{\gamma\beta})\arctan\left(\sqrt{\frac{\beta}{\gamma}}p^{1-\alpha}(z)\right)$, then the subordination (2.8) reduces to $P(z) + zP'(z) \prec h(z)$ and clearly $P \in \mathcal{H}[a, n]$. Now, from Lemma 2.1, we have $P(z) \prec Q(z) \prec h(z)$, which is equivalent to $\arctan\left(\sqrt{\frac{\beta}{\gamma}}p^{1-\alpha}(z)\right) \prec \sqrt{\gamma\beta}Q(z) \prec \sqrt{\gamma\beta}h(z)$. We have $q'(z) \neq 0$ as $h(z) \neq 0$ for all $z \in \mathbb{D}$, and therefore, $1 + zq''(z)/q'(z) =: g(z)$ is analytic in $\mathbb{D}$. The logarithmic differentiation of the function q, given in (2.10), yields

$$1 + \frac{zQ''(z)}{Q'(z)} = 1 + \frac{zq''(z)}{q'(z)} + B(z) =: \tilde{\psi}(g(z); z), \tag{2.11}$$

where

$$\begin{aligned} B(z) &= -\left(\alpha + 2(1-\alpha)\beta\frac{q^{2(1-\alpha)}(z)}{\gamma + \beta q^{2(1-\alpha)}(z)}\right)\left(\frac{zq'(z)}{q(z)}\right) \\ &= \frac{\alpha}{\alpha - 1}\left(\sum_{k=0}^{\infty}(-1)^k\frac{(2\sqrt{\gamma\beta}Q(z))^{2k}}{(2k+1)!}\right)^{-1}\left(\frac{zQ'(z)}{Q(z)}\right) \\ &\quad -2\sqrt{\gamma\beta}\tan(\sqrt{\gamma\beta}Q(z))zQ'(z), \end{aligned} \tag{2.12}$$

and we can write $\tilde{\psi}(r) = r + B(z)$. By Lemma 2.1, the function Q is convex, thus, from Conclusion 1.2, we have $\operatorname{Re}\tilde{\psi}(g(z)) > 0$. To prove q is convex, by Conclusion 1.2, it suffices to show $\operatorname{Re} g(z) > 0$. Now, to achieve this, we use [8, Theorem 2.3i, p. 35], so then we only need to show $\tilde{\psi} \in \Psi_n\{1\}$. Furthermore, using the admissibility condition for the same, it is equivalent to show that $\operatorname{Re}(\tilde{\psi}(\rho i; z)) \not\geq 0$. From Eq. (2.11), we obtain

$$\operatorname{Re}(\tilde{\psi}(\rho i)) = \operatorname{Re}(\rho i + B(z)) = \operatorname{Re} B(z).$$

Now, using the fact $Q(z) \neq 0$ for all $z \in \mathbb{D}$, we get B is analytic in $\mathbb{D}$. Since $P(z) \prec Q(z)$, we have $Q(z) \in \mathcal{H}[a, n]$, thus

$$\left(\sum_{k=0}^{\infty}(-1)^k\frac{(2\sqrt{\gamma\beta}Q(z))^{2k}}{(2k+1)!}\right)^{-1}\left(\frac{zQ'(z)}{Q(z)}\right)\Bigg|_{z=0} = 0$$

and

$$\tan(\sqrt{\gamma\beta}Q(z))zQ'(z)|_{z=0} = 0,$$

which ensures either $B(z) \equiv 0$ or lies on either side of the imaginary axis. Therefore, from Eq. (2.12), we have $\operatorname{Re} B(z) \not> 0$, which proves q is convex in $\mathbb{D}$. Now, similarly, we can prove that the function H is convex in $\mathbb{D}$. We now show $p(z) \prec q(z) \prec H(z)$ and q is the best (a_0, n)-dominant. Now, for this, we assume

$$\psi_2(r,s) := \frac{1}{\sqrt{\gamma\beta}} \arctan\left(\sqrt{\frac{\beta}{\gamma}} r^{1-\alpha}\right) + \frac{(1-\alpha)s}{r^{\alpha}(\beta r^{2(1-\alpha)} + \gamma)},$$

which corresponds to left-hand side of the subordination (2.8). Now, first, we show $p \prec H$ by proving $\psi_2 \in \Psi_n[h, H]$, or equivalently

$$\begin{aligned}&\frac{1}{\sqrt{\gamma\beta}} \arctan\left(\sqrt{\frac{\beta}{\gamma}} H^{1-\alpha}(\zeta)\right) + \frac{(1-\alpha)m\zeta H'(\zeta)}{H^{\alpha}(\zeta)(\beta H^{2(1-\alpha)}(\zeta) + \gamma)} \\ &\quad = \psi_2(H(\zeta), m\zeta H'(\zeta)) =: \psi \notin h(\mathbb{D}),\end{aligned}$$

whenever $|\zeta| = 1$ and $m \geq n$. Now, replacing H by its expression in terms of h in ψ, we get

$$\left|\arctan\left(\frac{\psi - h(\zeta)}{\zeta h'(\zeta)}\right)\right| = |\arg m| < \frac{\pi}{2}.$$

Since the function h is convex, $h(\zeta) \in h(\partial\mathbb{D})$ and $\zeta h'(\zeta)$ is the outer normal to $h(\partial\mathbb{D})$ at $h(\zeta)$, thus we obtain $\psi \notin h(\mathbb{D})$, which further implies $p(z) \prec H(z)$. A computation shows that q, given by (2.9), satisfies the differential equation

$$\begin{aligned}h(z) \;&=\; \frac{1}{\sqrt{\gamma\beta}} \arctan\left(\sqrt{\frac{\beta}{\gamma}} q^{1-\alpha}(z)\right) + \frac{(1-\alpha)nzq'(z)}{q^{\alpha}(z)(\beta q^{2(1-\alpha)}(z) + \gamma)} \\ &=: \psi_2(q(z), nzq'(z)).\end{aligned} \tag{2.13}$$

Now, we apply [8, Theorem 2.3f, p. 32] to show q is the best dominant. Without loss of generality, we can assume that h and q are analytic and univalent on $\overline{\mathbb{D}}$ and $q'(\zeta) \neq 0$ for $|\zeta| = 1$, which shows that $q \in \widetilde{Q}$. To complete the proof, it suffices to show $\psi_2 \in \Psi_n[h, q]$, which is equivalent to show that

$$\begin{aligned}&\frac{1}{\sqrt{\gamma\beta}} \arctan\left(\sqrt{\frac{\beta}{\gamma}} q^{1-\alpha}(\zeta)\right) + \frac{(1-\alpha)m\zeta q'(\zeta)}{q^{\alpha}(\zeta)(\beta q^{2(1-\alpha)}(\zeta) + \gamma)} \\ &\quad = \psi_2(q(\zeta), m\zeta q'(\zeta)) =: \acute{\psi} \notin h(\mathbb{D}),\end{aligned}$$

whenever $|\zeta| = 1$ and $m \geq n$. Equations (2.1) and (2.13) yield

$$Q(\zeta) + \frac{m}{n}(h(\zeta) - Q(\zeta)) = \acute{\psi}.$$

As we have $m/n \geq 1$ and from Lemma 2.1, we have $Q(z) \prec h(z)$; therefore, evidently, we get $\acute{\psi} \notin h(\mathbb{D})$. This completes the proof. □

Taking $n = 1$ in Theorem 2.8, Eq. (2.13) reduces to the differential equation (1.8) with $M(z,q) := (1/\sqrt{\gamma\beta}) \arctan\left(\sqrt{\frac{\beta}{\gamma}} q^{1-\alpha}\right) - h(z)$ and $N(z,q) := \frac{(1-\alpha)z}{q^{\alpha}(\beta q^{2(1-\alpha)}+\gamma)}$. Furthermore, we have $\frac{\partial M(z,q)}{\partial q} = \frac{\partial N(z,q)}{\partial z}$, which

shows that the differential equation, given in (2.13) is exact. Consequently, Eq. (1.4) yields the solution for equation (2.13) whenever $n = 1$ as follows:

$$(1/\sqrt{\gamma\beta})z \arctan\left(\sqrt{\frac{\beta}{\gamma}}q^{1-\alpha}(z)\right) - \int_0^z h(t)\mathrm{d}t = 0,$$

which coincides with (2.9) with $n = 1$. We now obtain the following corollary for different choices of h in Theorem 2.8, when β is a real number.

Corollary 2.9. *Let $-1 \leq \alpha \leq 0$, $\beta > 0$ and $\gamma(\neq 0)$ be a complex number. If $p \in \mathcal{H}[(\sqrt{\gamma/\beta}\tan(\sqrt{\gamma\beta}a))^{1/(1-\alpha)}, n]$ and $\psi_2(p(z), zp'(z))$, given by (1.6) satisfies the following:*

(i) *$\psi_2 \prec (1 + Az)/(1 + Bz)$, where $-1 \leq B < A \leq 1$ such that $|\sqrt{\gamma\beta}(1 + Az)/(1 + Bz)| < \pi/2$*
(ii) *$\psi_2 \prec e^{\mu z}$, where $|\mu| \leq 1$ such that $|\sqrt{\gamma\beta}e^{\mu z}| < \pi/2$*
(iii) *$\psi_2 \prec \sqrt{1 + \kappa z}$, where $\kappa \in [0, 1]$ such that $|\sqrt{\gamma\beta}\sqrt{1 + \kappa z}| \leq \pi/2$*
(iv) *$-\rho_2\pi/2 < \arg\psi_2 < \rho_1\pi/2$, $\rho = \frac{\rho_1-\rho_2}{\rho_1+\rho_2}$, where $0 < \rho_1, \rho_2 \leq 1$, and $c = e^{\rho\pi i}$ such that $|\sqrt{\gamma\beta}((1 + cz)/(1 - z))^{\rho'}| < \pi/2$, $(\rho' = (\rho_1 + \rho_2)/2)$,*

then, respectively, for the above parts (i) − (iv), we have $\mathrm{Re}\, p(z) > \xi_i(\alpha, \beta, \gamma, \lambda_i)(i = 1, 2, 3, 4)$, where

$$\xi_i(\alpha, \beta, \gamma, \lambda_i) = \mathrm{Re}\left(\sqrt{\gamma/\beta}\tan\left(\sqrt{\gamma\beta}\lambda_i\right)\right)^{\frac{1}{1-\alpha}},$$

and λ_i is same as in the Corollary 2.6. The result is sharp.

Proof. The proof of each part of this Corollary is on the similar lines of the proof of the Corollary 2.6. Therefore, omitted. □

3. Specific Examples

We provide here below some illustrations to our results.

Example 1. If $p \in \mathcal{H}[1, 1]$ satisfies

$$p(z) + zp'(z) \prec \frac{1 + z}{1 - z}, \tag{3.1}$$

then $p(z) \prec q(z) := \frac{-2\log(1-z)}{z} - 1 \prec \frac{1+z}{1-z}$. Moreover, q is the best $(1, 1)$-dominant and is convex in $\mathbb{D}$.

Proof. By taking $h(z) = \frac{1+z}{1-z}$, $n = 1$, and $\alpha = 0$ in Corollary 2.3, we obtain $Q(z) = q(z) = -2(\log(1 - z))/z - 1$. Now, the assertion follows at once from Corollary 2.3 as $h(0) = 1$ and $h(z) \neq 0$ for all $z \in \mathbb{D}$. Moreover, the quantities $zh'(z)/h(z)$ lie on either side of the imaginary axis as proved in the proof of Theorem 2.2. □

Remark 3.1. The above result can also be stated as if $p \in \mathcal{H}[1, 1]$ satisfies

$$\mathrm{Re}(p(z) + zp'(z)) > 0,$$

then

$$\mathrm{Re}\, p(z) > 2\log 2 - 1$$

$$= {}_2F_1\,(1,1,2;-1) - \frac{1}{2}\,{}_2F_1\,(1,2,3;-1)\,,$$

which is also a special case of Corollary 2.6(i) with $A = 1$, $B = -1$, $\beta = 1$, $\gamma = 0$, and $\alpha = 0$.

Example 2. Let $a_0 = 2((2/3)^{3/4} - 1)$. If $p \in \mathcal{H}[a_0, 1]$ satisfies the differential subordination

$$\frac{3(p(z)/2+1)^{4/3}}{2} + zp'(z)(p(z)/2+1)^{1/3} \prec \exp(z),$$

then $p(z) \prec q(z) \prec H(z)$, where $H(z) = 2((2e^z/3)^{3/4} - 1)$ and $q(z) = 2((2(e^z-1)/(3z))^{3/4} - 1)$. Moreover, the function q is convex and is the best $(a_0, 1)$-dominant.

Proof. Let $h(z) = \exp(z)$, $n = 1$, $\alpha = -1/3$, $\beta = 1/2$, and $\gamma = 1$ in Theorem 2.2, then we obtain $Q(z) = (e^z - 1)/z$. Now, h here with the given constants satisfy the hypothesis of Theorem 2.2, the assertion follows from the same. Moreover, in accordance with the proof of Theorem 2.2, we observe that

$$\operatorname{Re}\frac{zQ'(z)}{Q(z)} = \operatorname{Re}\left(\frac{1 + e^z(z-1)}{e^z - 1}\right) \not> 0$$

and similarly the quantity $\operatorname{Re} zh'(z)/h(z) = \operatorname{Re} z \not> 0$, therefore, both the quantities lie on either side of the imaginary axis. □

Example 3. Let $a_0 = ((0.5)\tan(0.5))^{3/5}$. If $p \in \mathcal{H}[a_0, 1]$ satisfies

$$2\arctan\left(2p^{5/3}(z)\right) + \frac{20}{3}\left(\frac{zp^{2/3}(z)p'(z)}{1 + 4p^{10/3}(z)}\right) \prec \frac{2+z}{2-z},$$

then

$$p(z) \prec q(z) := \left(\tan\left(\frac{-1}{2} - \frac{2}{z}\log\left(\frac{2-z}{2}\right)\right)\Big/2\right)^{3/5}. \tag{3.2}$$

Moreover, the function q is the best $(a_0, 1)$-dominant and is convex in $\mathbb{D}$.

Proof. Let $h(z) = \frac{2+z}{2-z}$, $n = 1$, $\alpha = -2/3$, $\beta = 1$, and $\gamma = 1/4$ in Theorem 2.8, then we obtain $Q(z) = -1 - ((4/z)\log((2-z)/z))$ and $q(z)$ is given by (3.2). As the function h and all the constants satisfy the hypothesis of Theorem 2.8, the result follows at once. Moreover, we observe that the quantity in (2.12) for the present case also lies on either side of the imaginary axis, which is in accordance with the proof of Theorem 2.8 □

3.1. Applications to Univalent Function

We obtain a sufficient condition for univalence of $f \in \mathcal{H}$ in the following theorem.

Theorem 3.2. *Let $f \in \mathcal{H}$. If f satisfies*

$$\operatorname{Re}(f'(z) + zf''(z)) > 0, \tag{3.3}$$

then f is univalent in $\mathbb{D}$ and in fact, $\operatorname{Re} f'(z) > 2\log 2 - 1$.

Proof. Let $p(z) = f'(z)$; then, we have $p \in \mathcal{H}[1,1]$. We observe that Eq. (3.3) becomes equivalent to equation (3.1) for $p(z) = f'(z)$. Since p satisfies the hypothesis of Example 1, we obtain $\operatorname{Re} f'(z) > 2\log 2 - 1 > 0$. Thus, Noshiro–Warschawski result yields the function f is univalent. □

The result by Miller Mocanu [7] with $n = 1$ is as follows:

Corollary 3.3. [7] *Let $a \in [0,1)$ and $\eta = \eta(a)$ be defined as*

$$\eta = (2(1-a)\beta(1) + (2a-1))^{1/2},$$

where $\beta(x) = \int_0^1 \frac{t^{x-1}}{1+t}\mathrm{d}t$. If $f \in \mathcal{H}$, then

$$\operatorname{Re} f'(z) > a \Rightarrow \operatorname{Re}\sqrt{\frac{f(z)}{z}} > \eta(a).$$

Corollary 3.4. *Let $f \in \mathcal{H}$ and $a = 2\log 2 - 1$. If f satisfies the Eq. (3.3), then*

(i) $\operatorname{Re}\sqrt{\frac{f(z)}{z}} > \eta(a)$.

(ii) $\operatorname{Re}\dfrac{f(z)}{z} > a$.

Proof. (i) The proof follows directly from Theorem 3.2 and Corollary 3.3.

(ii) Let $p(z) = f(z)/z$; then, we have $p \in \mathcal{H}[1,1]$. Thus, Remark 3.1 yields $\operatorname{Re} f'(z) > 0 \Rightarrow \operatorname{Re}(f(z)/z) > a$. Thus, the result follows now using Theorem 3.2. □

Acknowledgements

The authors are thankful to the editor and to the referees for their valuable and insightful suggestions.

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

References

[1] Ali, R.M., Cho, N.E., Ravichandran, V., Kumar, S.S.: Differential subordination for functions associated with the lemniscate of bernoulli. Taiwan. J. Math. **16**(3), 1017–1026 (2012)

[2] Ali, R.M., Ravichandran, V., Lee, S.K.: Subclasses of multivalent starlike and convex functions. Bull. Belg. Math. Soc. Simon Stevin **16**(3), 385–394 (2009)

[3] Banga, S., Kumar, S.S.: Applications of differential subordinations to certain classes of starlike functions. J. Korean Math. Soc. **57**(2), 331–357 (2020)

[4] Hallenbeck, D.J., Ruscheweyh, S.: Subordination by convex functions. Proc. Am. Math. Soc. **52**, 191–195 (1975)

[5] Kumar, S., Ravichandran, V.: Subordinations for functions with positive real part. Complex Anal. Oper. Theory **12**(5), 1179–1191 (2018)

[6] Ma, W.C., Minda, D.: A unified treatment of some special classes of functions. In: Proceedings of the Conference on Complex Analysis, pp. 157–169. Conference Proceedings and Lecture Notes in Analysis. I, International Press, Cambridge, MA, Tianjin (1992)

[7] Miller, S.S., Mocanu, P.T.: Differential subordinations and univalent functions. Mich. Math. J. **28**(2), 157–172 (1981)

[8] Miller, S.S., Mocanu, P.T.: Differential Subordinations: Theory and Applications, Pure and Applied Mathematics, vol. 225. Marcel Dekker, New York (2000)

[9] Naz, A., Nagpal, S., Ravichandran, V.: Star-likeness associated with the exponential function. Turk. J. Math. **43**(3), 1353–1371 (2019)

[10] Nunokawa, M., Sokół, J.: An improvement of Ozaki's condition. Appl. Math. Comput. **219**(22), 10768–10776 (2013)

S. Sivaprasad Kumar and Shagun Banga
Department of Applied Mathematics
Delhi Technological University
Delhi 110042
India
e-mail: shagun05banga@gmail.com

S. Sivaprasad Kumar
e-mail: spkumar@dce.ac.in

Received: July 3, 2020.

Revised: February 3, 2021.

Accepted: October 3, 2021.

Online Robust MPC based Emergency Maneuvering System for Autonomous Vehicles

Vivek Bithar *, Punit Tulpule *, Shawn Midlam-Mohler *
*Mechanical and Aerospace Engineering Department
The Ohio State University, Columbus, OH, USA

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

arXiv:2109.11959v2 [eess.SY] 27 Sep 2021

Abstract

Nonlinear Robust Model Predictive Control (RMPC) provides a very promising solution to the problem of automatic emergency maneuvering, which is capable of handling multiple possibly conflicting objectives of robustness and performance. Even though RMPC gives a suboptimal solution, the key challenge in real-time implementation is that it is computationally very demanding. In this paper a real-time capable robust tube MPC based framework for steering control during emergency obstacle avoidance maneuver is presented. The novelty of this framework lies in the robust integration of path planning and path following tasks of autonomous vehicles. A simulation study showcases the robust performance improvements due to the proposed strategy over a non-robust MPC in different extreme maneuvering scenarios.

Index Terms

Robust MPC, Collision Avoidance, Autonomous Vehicle, Motion Planning, Vehicle Control

.

I. Introduction

THE obstacle avoidance system for Autonomous Vehicles (AVs) deals with planning of emergency maneuvers to avoid the impending crash with the obstacle in the path [1] in addition to tracking the planned path and the stabilization [2] [3]. Early attempts at design of path planning and tracking focused on the hierarchical MPC framework that separated the task of path planning and tracking [4] [5] [6]. Authors in [7] presented a controller design, where high-level controller used a simplified point-mass vehicle model to generate a reference path based on information about the current position of the obstacle using NLMPC and fed the path to a low -level controller introduced in [8]. In [4], the authors used similar hierarchical structure for the trajectory planning and tracking tasks by applying an explicit MPC control law with 4-wheel model for the low level controller for high prediction accuracy. Another variant of this hierarchical approach is presented in [9] with MPC based motion planning and a PID based low level controller. Due to the use of simplified point mass models for motion planning, all these approaches reported dynamic infeasibility of the path planned at operating conditions such as high longitudinal velocity and for low road friction coefficient surfaces, leading to large tracking errors because of low-level controller prioritizing vehicle stabilization over tracking. Thus, limiting the range of operating conditions and functionality in emergency scenarios, for such frameworks.

To address this issue, the approaches that followed focused on utilizing a single MPC based framework to combine the tasks of the stabilization, trajectory generation/re-planning and tracking. Such a control system was proposed in [10] as a shared MPC controller for ground vehicles. The controller combined stable handling envelopes from [11] and environmental envelopes that representedd the obstacle and road edge constraints. This collision avoidance system computed obstacle free regions of the environments by joining feasible gaps in the prediction horizon as tubes and as a result transformed the non-convex optimization problem into a set of convex problems. The authors further employed a variable time step MPC execution to reduce the computational load, and to increase the look ahead time/horizon. Stability and collision avoidance objectives typically pose conflicting requirements on MPC design due to limited computational resources on board a vehicle. For stabilization, short time steps for capturing fast dynamics is recommended whereas collision avoidance requires preferably sufficient look ahead distance to react early. The controller developed in [10] shared a controller between the driver and the motion planning system, where the system was planned to intervene whenever it detected that the intended maneuver may result in the vehicle going outside the stability envelope.

The system design in [12] extended the framework developed by [10] and applied it to AV. The research pointed out that the controller for an AV has the advantage of having precise knowledge of the desired path instead of simply a projection of driver intent from the steering angle history as assumed in [10]. The authors worked to improve the vehicle modeling to take advantage of the availability of the information about upcoming path and its curvature. They employed successive linearization scheme for rear tire force generation and first order hold for linearization in the longer time step region of the controller. It continued using the safe handling and obstacle avoidance envelopes introduced in [11]. The research presented in [13] further

developed initial approach in [10] and [12] by introducing a point mass longitudinal controller to the system. The model linearization was improved with refined rear tire force linearization using an average rear tire slip angle.

The approaches discussed so far did not consider the presence of uncertainties while designing the control system. The measurement errors from the Inertial Measurement Units (IMU), model parametric uncertainties such as friction coefficients, tire models and linearized dynamic model inaccuracies are present in the real environment. To be functional in such real situations, the controller is required to be robust against them. The design of collision avoidance system thus, become a motion planning problem in Uncertain and Dynamic Environment (UDE). Mobile robots subject to nonholonomic constraints in a UDE have been studied extensively in the field of the robotics, and AVs can be considered as extension of such systems.

[14] introduced a set-theoretic receding horizon control algorithm for robots modelled as Linear Time-Invariant (LTI) systems subject to input and state constraints and disturbances. This algorithm computes ellipsoidal approximations of exact one-step controllable sets for the LTI systems considering all the possible obstacle scenarios that the system can encounter. Then it exploited these sets online to determine specific control actions to be applied on the system to have a collision free trajectory in an oriented graph space. The resulting framework guaranteed collision free and bounded feasible trajectory regardless of any obstacle scenario occurrence. Offline computation of the sets assumed that obstacle locations on the working area are known, but the current obstacle configuration to be unpredictable at each time instant. Due to this restrictive assumption, the pre-computed sets were comprehensive in nature to provide uniform ultimate boundedness and feasible solution at each time instant. This publication showed an offline methodology that brought obstacle configuration uncertainty in addition to system state measurement errors and disturbances in the system design.

Work presented in [15] also explored the computation of disturbance invariant set and proposed a bi-level NLMPC optimization strategy. Disturbance-invariant are the regions in the state-space of a system that guarantee that if the system starts in the invariant set, then the system will lie in the invariant set under bounded disturbances. The strategy introduced first executed a feed-forward control optimization using approximate disturbance-invariant sets for a general nonlinear dynamic system at an instant and next performed an optimization to propagate the ellipsoid for the updated controller and nominal trajectory for the next instant of time. It showed the framework was able to compute and execute a trajectory for a quadrotor system with a nonlinear backstepping controller in a space where the obstacles were stationary. This paper, in addition to attempting online robust MPC algorithm also attempted to provide a solution for a general nonlinear system. The work did not provide any computational performance information of the controller system for the experiment performed.

In literature, the uncertainty in the MPC framework has also been treated with Tube Based MPC Scheme (TMPC), for example, [16] used this structure for a general Lipschitz nonlinear systems to show that invariant sets for the system can be computed offline, and does not require recomputing the sets for every nominal trajectory provided The invariant set can be specified as an invariant space around the nominal trajectory. [16] used offline computed control invariant sets and demonstrated obstacle avoidance by combining tasks of planning and lane keeping on straight icy roads with stability bounds designed to limit vehicle states to linear regime of tire force generation. It developed a non-linear force input Lipschitz bicycle model with the assumption of additive uncertainty and simplified rear and front tire force linear model. For robust obstacle avoidance, it projected the computed control invariant set on the obstacle position ellipsoid constraint and solved nominal MPC with updated obstacle constraints. Although the controller execution rate was slow (approx. 100 [ms]) compared to the dynamics involved due to NLMPC optimization, this did present a functional approach of applying robust MPC for developing obstacle and path tracking system of an AV.

Existing researches have not presented a real-time implementable control system that can arbitrate between objectives of obstacle avoidance, stabilization and path tracking, and be robust against disturbances for emergency maneuvering. In this paper we propose a robust obstacle avoidance control which integrates the tasks of path planning and path tracking. The proposed approach improves the controller performance in terms of tracking error and vehicle stability in presence of uncertainties. While ensuring the robustness, this approach addresses the controls' path infeasibility problem. TMPC is selected as an obstacle avoidance and control system design methodology due to its efficient multiple-objective and explicit robust constraint handling capabilities. In the absence of obstacles in the perception region of the vehicle, the admissible region is restricted to a lane width around the prescribed reference path and the problem transforms to a reference path tracking. The objective of path tracking is conflicting with the obstacle avoidance objective and the implemented novel robust MPC (RMPC) system described in this study provides an effective robust solution for both problem with a single system design. The approach is employed in a simulation study to demonstrate improvements over a deterministic MPC and real-time computation capabilities. We used CarSIM and Simulink® co-simulation environment to test the proposed approach on three different extreme maneuvering scenarios.

The structure of the paper is as follows. Section II introduces the mathematical models used in the proposed control system. Section III presents the TMPC based robust controller and details the design process. Section IV elaborates the simulation setup, estimation of the system disturbance, evaluates and compares the performance of the designed robust control system with a deterministic non-robust control system in different test cases. Section V provides a conclusion and a discussion on the potential future improvements of the designed control system.

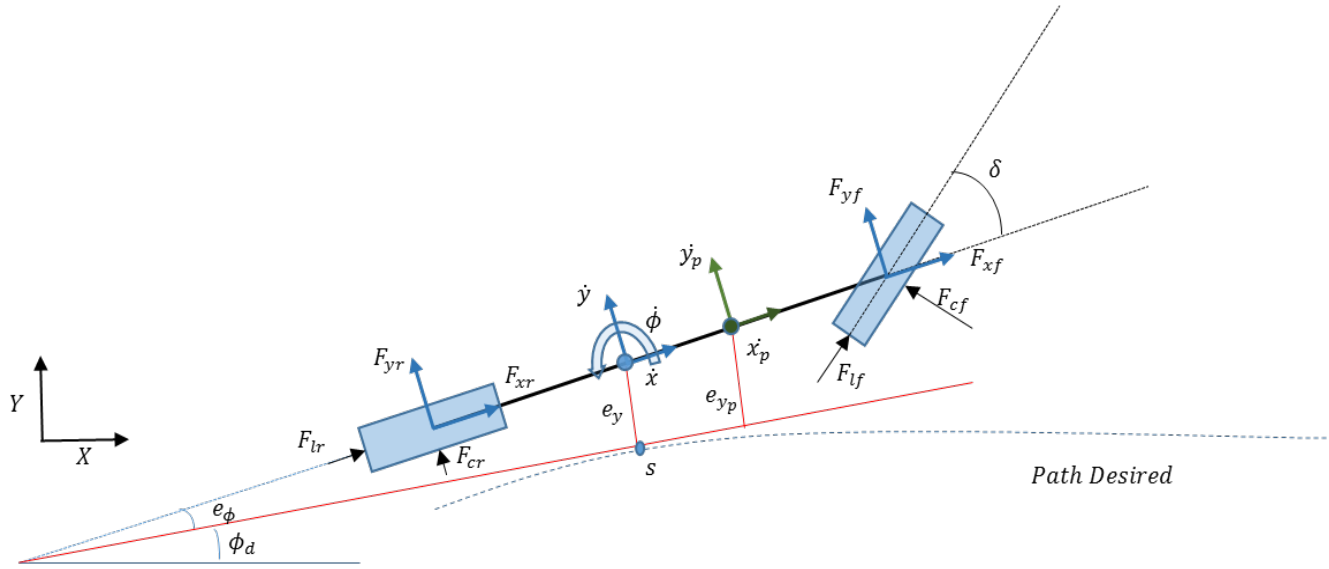


Fig. 1: Single Track Vehicle Model. (F_{xf}, F_{yf}, F_{xr} and F_{yr}) depict the forces in the vehicle body-fixed frame, (F_{lf}, F_{cf}, F_{lr} and F_{cr}) depict forces in the tire-fixed frame, $\dot{\phi}$ depict the rotational velocity about the CG, and ($\dot{y}$ and $\dot{x}$) represent translational velocities. The desired path reference frame coordinates e_y and e_ϕ are illustrated along with ϕ_d. p denotes the distance from CG to CP and ($\dot{y}_p$, $\dot{x}_p$) are the translational velocities about the CP.

II. Vehicle Dynamics and Tire Model

This section presents the vehicle dynamics model and tire model of the control system, and presents the process of linearization and discretization for developing the Linear Time Varying (LTV) system model of the vehicle. It presents a nonlinear single track model, describing the lateral dynamics [13], and introduces the concept of Center of Percussion (CP) for reducing the modeling error due to linearization. Further, it introduces the tire model and details the process of linearization and discretization of the nonlinear model using the tire model presented.

A. Single Track Vehicle Model

Fig. (1) shows a schematic diagram of a single track model. The force balance and path tracking error dynamics leads to the following set of differential equations to describe the vehicle motion.

$$m\ddot{y} = -m\ \dot{x}\dot{\phi} + 2F_{yf} + 2F_{yr} \tag{1}$$
$$I_z\ddot{\phi} = \ 2aF_{yf} - 2bF_{yr} \tag{2}$$
$$\dot{e}_\phi = \dot{\phi} - \dot{\phi}_d \tag{3}$$
$$\dot{e}_y = \dot{y}\ cos\,(e_\phi) + \dot{x}\ sin(e_\phi) \tag{4}$$
$$\dot{s_d} = \frac{\dot{x}\ cos\,(e_\phi)\ - \dot{y}\ sin(e_\phi)}{(1 - K(s_d)e_y)} \tag{5}$$

where m and I_z denote the vehicle mass and yaw inertia, respectively, a and b denote the distances from the vehicle center of gravity to the front and rear axles, respectively. $\dot{x}$ and $\dot{y}$ denote the vehicle longitudinal and lateral velocities, respectively, and $\dot{\phi}$ is the yaw rate around a vertical axis at the vehicle's center of gravity. e_ϕ and e_y in fig. (1) denote the vehicle orientation and lateral positions, respectively, in a path aligned coordinate frame and ϕ_d is the angle of the tangent to the path in a fixed coordinate frame. s_d is the vehicle longitudinal position along the desired path. F_{yf} and F_{yr} are the front and rear tire forces acting along the vehicle lateral axis, F_{xf} and F_{xr} forces acting along the vehicle longitudinal axis, and $K(.)$ denotes the radius of curvature at path distance s_d. Equations (3), (4) and (5) are from the path centered reference frame which directly provides tracking error. Here, the assumption is that, the steering system in the vehicle can only control the angle of front wheels. The longitudinal and lateral tire force components are modelled as

$$F_{yf} = F_{lf}\ sin\,(\delta_f) + F_{cf}\ cos(\delta_f) \tag{6}$$
$$F_{yr} = F_{cr} \tag{7}$$

The steering angle is assumed small and therefore, $F_{yf} \approx F_{cf}$ in (6). Tire forces F_{lf} and F_{yf} are defined by the brush tire model [17] and modified in [18] as a function of slip angles, α_*, where $(*)$ denotes either the front (f) or the rear (r) tire model specific parameters.

$$
\begin{aligned}
F_{y[f,r]} &= \begin{cases} \begin{bmatrix} -C_{\alpha_*} \tan\alpha_* \\ +\frac{C_{\alpha_*}}{\theta} |\tan\alpha_*| \tan\alpha_* \\ -\frac{C_{\alpha_*}}{3\theta^2} \tan^3\alpha_* \end{bmatrix}, & |\alpha_*| < \tan^{-1}\theta, \\ & \theta = \frac{3\mu F_{z_*}}{C_{\alpha_*}} \\ -\mu F_{z_*} sgn(\alpha_*), & \text{otherwise} \end{cases} \\
&= f_{tire}(\alpha_{[f,r]}, F_{z[f,r]})
\end{aligned}
\tag{8}
$$

where μ is the tire-road surface coefficient of friction, F_{z_*} is the normal force and C_{α_*} is the tire cornering stiffness. The parameter C_{α_*} can be estimated from the experimental ramp steer data [19], the coefficient of friction can be estimated online using approaches such as mentioned in [20] [21]. Due to assumption of constant speed in this problem setup, $F_{z[f,r]}$ has been calculated from the geometrical static weight distribution.

The tire slip angles $\alpha_{[f,r]}$ are described in terms of the vehicle states and steering angle input by using small angle approximation as

$$
\begin{aligned}
\alpha_f &= tan^{-1}\left(\frac{\dot{y}+a\ \dot{\phi}}{\dot{x}}\right) - \delta \approx \frac{\dot{y}+a\ \dot{\phi}}{\dot{x}} - \delta \\
\alpha_r &= tan^{-1}\left(\frac{\dot{y}-b\ \dot{\phi}}{\dot{x}}\right) \approx \frac{\dot{y}-b\ \dot{\phi}}{\dot{x}}
\end{aligned}
\tag{9}
$$

Clearly, from (6), (8) and (9), it can be observed that the relationship between the front lateral force and the front slip angle is non-convex. To linearize the input-output system, the control action could be directly front lateral force command (instead of steering angle). The relationship between steering angle, which is the real input, to front lateral force depends on vehicle states, but can be extracted outside the optimization solver. A low level look-up table then can be designed to extract the steering angle from of the lateral force command and current states, or a tire model similar to (8) can be utilized. In this paper, the tire model approach has been followed.

For the rear tire , this technique cannot be applied due to lack of actuation at the rear axle (9). The rear lateral force is a function of the vehicle states and the function is a non-convex. As a result, typically a linearization of rear tire model is employed using a constant linear gain [22].

$$
F_{yr} \approx F_{cr} = C_{\alpha r}\alpha_r = \frac{C_{\alpha r}}{\dot{x}}(\dot{y} - b\dot{\phi}) \tag{10}
$$

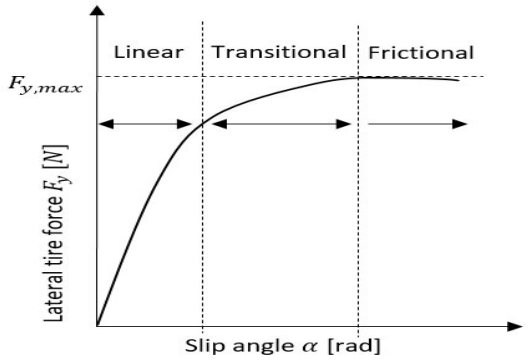
where gain C_{α_r} is called cornering stiffness.

Using a linear gain for representation of linear tire force limits the tire fidelity to the linear region of the tire force curve as illustrated by the fig. (2a). The linear gain model for the non-linear curve region can overestimate the tire force which can result in the instability of the vehicle [23]. To reduce the tire modeling error the rear tire force can be determined by an affine linearization approach. The affine linearization is expressed in (11) after linearization of the (8) at a slip angle of $\overline{\alpha}_r$ which is estimated from the measurement of the states using (9), where $\overline{C}_{\overline{\alpha}_r}$ is the resulting cornering stiffness at the same slip angle. In this paper $(\overline{*})$ notation represents an estimated parameter for the vehicle dynamics model based on available state measurements.

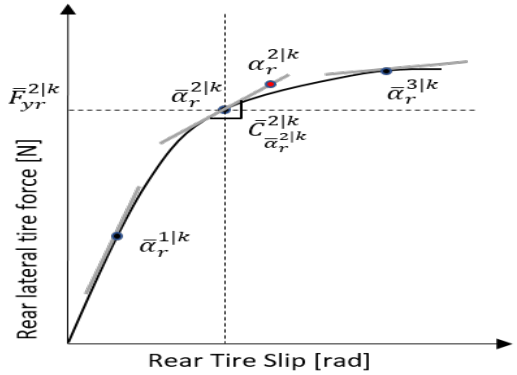
Model predictive controller computes a sequence of inputs at different prediction steps in its horizon resulting in an optimized trajectory at each controller execution step and this further facilitates determination of the rear tire force by successive linearization of the tire model (8) expressed in (12) where $\overline{C}^{i|k}_{\overline{\alpha}_r^{i|k}}$ is the local cornering stiffness estimated at a predicted slip angle $\overline{\alpha}_r^{i|k}$ for a prediction step i at the control execution step k. Fig. (2b) illustrates the process of determination of rear tire force model for each prediction step using the sequence of predicted slip angles $\overline{\alpha}_r^{i|k}$. This successive linearization of the tire model for the MPC converges within a few initial controller execution steps [13][24] and although, it changes the dynamic model from time invariant to time varying, it reduces the modelling errors for lateral force estimation by accurately capturing the non-linear relationship of vehicle states and the rear tire force.

$$
F_{yr} \approx F_{cr} = \overline{F}_{yr} - \overline{C}_{\overline{\alpha}_r}(\alpha_r - \overline{\alpha}_r) \tag{11}
$$

$$
F_{yr}^{i|k} \approx F_{cr} = \overline{F}^{i|k}_{yr} - \overline{C}^{i|k}_{\overline{\alpha}_r^{i|k}}(\alpha_r - \overline{\alpha}_r^{i|k}) \tag{12}
$$



(a) Characteristic curve between tire slip angle and lateral tire force.



(b) Rear tire model, based on previous control step k-1, rear slip angles ($\bar{\alpha}_r^{i|k}$, i=1,2,3) predictions, represented by heavy dot marker with potential linearizations (represented by solid grey lines) at each step. The $\alpha_r^{2|k}$ depicts the current step k, rear slip angle solution at prediction step 2, represented by red dot marker.

Fig. 2: Lateral tire force model

Further, loss of accuracy loss due to linearization of rear tire force can be restricted by writing the equation of lateral motion (1) in vehicle body reference frame with respect to Center of Percussion (CP) instead of Center of Gravity (CG) [25]. CP, as shown in fig. (1), is a point along the vehicle longitudinal axis. As observed in (1) and (2), the rear tire lateral forces have two effects on dynamics, a constant lateral acceleration on the vehicle body and an angular acceleration around the CG. At CP these two effects cancel each other out and thus, lateral velocity at CP is not influenced by the rear tire forces [26]. The dynamic equations with single rear tire force linearization expressed at CP are:

$$\dot{y}_p = \dot{y} + p\dot{\phi}, \quad \dot{x}_p = \dot{x}, \quad p = \frac{I_z}{mb} \tag{13}$$

$$\ddot{y_p} = -\dot{x_p}\dot{\phi} + \left(\frac{2}{m} + \frac{2a}{mb}\right) F_{yf} \tag{14}$$

$$\begin{aligned} \ddot{\phi} &= \frac{2aF_{yf}}{I_z} - 2b\overline{F}_{yr} + 2b\overline{C}_{\overline{\alpha}_r}\left(\alpha_r - \overline{\alpha}_r\right) \\ &= \frac{2aF_{yf}}{I_z} - \frac{2b\overline{F}_{yr}}{I_z\dot{x}_p} + \frac{2b\overline{C}_{\overline{\alpha}_r}}{I_z\dot{x}_p}\left(\dot{y}_p - (p+b)\,\dot{\phi}\right) \\ &\quad - \frac{2b\overline{C}_{\overline{\alpha}_r}}{I_z}\,\overline{\alpha}_r \end{aligned} \tag{15}$$

$$\dot{e}_\phi = \dot{\phi} - \dot{\phi}_d \approx \dot{\phi} - \dot{x_p}\kappa(\overline{s}_d) \tag{16}$$

$$\dot{e}_y = \left(\dot{y}_p - p\dot{\phi}\right)\, cos\left(e_\phi\right) + \dot{x}_{1p} sin\left(e_\phi\right) \;\approx \dot{y}_p - p\dot{\phi} + \dot{x}_p e_\phi \tag{17}$$

$$\begin{aligned} \dot{s_d} &= \frac{\dot{x}_p\ cos\left(e_\phi\right) - \left(\dot{y}_p - p\dot{\phi}\right) sin\left(e_\phi\right)}{(1-\kappa(s_d)e_y)} \\ &\approx \frac{\dot{x}_p - \dot{y}_p e_\phi + p\dot{\phi}e_\phi}{(1-\kappa(s_d)e_y)} \approx \frac{\dot{x}_p - \dot{y}_p \overline{e}_\phi + p\dot{\phi}\overline{e}_\phi}{(1-\kappa(\overline{s}_d)\overline{e}_y)} \end{aligned} \tag{18}$$

where $\dot{x}_p$ represents longitudinal velocity at CP and assumed to be constant and $[\overline{e}_\phi, \overline{e}_y, \overline{s}_d]$ is the heading error, lateral error and path distance measurements respectively for the controller execution step. $\kappa(s_d)$ represents the radius of curvature of the reference path at s_d and $\overline{s_d}$ denotes the distance measurement for the controller execution step.

Assumption 1: Approximations for $\dot{e}_y$ and $\dot{s_d}$ assume small e_ϕ. Equation (18) assumes constant $[\overline{e}_\phi, \overline{e}_y, \overline{s}_d]$, hence e_ϕ, e_y, and s_d preserve the linearity.

Assumption 2: The signal $\dot{\phi}_d$ is known at every time step over a finite time horizon and is approximated as shown in (16).

The equations (13-18) can be organized in a state space form as:

$$\dot{X} = AX + Bu + L + w \tag{19}$$

where, $X = [\dot{y}_p, \dot{\phi},\ e_\phi,\ e_y, s_d]$ represents the state vector, w represents the additive disturbance in the model, and the matrices A, B and L are shown in the following equation.

$$\begin{aligned} A &= \begin{bmatrix} 0 & -\dot{x}_p & 0 & 0 & 0 \\ \frac{2b}{I_z \dot{x}_p} & -\frac{2b(b+p)}{I_z \dot{x}_p} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & -p & \dot{x}_p & 0 & 0 \\ -\frac{\overline{e}_\phi}{(1-\kappa(\overline{s}_d)\overline{e}_y)} & \frac{p\overline{e}_\phi}{(1-\kappa(\overline{s}_d)\overline{e}_y)} & 0 & 0 & 0 \end{bmatrix} \\ B &= \begin{bmatrix} \frac{2}{m} + \frac{2a}{mb} \\ \frac{2a}{I_z} \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad L = \begin{bmatrix} 0 \\ -\frac{2b\overline{F}_{yr}}{I_z \dot{x}_p} - \frac{2b\overline{C}_{\overline{\alpha}_r}}{I_z}\ \overline{\alpha}_r \\ -\dot{\phi}_d \\ 0 \\ \frac{\dot{x}_p}{(1-\kappa(\overline{s}_d)\overline{e}_y)} \end{bmatrix} \end{aligned} \tag{20}$$

The pair (A,B) is not stabilizable because path distance state, s_d, is not controllable by the steering input. The state is therefore treated as an augmented state in the system model and the system matrices are partitioned according to stabilizable and augmented states as shown below

$$\dot{X} = \begin{bmatrix} A_{stab} & 0 \\ A_{aug} & 0 \end{bmatrix} X + \begin{bmatrix} B_{aug} \\ 0 \end{bmatrix} u + \begin{bmatrix} L_{stab} \\ L_{aug} \end{bmatrix} + w \tag{21}$$

After incorporating the sequence of rear tire force linearizations (12) and the sequence of state information from the previous time instant k, the system model can then be expressed as sequence of models:

$$\dot{X} = \begin{bmatrix} A_{stab}^{i|k} & 0 \\ A_{aug}^{i|k} & 0 \end{bmatrix} X + \begin{bmatrix} B_{aug}^{i|k} \\ 0 \end{bmatrix} u + \begin{bmatrix} L_{stab}^{i|k} \\ L_{aug}^{i|k} \end{bmatrix} + w \tag{22}$$

where,

$$A^{i|k} = \begin{bmatrix} 0 & -\dot{x}_p & 0 & 0 & 0 \\ \frac{2b}{I_z \dot{x}_p} & -\frac{2b(b+p)}{I_z \dot{x}_p} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & -p & \dot{x}_p & 0 & 0 \\ -\frac{\overline{e}_\phi^{i|k}}{(1-\kappa(\overline{s}_d^{i|k})\overline{e}_y^{i|k})} & \frac{p\overline{e}_\phi^{i|k}}{(1-\kappa(\overline{s}_d)\overline{e}_y^{i|k})} & 0 & 0 & 0 \end{bmatrix} \tag{23}$$

$$L^{i|k} = \begin{bmatrix} 0 \\ -\frac{2b\overline{F}^{i|k}{}_{yr}}{I_z \dot{x}_p} - \frac{2b\overline{C}^{i|k}}{I_z} \overline{\alpha}^{i|k}{}_r \\ -\dot{x_p}\kappa(\overline{s}_d^{i|k}) \\ 0 \\ \dot{x}_p \end{bmatrix}$$

B. Model Discretization

After linearization, discretization is the next critical step in the model formulation. In this paper a Zero Order Hold (ZOH) is applied to discretize the model using matrix exponential method [27]. In addition, similar to [13], prediction horizon is discretized with variable time steps approach. To meet the real time computation requirement, the number of discretization steps is limited to $N_p = 33$ in accordance with the computational power. Hence, the step size depends on the tradeoff between modeling and prediction accuracy, and length of prediction horizon. Obviously larger prediction horizon is preferred, but it results into larger step size reducing prediction accuracy. In this study, two time steps were used namely, short time-step ($T_{ss} = 30\ ms$) and long time-step ($T_{ls} = 200\ ms$). 27 short time steps, i.e. N_{ss}= 27 and 6 long time steps i.e. N_{ls}= 6 are selected to provide satisfactory path tracking and vehicle stabilization performance. The total of N_p=N_{ss}+N_{ls}= 33 steps results in a time horizon of 2.1 s. The number of short time steps selected are more than the control horizon steps $N_c = 10$ selected for the proposed system, thereby, providing the controller a sufficient control over the dynamics involved. Moreover, the the designed control system's execution rate is set at 30 ms, and the short time steps at this rate provides required fidelity for vehicle dynamics modeling and fast convergence in the case of rear tire force successive linearization. In our evaluation, the computation time for the convex optimization was approximately recorded as 10 ms on a single core of CPU E5-1650 @ 3.5 GHz. Therefore, the 30 ms execution rate of designed control system, provides more time for the developed system to evaluate multiple admissible convex tube spaces as discussed in (III-C)

Equation (24) represents the model after discretization with the ZOH and this Linear Time Varying (LTV) model is utilized for prediction in RMPC based control system in this paper.

$$\begin{aligned} X_{i+1|k} = & \begin{bmatrix} A_{stab,d}^{i|k} & \overline{A}_{aug,d}^{i|k} \\ A_{aug,d}^{i|k} & \mathbf{1} \end{bmatrix} X_{i|k} + \begin{bmatrix} B_{aug,d}^{i|k} \\ 0 \end{bmatrix} u_{i|k} \\ & + \begin{bmatrix} L_{stab,d}^{i|k} \\ L_{aug,d}^{i|k} \end{bmatrix} + w_{i|k}, \quad for\ i\ =\ 0\ \ldots\ N_p \end{aligned} \tag{24}$$

III. Robust Model Predictive Controller (RMPC)

A. Control Structure and feedback Gains

The control input structure for the robust MPC framework in this study is defined by:

$$u_{0|k}^* = \begin{bmatrix} K_{aug}^{0|k} & 0 \end{bmatrix} X_{0|k} + c_{0|k} \tag{25}$$

where, $u_{0|k}^*$ represents the control input computed at controller execution step k, $K^{0|k}{}_{aug}$ is the infinite horizon Linear Quadratic Regulator (LQR) optimal feedback gain for the pair $((A_{stab}^{0|k}, B_{stab}^{0|k})$. The $c_{0|k}$ is the first output of the open loop MPC optimization. The LQR optimal gain $K^{i|k}{}_{aug}$, needs to be obtained for each prediction step i using the method proposed in [28] due to the time-varying nature of the linear model described by $((A_{stab}^{0|k}, B_{stab}^{0|k})$. In order to meet the real time constraint, optimal $K^{i|k}{}_{aug}$ are computed only for first N_c short time steps and 1 long time step as shown in the equations below.

$$u_{i|k} = \begin{bmatrix} K_{aug}^{i|k} & 0 \end{bmatrix} X_{i|k} + c_{i|k}, \quad \forall\ i = 0 \ldots .\ N_c \tag{26}$$

$$u_{i|k} = \begin{bmatrix} K_{aug}^{N_c|k} & 0 \end{bmatrix} X_{i|k} + c_{N_c|k}, \ \forall\ i = N_c\ +1, \ldots, N_{ss} - 1 \tag{27}$$

$$u_{i|k} = \left[K_{aug}^{N_{ss}|k} \; 0 \right] X_{i|k} + c_{Nc|k}, \; \forall \, i = N_{ss}, \ldots, N_p - 1 \tag{28}$$

For the optimal LQR at each step, there are two (possibly) conflicting objectives, namely the stability of the closed loop system in presence of disturbances and physical limits on steering angle. The typical trade-off between these objectives is handled by appropriately choosing weights on each objective.

B. Stabilization Constraints

Each vehicle has limitations in its handling capabilities due to existence of limitations in the grip of the tire. These limitations come forth whenever the vehicle requires forces from the nonlinear regions of the tire force curves during a maneuver. To operate the vehicle at these limits without allowing it to destabilize is a high skill task. In the case of autonomous vehicles, the controller can be expected to have an envelope of handling limits. Various ways have been proposed in past to consider vehicle stability limits.

For example, [29] and [10] use states phase plots with side-slip angle for constant steering input to find the unstable regions. Based on the tire maximum force generation and understanding of the equilibrium point trends for the range of steering and braking inputs, [10] introduced an invariant stable handling envelope for the vehicle. With this knowledge of a safe handling envelope , [12] further showed that unlike Electronic Stability Control (ESC), which only acts when the vehicle is in the unstable region, the approach for stabilizing the autonomous vehicle should prevent it from entering the unstable region. This is the approach we have followed in the proposed control.

Similar to [10], we bound our vehicle's lateral velocity $\dot{y}_p$ and r (vehicle yaw rate at CG) by using the maximum steady-state force generation of the tires. In order to ensure the stability of the vehicle at a given longitudinal speed, the rear tire slip angle is constrained to the linear region of tire force curve. In this study, the rear tire saturation slip angle (α_r) is converted into a bound on $\dot{y}$ and by using (13), it is translated then to a bound on $\dot{y}_p$. It is formally expressed for a given longitudinal speed ($\dot{x}_p$) as

$$\dot{y}_{p,max} = \dot{x}_p \alpha_{r,sat} + (p + b)\, r \tag{29}$$

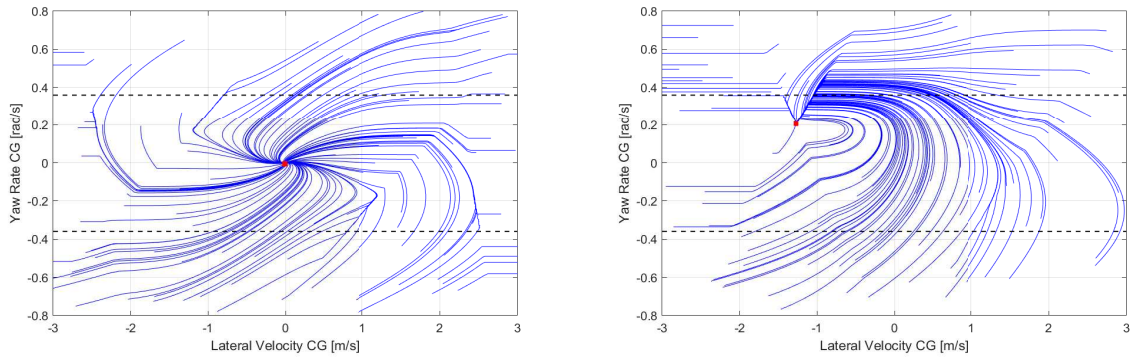
where, $\dot{x}_p$ represents the longitudinal velocity at CP, as shown in fig. (1), and $\alpha_{r,sat}$ is the rear tire saturation angle defined in the tire model (8).

The other constraint is on the yaw rate (r). The bounds as stated in (30) is based upon the steady state yaw rate condition. If the maximum force at front axle $F_{yf,max}$ is less than the rear maximum force $F_{yr,max}$, the maximum yaw rate is then governed by the front axle, $F_{yf,max}$. In this work, constant longitudinal speed has been assumed and therefore, the vehicle is assumed to have a neutral steer. The neutral steer assumes simultaneous tyre saturation of the front and rear axles. Also the nominal forces on the rear and front tires are approximated by the static forces F_{zf} and F_{zr}, respectively. This assumption hold true due to no brake actuation utilized by the controller. This transforms the minimum condition in (30) to a single value of $\mu g / \dot{x}_p$, where μ is the tire road friction coefficient and g denotes the acceleration due to gravity. The phase plane plots of yaw rate v/s slip angle for a neutral steer has been shown in [19]. These plots shows the trajectories for a range of yaw rate and slip angle combination, with a fixed steering and longitudinal speed. For the neutral dynamics [29] [19], showed that there is a single, stable equilibrium point and two saddle points for steering angles upto $10°$, at a speed of 10 m/s. Fig. (3) shows the progression of the neutral steer dynamics for the two steering angles, at a speed of $15m/s$, in the phase plane plots similar to [19]. Due to presence of saddle equilibrium in the region of high yaw rate(greater than the dotted lines) and high side slip angles, the trajectories diverge and for the initial conditions starting within the yaw rate bound converges to the stable equilibrium point. This shows that the steady state yaw rate bifurcates the phase plane plot into stable region and unstable region for different steering angles at a fixed longitudinal speed. Therefore, the bound on yaw rate, described by (30), is selected as the second set of stability constraints in this study.

$$\begin{aligned} r_{max} &= min\{ \; \frac{F_{yf,max}(1 + a/b)}{m\dot{x}_p} \; , \; \frac{F_{yr,max}(1 + b/a)}{m\dot{x}_p} \; \} \\ &= \frac{\mu g}{\dot{x}_p} \end{aligned} \tag{30}$$

The bound on CP lateral velocity in (29) and the yaw rate in (30), creates a closed bound set for the phase plane of lateral velocity (CP) and yaw rate, as shown in fig. (4). Vehicle stability is guaranteed for all states in the envelope formed by (30) and (31). The invariance of the set has been proved in the [10]. Exceeding not necessarily cause the instability of the vehicle, as it was shown in [13] [14], however, for the states outside the envelope, the availability of a steering control input to move the vehicle closer to the boundary in the next time step cannot be guaranteed.

In this work, during optimization the yaw rate constraint is penalized more than on the lateral velocity due to designed control system planning for trajectories that utilizes the maximum rate of change of force inputs for obstacle avoidance.



(a) (b)

Fig. 3: The phase portrait of a neutral steering vehicle at $\dot{x}_p$=15 m/s, $\mu = 0.55$, static front axle load $Fz_f = 7.4$ kN and rear axle load $Fz_r = 4.93$ kN, and (a) 0° and (b) 5°. The thicker dotted lines represents, the maximum-yaw rate computed in (30), examples of the trajectories as thin lines and stable equilibrium as heavy dot marker.

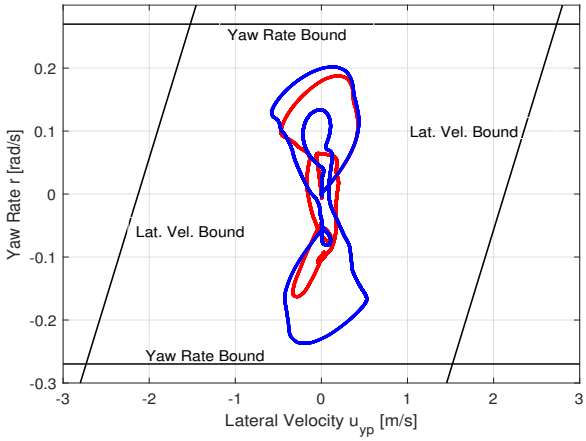


Fig. 4: Stability Envelope. A closed bounded set represented by the yaw rate and lateral velocity (CP) bounds in a yaw rate v/s lateral velocity phase plane. The blue and red are examples of two safe trajectories of the vehicle that adhered to the safety envelope.

The stabilization constraint for the controller is organized in the form expressed as:

$$E_{stab} X_{i|k} \leq G_{stab} \tag{31}$$

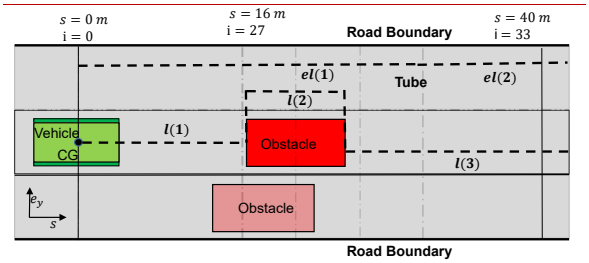
where the $E_{stab} \in \mathbb{R}^{l \times n}$ and $G_{stab} \in \mathbb{R}^{l \times 1}$.

C. Collision Constraints

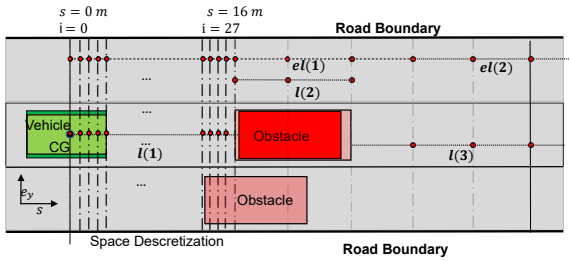
In order to ensure that the planned trajectories avoid the obstacles and stay within the edges of the road, the collision constraints are defined for each control execution step. Obstacles and road edges combined together form an admissible region (search space) for the path planning objective around the ego vehicle. In real-world scenarios, each obstacle provides two options to pass an obstacle, either from the right or from the left (if the space permits). In this section, only left pass constraint computation has been explained. It is assumed the decision to take left or right can either be an output of the higher level decision layer or could be an outcome of simultaneously solving the optimization for both set of admissible regions and selecting the direction based on the one with the reduced cost. Therefore, without loss of generality, we have shown a maneuver from left of the obstacle in fig. (5).

The admissible region can be defined by the constraints called active constraints. In this study we assume that the perception layer (Camera, RADAR etc. along with detection algorithm and coordinate transformations) is capable of generating the locations of obstacles and road edges in the path reference frame [e_y,s_d]. These locations are then expressed in form of linear inequalities on path error states as shown in fig. (5a).

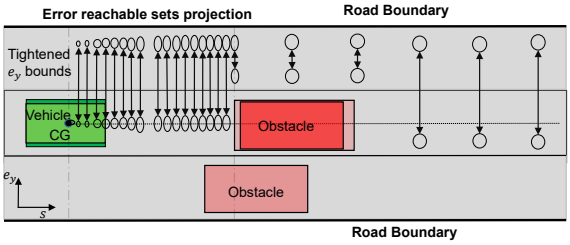
The introduction of error states simplifies the constraint process, as the stability constraints are defined using the states represented in body centered frame of reference, whereas collision constraints on the error states.



(a) Active Constraints. $el(1)$, $el(2)$, $l(1)$, $l(2)$ and $l(3)$ represent active constraints in the prediction horizon. Tube represented by combining the active constraints.



(b) Time Varying Space Constraints. The heavy red dots at a fixed path distance, forms the bound on the lateral error e_y. The transparent red box around the obstacle represents the stretched obstacle lengths. Dash-dot lines illustrate Prediction horizon space discretization with both time steps, T_{ls} long, and T_{ss} short.



(c) Robust Collision Constraints. The ellipsoids represent the error reachable set projection on the time varying space constraints. The double arrow line at a prediction step i represents the tightened bound for the lateral error state $e_y^{i|k}$. Dashed line represents the desired path.

Fig. 5: Illustration of robust collision constraints development process at a controller execution step k. Vehicle assumed to have longitudinal speed $\dot{x}_p$ = 20 m/s

$$E^j_{coll} X_{i|k} \leq G^j_{coll}\ ,\ j = 1 \dots l \tag{32}$$

where l denotes number of active constraints for the states at instant k, that can be combined to form a tube or a search space around the ego vehicle as shown in fig. (5a). The tube defined is then discretized using the path distance estimation $s_{di|k}$ for each prediction step. The path distance is estimated using the discretized distance equation in (24) and the information of time steps T_{ss} and T_{ls}. The transformed tube is represented as

$$E^{i|k}_{coll} X_{i|k} \leq G^{i|k}_{coll} \tag{33}$$

where $E^{i|k}_{coll} \in \mathbb{R}^{q \times n}$ and $G^{i|k}_{coll} \in \mathbb{R}^q$ consists of constraints on the normal path error state $e_{y_{i|k}}$, defining the gap in the search space for the i^{th} prediction step as shown in fig. (5b). It should be be noted that before the discretization step the obstacle lengths are stretched, illustrated in fig. (5b). The example illustrated in fig. (5) is shown for the obstacles that are stationary

and it is assumed that no information of their intended trajectories are available. However, if the prediction is available, the tube could be adapted due to its time varying nature.

Now that the discretized admissible region in established, the next step is to include projection of the errors induced by the additive disturbance by tightening the constraints in (33). The equation (33) can be expressed with constraints tightening as:

$$E_{coll}^{i|k}\left(s_{i|k}+e_{i|k}\right) \leq G_{coll}^{i|k}, \qquad i=0,1,\ .\ \ldots N_p \tag{34}$$

where $s_{i|k}$ and denote $e_{i|k}$ the nominal state and error in the state induced due to disturbance at the prediction step i, respectively. The time varying constraint (34) can be stated as

$$E_{coll}^{i|k} s_{i|k} \leq G_{coll}^{i|k} - h_i, \qquad i=0,1,\ .\ \ldots N_p \tag{35}$$

where h_i is defined as

$$h_i = \underset{e_{i|k}\in S^i}{max}\ E_{coll}^{i|k} e_{i|k}\ , \qquad i=1,2,\ \ldots N_p \tag{36}$$

The S^i in (36) represents the reachable sets for the state error due to disturbance at each prediction step till step i. These reachable sets follow the constraint

$$\phi^i S^i \oplus W \subseteq S^{i+1} \tag{37}$$

where $\phi^i = A^i + B^i K^i$. The (36) using the (37) constraint can be redefined as

$$h_i = \sum_{m=0}^{m=i-1} \underset{w_{m|k}\ \in W}{max} \left[E_{coll}^{i|k} . \left\{ \prod_{k=i-m-1}^{k=0} \phi^k \right\} . w_{m|k} \right] , \ i=1,2,\ \ldots N_p \tag{38}$$

In equation (38), computation of h_i for N_p steps require total number of $(2N_p+1)q$ linear programs, where N_p denotes the prediction horizon and q denotes the number of inequalities present at step i. Now, due to discretization of space, required for formulation (34), reduced the number of linear inequalities to $q=2$ for each prediction step i, as shown in fig. (5b) by the red dots. This reduction in q resulted in limiting the number of linear programs required for h_i to $(4N_p+2)$. The result shows that the number of linear programs to be solved are independent of any number of obstacles in the search space of the ego vehicle. This is a significant result from the point of keeping the additional computational load required for providing robustness in check.

Note that in equation (38), the h_i are defined for the entire prediction horizon. In practice the MPC control considers only N_c steps. This helps in avoiding the excessive reduction of the search space for prediction steps beyond the control horizon for which controller is incapable to plan. Also, in order to reduce computation time, the h_i terms for N_c+1 to N_p are considered constant, while the h_i are explicitly computed for first N_c steps as shown in the equation below.

$$\begin{aligned} h_i = \sum_{m=0}^{m=i-1} \underset{w_{m|k}\ \in W}{max} \left[E_{coll}^{i|k} . \left\{ \prod_{k=i-m-1}^{k=0} \phi^k \right\} . w_{m|k} \right] , \\ i=1,2,\ \ldots N_c,\ G_{coll}^{i|k} \geq h_i \geq \mathbf{0} \\ h_i = h_{N_C}, \qquad i=N_c{+}1 \ldots\ldots N_p,\ \ G_{coll}^{i|k} \geq h_i \geq \mathbf{0} \end{aligned} \tag{39}$$

In (39), the selected bounded disturbance set W is a simple bound based on the expected modelling error and state measurements error. It is a critical set and the next section (IV) in this paper covers the process of determining it in more detail.

We further simplified the constraints and implemented as a time varying bound on the lateral error, e_y, state, as shown in fig. (5c) and expressed as

$$\overline{e}_{y,\ k+i|k,min} \leq e_{y,i|k} \leq \overline{e}_{y,k+i|k,max}\ , \tag{40}$$

The tube formed by the time dependent constraints in (40) are not convex and thus, further requires convexification for the convex optimization. Similar to successive linearization concept that uses MPC's previous step optimization information, we use heading error angle state (e_ϕ) from the last step to adjust the bounds in (40). It is expressed as

$$\begin{aligned} \overline{e}_{y,\ k+i|k,min} \ + f_{width}(e_{\phi_{i|k-1}}) \leq e_{y,i|k} \\ \overline{e}_{y,k+i|k,max} \ - f_{width}(e_{\phi_{i|k-1}}) \geq e_{y,i|k} \end{aligned} \tag{41}$$

where, the function $f_{width}\left(\phi_{i|k-1}\right)$ approximates the vehicle's effective width [30] based on the orientation information gleaned from the prediction step i of last time instant $k-1$ optimization as the path changes. If $e_\phi = 0$, the $f_{width}\left(e_\phi{=}\ 0\right){=}\frac{w}{2}$, while at $e_\phi{=}\pi/2$, the approximate width is $f_{width}\left(e_\phi{=}\pi/2\right){=}\ a$, the distance from the center of gravity to the front axle.

With the time varying perception constraint defined in (41), the next section formally introduces the optimal problem solved at every time instant k for the designed control system and it will be followed by the section on results and discussions.

D. Optimization Problem

In this section we formulate the path tracking and obstacle avoidance online RMPC problem for an autonomous vehicle. RMPC, at each sampling time instant k, computes an optimal input sequence or specifically, a lateral force input sequence, by solving a constrained finite horizon optimal control problem. This computed optimal input sequence and its corresponding optimized state trajectory are kept as the nominal input and state trajectories, $\mathbf{c}$ and $\mathbf{s}$ respectively. Using first input of the computed sequence $c_{0|k}$ and the control structure in (25), the augmented control input $u_{0|k}$ is calculated, which is then applied to the vehicle at current step. At the next time instant, the constrained optimization is setup again and solved using a new set of measurements.

$$\begin{aligned} s_{i|k+1} \ = (A^{i|k}{+}B^{i|k}K^{i|k})s_{i|k}{+}B^{i|k}c_{i|k}{+}L^{i|k} \\ = f^{i|k}(s_{i|k}, c_{i|k}) \end{aligned} \tag{42}$$

In this study, the cost function of the optimal control problem is designed to mediate between objectives defined by trajectory re-planning, stability constraints, and the collision constraints. The augmented input is the front tire force calculated based on the discretized model in (45) and converted to a steering command using the tire model in (8) and the relationship in (9). The nominal optimization problem for the optimal control problem is represented as

$$\min_{\mathbf{c},\epsilon_{stab},\epsilon_{coll}} \sum_{i=0}^{N_p-1} \left\|s_{i|k}\right\|_Q^2{+}\left\|c_{i|k}\right\|_R^2{+}\left\|\Delta c_{i|k}\right\|_S^2 \tag{45a}$$

$$+\|\epsilon_{stab}\|_{\lambda_{stab}}^2{+}\|\epsilon_{coll}\|_{\lambda_{coll}}^2$$

$$s_{i|k+1}{=}f^{i|k}(s_{i|k}, c_{i|k}),\ i = 0, 1, \ldots\ldots, N_p{-}1 \tag{45b}$$

$$\overline{e}_{y,i|k,min}{-}\epsilon_{coll}{\leq}\ Hs_{i|k}{\leq}\ \overline{e}_{y,i|k,max}{+}\epsilon_{coll} \quad ,\ i = 0, 1, \ldots..,\ Np \tag{45c}$$

$$E_{stab}^k s_{i|k}\ {\leq} G_{stab}^k{+}\epsilon_{stab} \quad ,\ i = 0, 1, \ldots..,\ Np \tag{45d}$$

$$\epsilon_{stab} \geq \mathbf{0},\ \epsilon_{coll} \geq 0 \tag{45e}$$

$$c_{i|k}{=}\Delta c_{i|k}{+}c_{i-1|k} \tag{45f}$$

$$c_{i|k}{\in}\mathbb{U}, \qquad i = 0, 1, \ldots..,N_c{-}1 \tag{45g}$$

$$\Delta c_{i|k}{\in}\Delta\mathbb{U}, \qquad i = 0, 1,\ldots...,N_c{-}1 \tag{45h}$$

$$\Delta c_{i|k}{=}\ 0 \qquad i{=}N_c, \ldots..,N_p \tag{45i}$$

where $s_{i|k}$ denotes the predicted nominal state at prediction time step i obtained by applying the control sequence $\boldsymbol{c_k}$ $=\{c_{0|k}, \ldots,\ c_{N_c-1|k}\}$ to the model in (45b) with $s_{0|k}{=}\ x_k$. The $s_{ref,i|k}$ represents the reference path of the vehicleThe collision constraints (41) have been imposed as soft constraints, by introducing the slack variable ϵ_{coll} and ϵ_{stab} in (45c) and (45d) respectively. The collision avoidance and stabilization are sometimes conflicting objectives. Therefore, the weights λ_{stab} and λ_{coll}, for the stability and collision constraints violation ensure problem feasibility, and their values encode the tradeoff between the two objectives. The robust collision constraints by including modeling errors proactively, take care of the search space constraints and therefore, in the optimization, the stability constraint is enforced over the collision constraint. The weights $0 \prec Q \in \Re^5 \times \Re^5$, $0 < R \in \Re$ and $0 < S \in \Re$ are penalises on the state tracking error, RMPC control action, and the rate of change of nominal the control action respectively.

It should be noted that the dynamic feasibility of the RMPC is difficult to guarantee under all the possible scenarios. This is due to the nonlinearities of the system and the changing search space at every time instant. However, proposed approach reduces the need for conservative fixed constraints (e.g. [22]) by dynamically updating the limits.

Parameter	Symbol	Value	Units
Vehicle Mass	m	1260	kg
Yaw Moment of Inertia	I_{zz}	1343.1	kg/m^2
Wheelbase	L	2.6	m
Front Axle- CG Distance	a	1.04	m
Rear Axle- CG Distance	b	1.56	m
Vehicle Width	wd	1.695	m
Cornering stiffness of front tire	C_{α_f}	51650	N/rad
Cornering stiffness of rear tire	C_{α_r}	38160	N/rad
Normal force front tire	Fz_f	2704.4	N
Normal force rear tire	Fz_r	2704.4	N
Road surface friction coefficient	μ	0.55	
Radius of the tire	R_{tire}	0.3	m
Height of CG	h	0.54	m

TABLE I: B-Class CarSim Vehicle Parameters

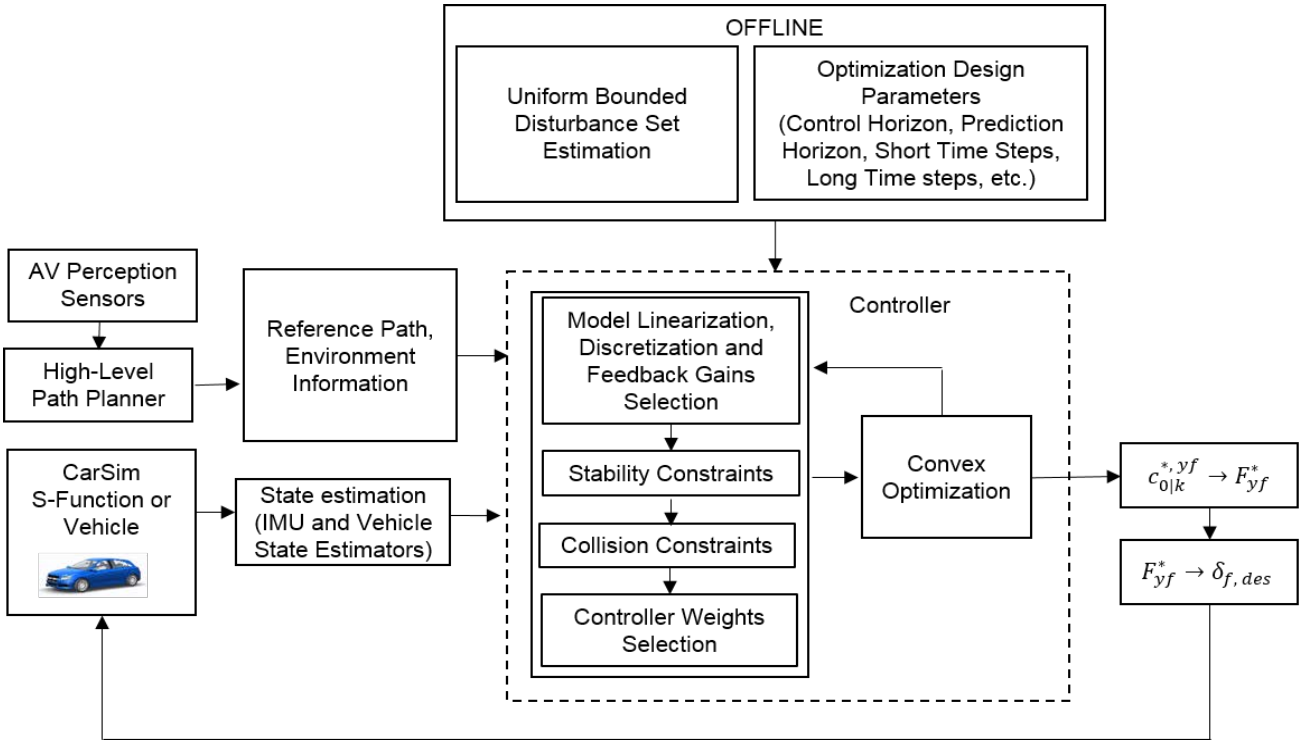


Fig. 6: The block diagram of robust model predictive control for obstacle avoidance and path tracking system of autonomous vehicle

IV. Simulation Study Results

In this section, the designed control system is tested for specific test cases that are provided with a prescribed reference path and the initial longitudinal velocity as inputs. The performance of the designed system is compared with a Linear Time Varying non-robust Deterministic MPC (DMPC) controller equipped with successive linearization of the rear tire forces in order to understand the advantages of robust control design. Three different test cases are presented to demonstrate the system's performance envelope. Simulation is performed using CarSim linked with Simulink software and B-Class model is chosen. The vehicle parameters are shown in Table I. By linking CarSim to Simulink, the full vehicle model with high DOF and system's non-linearities is linked to the designed controller in Simulink. The majority of the researchers use software simulation for testing and validating their designed vehicle control systems [31], [32], [9], [33], [34], [35], [36], [37], [38], [39]. The proposed system's block diagram with CarSim model-in–the-loop is shown in the fig. (6). The convex optimization problem in (45a-45i) and linear program in (39) are solved using CVXGEN solvers [40]. As shown in fig. (1), the augmented force input computed by the MPC is converted into a steering command using the tire model discussed in (8). The Table II lays out the weights for the convex optimization. The weights set in the controller prioritise the yaw rate bound enforcement, $\lambda_{stab}[3,4]$, over the collision constraint λ_{coll}. Table III shows the computation time to optimize the resulting optimization problem (45a-45i) on a single core of CPU E5-1650 @ 3.5 GHz [40].

Parameter Description	Symbol	Value (Avoidance,Tracking)	Units
Control System			
Number of Timesteps	N_p	33, 33	
Controller Timestep	T_s	0.03, 0.03	s
Short Timestep size	T_{ss}	0.03, .03	s
Long Timestep size	T_{ls}	0.2, 0.2	s
Control horizon step	N_c	10, 10	
Timestep change	N_{ss}	27, 27	
Weight Normalization			
Max. expected lateral error	$e_{y,max}$	3.8, 3.1	m
Max. expected heading error	$\Delta\phi_{max}$	0.2, 0.2	rad
Max. expected lateral velocity	u_{yp}	$0.2u_{xp}$, $0.2u_{xp}$	m/s
Max. expected yaw rate	r_{max}	$0.9\frac{\mu g}{u_{xp}}$, $0.9\frac{\mu g}{u_{xp}}$	rad/s
Max. expected force input	$\mathbb{U}_{max}$	μFz_f, μFz_f	N
Max. expected force input change	$\Delta\mathbb{U}_{max}$	12, 12	kN/s
Vehicle states' weights	Q_{diag}	[1,1,1,5,0], [1,5,1,10,0]	
Force input	R	4, 4	
Change in force input	S	4, 4	
Collision slack cost	λ_{coll}	10^5, 10^5	
Stability slack cost	λ_{stab}	$10^2\times$ $[5, 5, 10^3, 10^3]$, $10^4\times$ $[5, 5, 10^2, 10^2]$	

TABLE II: RMPC control system parameters and weights

Task	Time (Avoidance)	Time (Tracking)	Units
Convex Optimization	9 ± 1	7 ± 2	ms
Robust Constraint Setup	4 ± 0.5	4 ± 0.5	ms

TABLE III: Constrained optimization computation results

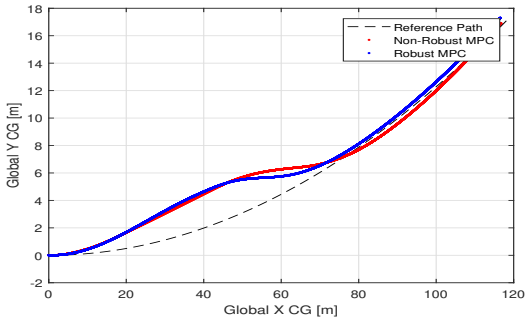
A. Estimation of Disturbance Bound

The RMPC requires estimation of the bounded disturbance set W of the signal $w_{i|k}$ which represents an additive disturbance to the system model in (24). We employ a process similar to [16] to identify this set. We compare the one-step prediction of the model in (24) with the measured vehicle states from the Carsim model. Multiple one-step predictions were performed that mimic the situations encountered during evasive steering with particular emphasis on the mismatch of road friction coefficient errors. The example of one such test trial can be a double lane change maneuver with model assuming $\mu = 0.7$ and Carsim model running at $\mu = 0.5$. In addition to modeling error embedded with residual model uncertainty due to linearization, the estimation of the bound also includes the IMU measurement errors information from the EURO NCAP test protocol [41] measuring equipment requirements for lane support systems. After including all the aforementioned uncertainties, the set W is defined by $W = \{w \in \Re^5 || w | < [0.2, 0.14, 0.0175, 0.025, 0.025]\}$.

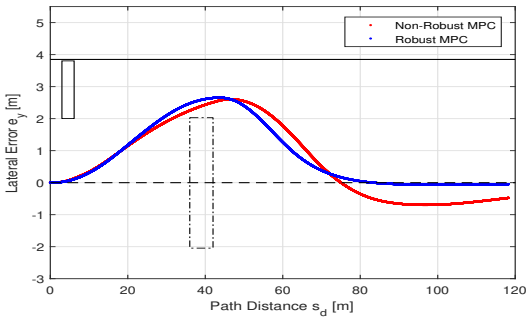
B. Test Cases

In this subsection, the test cases are defined for a curved road with a constant radius, $R = 400$ m and tire-road friction coefficient $\mu \approx 0.55$, as described by the reference path in fig. (7a). The vehicle with initial longitudinal speed $\dot{x}_p = 18$ m/s is expected to execute a lane change maneuver for avoiding the obstacle and another lane change maneuver to return the vehicle to the reference path again. The set of test cases are designed to pose additional challenge by adding the yaw rate requirement of a curved road.

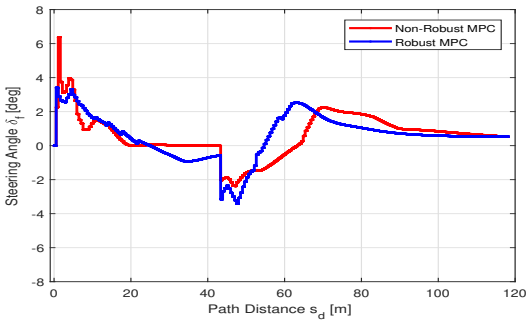
1) Test case: Obstacle at Prediction Horizon: In this example, there is a detection of the obstacle at the prediction horizon of the controller, i.e., $t = 2$ s, on a curved road. The results illustrated by fig. (7b-7e) reveal that both DMPC and RMPC successfully are able to evade the obstacle. The performance of the DMPC is poor, especially, in the post avoidance phase as shown in fig. (7b) , where the overshoot around the path distance $s_d = 80$ m is more than 40 cm. The controllers shown in fig. (7e) are successful in keeping the vehicle inside the imposed stability constraints. The trajectories are closer to the yaw rate bound constraints due to curved road requirement. Overall, due to early detection, both the controllers are able to plan a safe trajectory around the obstacle.



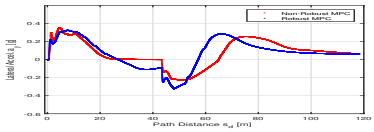
(a) Global reference frame



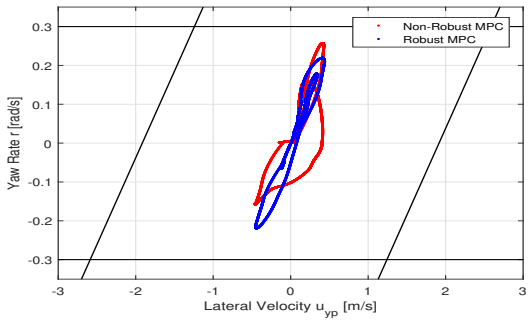
(b) Lateral error



(c) Steering angle input



(d) Lateral acceleration

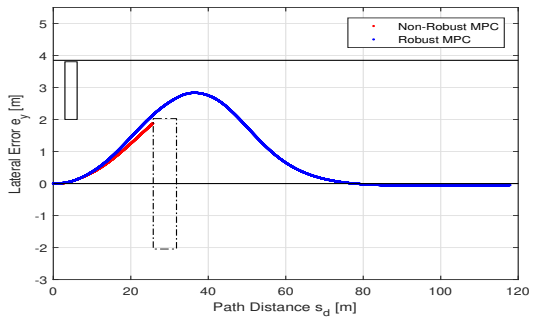


(e) Closed loop trajectories in yaw rate v/s lateral velocity (CP) phase plane with stability constraint

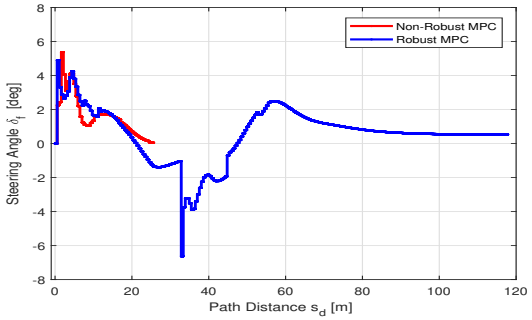
Fig. 7: Test case: Curved road, $\dot{x}_p = 18$ m/s, obstacle at prediction horizon. Both controllers are able to avoid the obstacle and return to the prescribed path because of early obstacle detection

2) Test case: Pop-up obstacle : This test case is an extension of the previous test case which represents either a case of delayed detection or a pop up obstacle in the reference path of the vehicle. Fig. (8) shows the resulting lateral errors, steering angle input, lateral acceleration and phase plane trajectory. Similar to the reasons described in pop-up obstacle on a straight road, the DMPC fails to avoid the obstacle and resulted in a collision. On the other hand, the RMPC controller managed to avoid the obstacle. The difference is the consideration of the model uncertainty in RMPC, resulting in necessary adjustment to the planned trajectory. However, this test case demonstrates the limitations of an obstacle avoidance and path tracking controller system relying only on steering actuation because the trajectory executed by RMPC could only manage to avoid the obstacle by a small gap, as shown in fig. (8a). This is understood as the artifact of not being able to sustain the high lateral acceleration for the initial path distance ($s_d = [5 - 20]$ m), as shown in fig. (8c), due to predicted impending violation of the yaw rate bound, described in fig. (8d).

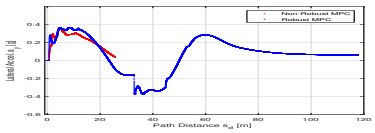
3) Test case: Pop-up Obstacle with disturbance: Fig. (9) shows the results of the pop-up obstacle test case on a curved road. The controllers are set up for the nominal $\mu_c \approx 0.55$ and the tire-road friction coefficient for the CARSIM vehicle solvers is set around $\mu_{act} = 0.35$. The vehicle for RMPC is able to avoid the obstacle and track the center line of the road at the speed of 18 m/s, however, barely able to keep the vehicle within the safe road boundary limits, as shown in fig. (9a). In fig. (9a), controller RMPC shows considerable overshoot $s_d = 85$ m, after avoiding the obstacle. This is because the RMPC maintained high steering command for longer than in previous test case to compensate for the reduced friction/lateral acceleration, as



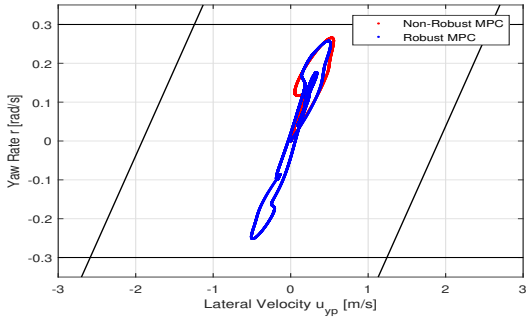
(a) Lateral error



(b) Steering angle input



(c) Lateral acceleration



(d) Closed loop trajectories in yaw rate v/s lateral velocity (CP) phase plane with stability constraint

Fig. 8: Test case: Curved road, $\dot{x}_p = 18$ m/s, pop-up obstacle at $t = 1.4$ s. RMPC controller able to avoid the obstacle and return to the prescribed path, whereas, DMPC failed to find a collision free path.

shown in fig. (8c) and fig. (9b) for $s_d = [5 - 20]$ m.

In a similar case with lower vehicle longitudinal speed, this would have resulted in less path distance covered for the same time duration required and could have resulted in steering angle reduction at path distance earlier than for the current velocity case leading to maneuver not close to the road boundary limit. Therefore, reduction in longitudinal velocity will be explored in future work by including the brake actuation and the result will be compared to the result for the integrated braking and steering actuated control system.

V. Conclusion

This paper presents a fast-online robust MPC framework based lateral control system for obstacle avoidance and path tracking of autonomous vehicle. A discrete linear time varying using successive linearization of a non-linear tire model is developed and used in the control design. A novel method of incorporating the error reachable sets information to a discrete convex search space is developed. A tube-based MPC is used with tightened search space to ensure obstacle avoidance and path tracking in the presence of unknown disturbances. Multiple simulations with different test cases have been run to evaluate the controller performance and its comparison to a deterministic linear MPC controller. The results highlighted the importance of robust design for planning and execution of emergency maneuvers and for path tracking in slippery conditions for autonomous vehicles.

References

[1] Z. Shiller and S. Sundar, "Emergency lane-change maneuvers of autonomous vehicles," *Journal of Dynamic Systems Measurement and Control-transactions of The Asme - J DYN SYST MEAS CONTR*, vol. 120, 03 1998.

[2] X. He, Y. Liu, C. Lv, X. Ji, and Y. Liu, "Emergency steering control of autonomous vehicle for collision avoidance and stabilisation," *Vehicle System Dynamics*, vol. 57, no. 8, pp. 1163–1187, 2019.

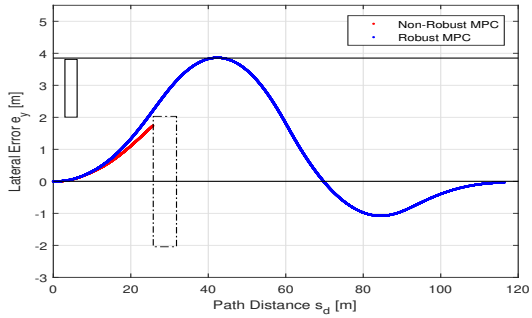
[3] J. Funke, M. Brown, S. M. Erlien, and J. C. Gerdes, "Prioritizing collision avoidance and vehicle stabilization for autonomous vehicles," in *2015 IEEE Intelligent Vehicles Symposium (IV)*, 2015, pp. 1134–1139.

[4] B. Falcone, P., A. F., T. J., H. E., and D. Hrovat, "A hierarchical Model Predictive Control framework for autonomous ground vehicles," *American Control Conference*, 2008.

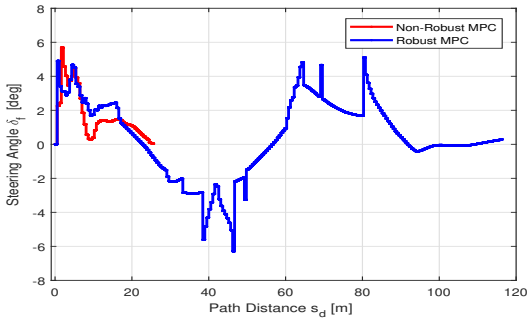
[5] B. Paden, M. Cap, S. Z. Yong, D. Yershov, and E. Frazzoli, "A Survey of Motion Planning and Control Techniques for Self-Driving Urban Vehicles," *IEEE Transactions on Intelligent Vehicles*, vol. 1, no. 1, pp. 33–35, 2016.

[6] O. S. Tas, F. Kuhnt, J. M. Zollner, and C. Stiller, "Towards a Functional System Architecture for Automated Vehicles," *2016 IEEE Intelligent Vehicles Symposium (IV)*, 2016.

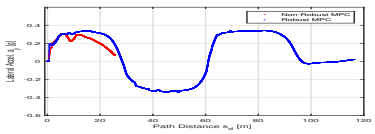
[7] Y. Gao, T. Lin, F. Borrelli, E. Tseng, and D. Hrovat, "Predictive Control of Autonomous Ground Vehicles With Obstacle Avoidance on Slippery Roads," *ASME 2010 Dynamic Systems and Control Conference*, vol. 1, 2010.



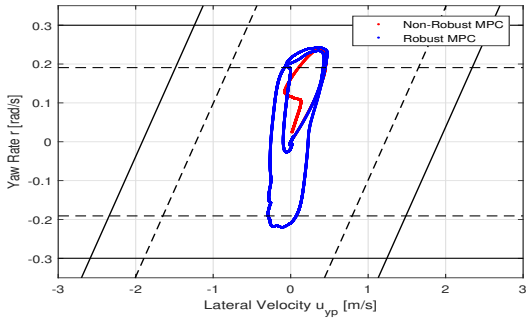
(a) Lateral error



(b) Steering angle input



(c) Lateral acceleration



(d) Closed loop trajectories in yaw rate v/s lateral velocity (CP) phase plane, solid lines represents controller stability constraint, and dotted lines represents actual constraints

Fig. 9: Test case: Straight road, $\dot{x_p} = 18$ m/s, pop-up obstacle at $t = 1.4$ s, $\mu_{act} = 0.35$ and $\mu_c = 0.55$. RMPC able to avoid the obstacle but failed to satisfy the road boundary and overshoot requirement. DMPC failed to find a collision free path.

[8] P. Falcone, F. Borrelli, J. Asgari, H. E. Tseng, and D. Hrovat, "Low complexity mpc schemes for integrated vehicle dynamics control problems," *9th International Symposium on Advanced Vehicle Control*, 2008.

[9] J.-M. Park, D.-W. Kim, Y. Yoon, and K.-S. Yi, "Obstacle avoidance of autonomous vehicles based on model predictive control," *Proceedings of The Institution of Mechanical Engineers Part D-journal of Automobile Engineering - PROC INST MECH ENG D-J AUTO*, vol. 223, pp. 1499–1516, 12 2009.

[10] S. M. Erlien, S. Fujita, and J. C. Gerdes, "Safe Driving Envelopes for Shared Control of Ground Vehicles," *IFAC Proceedings Volumes*, vol. 46, no. 21, pp. 831–836, 2013.

[11] C. E. Beal and J. C. Gerdes, "Model Predictive Control for Vehicle Stabilization at the Limits of Handling," *IEEE Transactions on Control Systems Technology*, vol. 21, no. 4, pp. 1258–1269, 2013.

[12] M. Brown, J. Funke, S. Erlien, and J. C. Gerdes, "Safe Driving Envelopes for Shared Control of Ground Vehicles," *Control Engineering Practices*, vol. 61, pp. 307–316, 2017.

[13] J. Funke, M. Brown, S. M. Erlien, and J. C. Gerdes, "Collision Avoidance and Stabilization for Autonomous Vehicles in Emergency Scenarios," *IEEE Transactions on Control Systems Technology*, vol. 25, no. 4, pp. 1204–1216, 2017.

[14] G. Franzè and W. Lucia, "The obstacle avoidance motion planning problem for autonomous vehicles: A low-demanding receding horizon control scheme," *Systems & Control Letters*, vol. 77, pp. 1–10, 2015.

[15] G. Garimella, M. Sheckells, J. Moore, and M. Kobilarov, "Robust Obstacle Avoidance using Tube NMPC," *Robotics: Science and Systems XIV*, 2018.

[16] Y. Gao, A. Gray, H. E. Tseng, and F. Borrelli, "A tube-based robust nonlinear predictive control approach to semiautonomous ground vehicles," *Vehicle System Dynamics*, vol. 52, no. 6, pp. 802–823, 2014.

[17] H. B. Pacejka, "Tire Characteristics and Vehicle Handling and Stability," in *Tire and Vehicle Dynamics, 3rd ed.* Butterworth-Heinemann, 2012, pp. 1–58.

[18] H. Pacejka, *Tire and Vehicle Dynamics*. Elsevier, 2005.

[19] C. G. Bobier-Tiu, C. E. Beal, J. C. Kegelman, R. Y. Hindiyeh, and J. C. Gerdes, "Vehicle control synthesis using phase portraits of planar dynamics," *Vehicle System Dynamics*, vol. 57, no. 9, pp. 1318–1337, 2018.

[20] C. E. Beal, S. P. Boyd, J. C. Gerdes, and S. Beiker, *Applications of model predictive control to vehicle dynamics for active safety and stability*. Dissertation,Stanford University, 2015.

[21] H. Liu, C.-S.and Peng, "Road Friction Coefficient Estimation For Vehicle Path Prediction," *Vehicle System Dynamics*, vol. 25, no. (supl), p. 413–425, 1996.

[22] Y. Gao, A. Gray, H. E. Tseng, and F. Borrelli, "A tube-based robust nonlinear predictive control approach to semiautonomous ground vehicles," *Vehicle System Dynamics*, vol. 52, no. 6, pp. 802–823, 2014.

[23] S. M. Erlien, J. C. Gerdes, S. P. Boyd, and M. Pavone, *Shared vehicle control using safe driving envelopes for obstacle avoidance and stability*. Stanford University, 2015.

[24] B. Yi, S. Gottschling, J. Ferdinand, N. Simm, F. Bonarens, and C. Stiller, "Real time integrated vehicle dynamics control and trajectory planning with mpc for critical maneuvers," *2016 IEEE Intelligent Vehicles Symposium (IV)*, 2016.

[25] P. Falcone, "Nonlinear model predictive control for autonomous vehicles," Ph.D. dissertation, University of Sannio, Benevento, Italy, 2007.

[26] K. Kritayakirana and J. C. Gerdes, "Using the centre of percussion to design a steering controller for an autonomous race car," *Vehicle System Dynamics*, vol. 50, no. sup1, pp. 33–51, 2012.

[27] G. F. Franklin, J. D. Powell, and M. L. Workman, *Digital Control of Dynamic Systems*. Addison-Wesley, 1998, vol. 3.
[28] A. Bunse-Gerstner, "Computational solution of the algebraic riccati equation." 1996.
[29] S. Inagaki, I. Kshiro, and M. Yamamoto, "Analysis of vehicle stability in critical cornering using phase-plane method," *Proceedings of the International Symposium on Advanced Vehicle Control*, 1994.
[30] J. Funke, M. Brown, S. M. Erlien, and J. C. Gerdes, "Collision Avoidance and Stabilization for Autonomous Vehicles in Emergency Scenarios," *IEEE Transactions on Control Systems Technology*, vol. 25, no. 4, pp. 1204–1216, 2017.
[31] H. Alipour, M. B. B. Sharifian, and M. Sabahi, "A modified integral sliding mode control to lateral stabilisation of 4-wheel independent drive electric vehicles," *Vehicle System Dynamics*, vol. 52, no. 12, pp. 1584—-1606, 2014.
[32] H. Jiang, Z. Wang, Q. Chen, and J. Zhu, "Obstacle avoidance of autonomous vehicles with CQP-based model predictive control," *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2016.
[33] S. Scalzi, Benine-Neto, N. A., P.-L. M., W., and S. Mammar, "Active steering control based on piecewise affine regions," *Proceedings of the 2010 American Control Conference*, 2010.
[34] H. Alipour, M. B. B. Sharifian, and M. Sabahi, "Gain-scheduled robust control for lateral stability of four-wheel-independent-drive electric vehicles via linear parameter-varying technique," *Mechatronics*, vol. 30, pp. 286–296, 2015.
[35] R. Marino, S. Scalzi, and M. Netto, "Integrated driver and active steering control for vision-based lane keeping," *European Journal of Control*, vol. 18, no. 5, pp. 473 – 484, 2012.
[36] D. Soudbakhsh, A. Eskandarian, and D. Chichka, "Vehicle Collision Avoidance Maneuvers With Limited Lateral Acceleration Using Optimal Trajectory Control," *Journal of Dynamic Systems, Measurement, and Control*, vol. 135, no. 4, 05 2013, 041006. [Online]. Available: https://doi.org/10.1115/1.4023943
[37] X. Jin, G. Yin, C. Bian, J. Chen, P. Li, and N. Chen, "Gain-Scheduled Vehicle Handling Stability Control Via Integration of Active Front Steering and Suspension Systems," *Journal of Dynamic Systems, Measurement, and Control*, vol. 138, no. 1, 10 2015.
[38] R. Wang, H. Jing, C. Hu, M. Chadli, and F. Yan, "Robust h∞ output-feedback yaw control for in-wheel motor driven electric vehicles with differential steering," *Neurocomputing*, vol. 173, pp. 676 – 684, 2016.
[39] A. Norouzi, R. Kazemi, and S. Azadi, "Vehicle lateral control in the presence of uncertainty for lane change maneuver using adaptive sliding mode control with fuzzy boundary layer," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 232, no. 1, pp. 12–28, 2018.
[40] J. Mattingley and S. Boyd, "CVXGEN: A code generator for embedded convex optimization," *Optimization and Engineering*, vol. 12, no. 1, pp. 1–27, 2012.
[41] E. NCAP, "Test protocol - lane support systems," 06 2017, lSS-1701-03 Euro NCAP. [Online]. Available: https://cdn.euroncap.com/media/29410/euro-ncap-lss-test-protocol-v20.pdf

Vivek Bithar
Vivek Bithar received the B.S. degree in production and industrial engineering from the Delhi College of Engineering, Delhi, India, in 2009, the M.S. degree in mechanical engineering from University of Michigan, Ann Arbor, MI, USA, in 2010, and the Ph.D. degree in mechanical engineering from the Ohio State University in 2020, with a focus on robust optimal predictive control systems and its application to obstacle avoidance for automated vehicles.

Punit Tulpule
Punit Tulpule received his Bachelors of Science degree in Physics and Master of Science degree in Atmospheric Sciences from University of Pune, India in 2006 and 2008 respectively. He received PhD degree in Mechanical Engineering from Iowa State University in 2014. He is currently a Research Assistant Professor of Mechanical and Aerospace Engineering Department at the Ohio State University (OSU). He has held different researcher positions at Simulation Innovation and Modeling Center and Center for Automotive Research at OSU. His research interests are in validation and verification of complex engineering systems, modeling and simulation of large scale systems, and optimal control.

Shawn Midlam-Mohler
Shawn Midlam-Mohler received his bachelors, masters and PhD degrees from The Ohio State University (OSU) in 1999, 2001 and 2005 respectively. He has held different research faculty and staff positions at OSU's Center for Automovite Research until 2014. Currently he is the Director of Simulation Innovation and Modeling Center and professor of practice in the Mechanical and Aerospace Engineering Department. His research interests are in the area of design, system modeling, applied controls, and systems engineering. He has over one hundred patents and peer-reviewed publications and his main research focus is in the area of advanced automotive systems with a focus on energy and emissions.

Advances in Natural Sciences:
Nanoscience and Nanotechnology



REVIEW

Ovarian Cancer: Potential biomarkers and nanotechnology based diagnostic tools

To cite this article: Shruti Sounkaria and Prakash Chandra 2021 *Adv. Nat. Sci: Nanosci. Nanotechnol.* **12** 033001

View the article online for updates and enhancements.

You may also like

- Cytotoxicity and cellular uptake of different sized gold nanoparticles in ovarian cancer cells
Dhiraj Kumar, Isha Mutreja, Kenny Chitcholtan et al.
- Three-dimensional texture analysis of optical coherence tomography images of ovarian tissue
Travis W Sawyer, Swati Chandra, Photini F S Rice et al.
- Epithelial-to-mesenchymal transition: lessons from development, insights into cancer and the potential of EMT-subtype based therapeutic intervention
Jane Antony, Jean Paul Thiery and Ruby Yun-Ju Huang

This content was downloaded from IP address 14.139.251.100 on 13/12/2021 at 05:42

IOP Publishing | Vietnam Academy of Science and Technology Advances in Natural Sciences: Nanoscience and Nanotechnology

Adv. Nat. Sci.: Nanosci. Nanotechnol. **12** (2021) 033001 (15pp) https://doi.org/10.1088/2043-6262/ac2741

Review

Ovarian Cancer: Potential biomarkers and nanotechnology based diagnostic tools

Shruti Sounkaria and Prakash Chandra

Department of Biotechnology, Delhi Technological University, New Delhi-110042, India

E-mail: prakashchandra@dtu.ac.in

Received 8 May 2021
Accepted for publication 30 June 2021
Published 4 October 2021



Abstract

Due to lack of effective diagnosis, most of the women are diagnosed for ovarian cancer in late stages, which reduces their survival chances. Hence, there is a dire need for development of effective biomarkers and diagnostic strategies that can help in predicting the disease in early stages. Such developments can provide the biomarkers which have the capacity to provide effective results in limited time. They can direct the treatment strategies thereby revolutionising the way ovarian cancer is dealt worldwide. The article focuses on the current diagnostic and prognostic methods available and the potential biomarkers for ovarian cancer detection. These markers are currently being studied and validated. Some of the recent nanotechnological approaches that can be used to develop diagnostic platforms are discussed, which may represent the next generation of ovarian cancer diagnostics and prognostics.

Keywords: ovarian cancer, biomarkers, diagnosis, prognosis, nanotechnology

1. Introduction

An uncontrolled growth of cells characterised by self-sufficient proliferation, evasion of immune system and apoptosis, resistance to anti-inflammatory signals, angiogenesis, invasion of tissues and metastatic spread is called cancer [1]. Every year many people die from cancers of different types like lungs, prostate, ovary, breast, blood, etc A few of them get diagnosed at early stages and survive the cancer but a lot of them get diagnosed at advanced stages, fight a hard battle with cancer and sometimes succumb to this disease. The treatment methods like radiation and chemotherapy have hugely helped in increasing the survival rate in past many years and the growing research has also revolutionised the field of cancer diagnosis and treatment [2].

A deadly gynecological disease that many women are battling today is ovarian cancer. American Cancer Society says that ovarian cancer has the highest case to fatality ratio and 69% people with this cancer will usually face death. Approximately 125,000 women are killed each year worldwide by the ovarian cancer. Early detection can provide a survival rate of 90 % but sadly the detection is seen in a mere 20% at such early stages due to lack of visible symptoms and knowledge [3].

The most common histological type of ovarian cancer is epithelial ovarian cancer accounting for 90% of all cases. The epithelial cancer can be subdivided into four types which are serous, endometrioid, clear- cell and mucinous cancers [4]. Ovarian carcinomas can originate from either the ovary or from sites like fallopian tube and peritoneal cavity lined with mesothelium. A well differentiated ovarian carcinoma belongs to type I category and the high-grade serous carcinoma that metastasises rapidly belongs to type II category [5]. The process of ovarian tumourigenesis involves the epithelial to mesenchymal transition caused by interaction between adherins and proteases. The late tumourigenesis involves the growth of tumour nodules on surfaces covered with mesothelium. The metastasis usually takes place when the cells from primary tumour detach and spread in the abdominal cavity via peritoneal fluid. Epidemiological data has revealed that the consumption of oral contraceptives and pregnancy can lower the risk of ovarian cancer while the conditions like endometriosis and polycystic ovarian syndrome can increase the risk of same [6]. Although smoking and infection with

 © 2021 Vietnam Academy of Science & Technology

human papilloma virus have been linked to the epithelial ovarian cancer, the germline genetic mutations also play an important role in this cancer. Along with inflammatory causes like endometriosis, pelvic inflammatory disease and hyperthyroidism can also lead to formation of ovarian cancer [7]. Continual ovulation and increased levels of gonadotropins and estrogen can also stimulate ovarian cancer. Such ovulation can be traumatic around the ovarian surface leading to neoplasm formation. Certain environmental factors like use of talcum powder in genital area, smoking, alcohol consumption, intake of red meat and caffeine and conditions like obesity can also cause ovarian cancer [7, 8]. Familial history, hereditary and genetic factors are also associated with ovarian carcinoma with an almost 10% of ovarian cancers being hereditary. A mutation in BRCA1/2 high penetrance genes also enhances the risk of epithelial ovarian cancers [9].

Ovarian cancer is known as a silent killer because it normally has no symptoms until it is advanced. Patients with epithelial ovarian cancer are typically diagnosed in stages 3 or 4, which are advanced stages. A consensus in 2007 stated that women with bloating, abdominal pain, difficulty in eating of feeling full and urinary symptoms daily for many weeks should consult their doctor' [10]. The diagnostic or screening strategies currently in practice are pelvic examination, transvaginal and abdominal ultrasonography and laparoscopic studies. Moreover, the pelvic examination gives a sensitivity and specificity of only 45% and 90% in distinguishing between a premalignant lesion from a healthy ovary [10]. Another screening method is the transvaginal sonography or TVS that detects changes in ovarian volume and morphology. According to published literature, the ovarian volume of more than 20 cubic centimeters in premenopausal women and more than 10 cubic centimeters in post-menopausal women is usually considered as abnormal. Presence of papillary projections in the cyst of ovary is considered a morphological abnormality in the TVS screening. TVS screening method is reliable in detecting the small abnormalities in volume and morphology of ovary but it has still not proven its accuracy in differentiating a benign ovarian tumour from a malignant one which is an integral piece of information to move forward in the treatment process [11]. TVS imaging is also an invasive approach and does not necessarily improve the chances of living for OC patients.

The less reliable, invasive and late diagnostic or screening methods have increased the need to identify biomarkers which would be more sensitive and non-invasive diagnostic tool. The biomarker should be able to diagnose all kinds of asymptomatic people who can probably develop ovarian cancer during their lifetime [12]. The diagnosis at earlier stages of ovarian cancer can predict the outcome of treatment and increase the survival rates which are quite low when the cancer is diagnosed at advanced stages. Development of reliable and effective biomarkers can help in disease detection in stage I tumour patients and thereby improving the 5-year survival rate by 95% [13]. Biomarkers can help in the screening of a disease at much earlier and curable stage, which is quite beneficial in the case of ovarian cancer since it is a major health problem, prevalent in the population and can be cure if diagnosed during period [12]. Determination of prognosis, disease relapse, reaction to treatment and efficacy of a therapy can also be tested by an efficient biomarker [14].

A major scientific advancement that has enabled point-of-care diagnosis is the development of biosensors. Nanobiosensing is an integration of biosensing and nanotechnology which has today become extremely significant in detection of ovarian cancer. A biosensor analyses a biomarker's concentration or presence and converts it into a signal generated electrically which can be read. The result generation is fast, easy and just a-few-step process [15]. Certain metabolites, miRNA's and circulating tumour cells have become vital in cancer diagnosis, but due to limitations of the conventional diagnostic strategies, the focus of research has shifted to nano-biosensors. These tools are robust, efficient, easy-to-use, specific and have an ability to provide ultrasensitive detection. The different nanomaterials employed in such tools include graphene, carbon nanotubes, conducting polymers, etc and molecules like an antibody, an aptamer and an enzyme are usually used as receptors. Other factors like reproducibility, portability and cost effective nature of biosensors have increased their interests in commercial setups [16, 17].

2. Biomarkers for ovarian cancer

Any biological molecule found in body fluids or tissue that can differentiate a normal person from an affected one is called a biomarker. It signifies an abnormality or a diseased condition and can detect malignancy, if any in a healthy person. The biomarkers can be categorised into proteins, nucleic acids, antibodies, peptides or genetic, proteomic and metabolomic signatures. Biomarkers can be considered as variables that are quantifiable and may or may not be correlated with a person's state of health and wellbeing [18]. The assessment is usually easy and non-invasive but can sometimes require biopsy [14].

The potential uses of biomarkers which have been established make their development in clinical setting rather necessary. For timely diagnosis and for a clinician to make decision regarding patient care, accurate point-of-care diagnostic markers are required. These markers are robust and non-invasive as they can be found in sera or urine of patients. Such markers are usually involved in the signaling cascade of progression of ovarian tumours and can be the byproducts of such pathways. The research and development in these sorts of markers will increase detection, treatment and probably enhance the chances of a disease-free survival [3]. The developing genomics and proteomics technology have helped in identification of many DNA, RNA and protein biomarkers. DNA biomarkers include circulating DNA and tumour cells in serum and blood respectively that can dictate or represent the metastatic spread of cancer [19, 20]. For example, a mutation in TP53 gene and its inheritance are known to be a cause of many cancers. Similarly, mutations in oncogenes like RAS and tumour-suppressor genes like APC and RB1 can also be probable markers for cancer prognosis and treatment

Table 1. Ovarian cancer biomarkers, their type, source, sensitivity and specificity.

Biomarker	Type	Source	Sensitivity	Specificity	References
CA-125	Glycoprotein	Serum	50%–60%	90%	[3, 21]
HE-4	Protein	Serum	72.9%	95%	[22]
MMPs	Endopeptidases	Urine	82%	75%	[23]
β_2-MICROGLOBULIN	Protein	Serum	87%	48%	[3]
M-CSF [Macrophage colony stimulating factor]	Glycoprotein	Serum	61%	92.7%	[3]

selection. Most of the biomarkers that are approved by FDA and are clinically in use are proteins in nature and usually extracted from serum. The protein expression can reflect the prognosis as in case of breast and colon cancers, the expression of HER2/NEU and EGFR can predict their response to treatment by trastuzumab and imatinib respectively [19].

An ovarian carcinoma can remain undiagnosed due to asymptomatic early stages and also due to different sources of origin like stromal cells, germ cells and epithelial cells which can increase the diagnostic complications [3]. To counter this lag in healthcare arena, many biomarkers have since been established so that the diagnosis and prognosis of ovarian cancer can become easier. Some of the biomarkers with the level of sensitivity and specificity they show are listed in table 1.

2.1. CA125

One of the most used tumour biomarkers for ovarian cancer is antigenic protein CA125. It increases in 80%–90% of patients with EOC. Its specificity stands at 90% for early postmenopausal women [21]. The antigen is a glycoprotein in nature and has three domains. The second domain is extracellular domain and is recognised by the monoclonal antibodies and the cytoplasmic domain is involved in the signal transduction [2, 3]. Currently it is the only biomarker which is effectively used as a measure of prognosis and disease stabilisation in ovarian cancer patients. The discovery of CA125 as an antigen was made possible by the monoclonal technology where antibodies were developed to determine treatment for ovarian cancer. In the antigen-antibody interaction, it was observed that there were high levels of circulating CA 125 in ovarian cancer cells [24]. The gene coding for CA125 is MUC16 which was cloned by Ken Lloyd and Tim O'Brian. CA125 has been a serum biomarker for ovarian cancer detection for many years, to assess the response of patient to chemotherapy, to differentiate between pelvic benign and malignant masses, for designing clinical trials and to check the disease recurrences [1, 3]. The factors which limit the sensitivity of this marker are endometriosis and cirrhosis [25]. Therefore, another marker called HE4 is used in combination with CA125 to assess the probability of development of ovarian malignancy.

2.2. HE4

The HE4 biomarker in combination with CA125 is a potential biomarker for EOC and is hence the only one to be approved by the FDA. It is also clinically available to assess the progression of the cancer [25]. Most of the malignant epithelial ovarian tumours are of serous and endometrioid types and an overexpression of HE4 protein is observed in them [26]. The sensitivity and specificity of HE4 alone are 72.9% and 95% whereas when it is used in conjunction with CA125 as dual marker the sensitivity increased to 76.4% [21]. Structurally HE4 has 2 domains of whey acidic protein and four disulfide cores which are made up of 8 cysteine residues. The HE4 gene secretes an N-glycosylated protein product in the bloodstream which can then be detected. In prospective case studies done on patients with recurrent epithelial ovarian cancer, it was observed that HE4 is a potential biomarker and can predict the recurrence better than CA125 [22].

2.3. AFP

Belonging to albuminoid family of proteins, alpha-fetoprotein is a multifunctional glycoprotein which can help in diagnosis of cancers and their growth. Many commercial diagnostic tests can detect an increase level of AFP in patients with cancer. In patients with liver cancer, AFP is raised by 70% [27]. It is usually a marker for non-epithelial ovarian malignancy.

2.4. MMPs

MMPs are zinc dependent enzymes which are involved in cleavage of protein substrates. They regulate the extracellular matrix by doing so and also secrete proteins like cytokines, chemokines and growth factors from their platforms which are anchored to the membrane [28]. They are thus endopeptidases and consist of more than 20 of them, among which MMP-9 is the one which is majorly involved in cancer progression. The main processes of carcinogenesis like migration, metastasis, invasion and angiogenesis are associated with extracellular environment and MMPs main function is ECM remodelling. Figure 1 shows how MMP-9 and MMP-2 are involved in the development of a Cancer. Therefore, MMP-9 inhibitors are being developed for many anti-cancer therapies. MMP-9 degrades the type IV collagen protein which is a component of nearly all basement membranes in the cell. By degradation MMP-9 allows the vascular and stromal invasion by the tumour cells. This can be a primary reason of choosing MMP-9 as a therapeutic target [29].

MMP-9 can be considered as a potential biomarker owing to its sensitivity and specificity when used in combination with other biomarkers. MMP-9 can be detected in the extracellular vesicles ascites of high grade serous ovarian

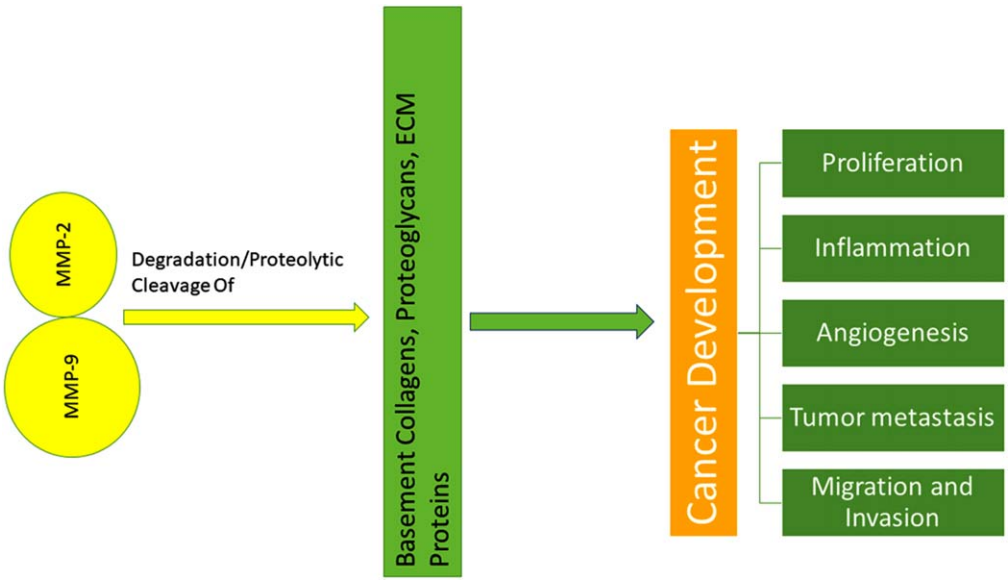


Figure 1. The two major MMPs, MMP-9 and MMP-2 are involved in degradation or proteolytic cleavage of basement collagens like type IV collagen, ECM proteins like fibronectin and proteoglycans. By cleaving them in extracellular space and degrading the extra tissue structure, MMPs can release more of the factors associated with angiogenesis, thus allowing the already expanding tumour to increase angiogenesis [30–32].

cancer at high levels [32]. Along with MMP-9, MMP-2 was also expressed in ovarian malignant tumours. The levels of both of them were although different from those found in benign tumours and normal ovarian tissues. In a statistical analysis serum levels of MMP-9 were high in patients with ovarian cancer and those having a relapse. These levels usually decreased after surgery proving that MMP-9 can be useful biomarker. Another meta-analysis found an increased expression of MMP-9 in ovarian cancer than in non-malignant tumours. Thus, an overexpression of MMP-9 can help clinicians in differentiating between tumours, assessing the tumour stage and the metastatic ability [29]. In ovarian cancer cell lines, the role of MMP-9 in invasion and adhesion was investigated using RNAi. This also suggests a promising strategy for reducing cancer invasiveness by targeting MMP-9 with gene therapy [33]. MMP-2 expression was also responsible for poor prognosis in epithelial ovarian cancer [34]. Both MMP-9 and MMP-2 are extensively studied in association with cancer though there is differential expression for both of them with MMP2 higher in benign tumours and MMP9 higher in malignant tumours [35].

Other MMPS like 1, 3, 7,8,13, 14 and 15 also promote tumour development and advancement in ovarian cancer. By a research performed on patients with late stages of ovarian cancer by detecting MMPs in urine, 82% sensitivity and 75% specificity were achieved [23].

2.5. HSP70

Proteomic technology to explore new biomarkers for cancer is rapidly developing and provides great avenues for assessing the tumour microenvironment. It helps in detection of new prognostic and diagnostic markers by characterising all the proteins in the biological system of malignant cells in their different forms. HSP70 comes into action for cell survival when it is stressed, which is why its expression is higher in tumours. It might be involved in tumour development, reduced survival, proliferation of cancer cells and advancement in cancer stage. Because the ovarian cancer sera showed stronger HSP70 autoantibody response than normal sera, it can potentially serve as a diagnostic biomarker in coming times [36]. HSP70 promotes cancer development because it has anti-apoptotic properties and also provides the cancer cell resistance to chemotherapeutic drugs. Membrane bound HSP70 has been detected in exosomes and has opened an avenue for them being researched as a cancer biomarker. They are found in more quantity as compared to circulating tumour cells and are usually translocated to membrane, which makes them easily detectable in exosomal membranes. It is believed that HSP70 is secreted in large amount in blood of patients with different cancers like breast, ovarian and non-small cell lung cancers. By detecting the amount of circulating exosome with membrane bound HSP70, it can establish whether it correlates with cancer progression or regression. Clinically, this can help in directing the therapeutic regime for patients [37].

2.6. Prostasin

The probable use of prostasin, a serine protease present in the prostate gland and secreted in seminal fluid as an ovarian cancer biomarker was investigated. In a small subset, researchers discovered elevated levels of prostasin in the blood of ovarian cancer patients as compared to controls. This study was possible due to the microarray technology which was used to generate complementary DNA sequences of the RNA isolated from ovarian cancer and normal human ovarian surface epithelial cell lines. Anti-prostasin antibodies were

used to test the serum levels of prostasin after the overexpressed gene was detected. Many other serine proteases could also be elevated in ovarian cancer patients, according to research. While the sample size was insufficient to show that prostasin is a good biomarker, when combined with CA125, the sensitivity and specificity increased [38].

2.7. Creatine kinase B

Another gene which was found to be overexpressed in ovarian tumours was creatine kinase B [CKB]. This was also reported using the micro array technology. CKB enzyme expression was greater in ovarian cancer patients' sera, particularly in those with level I disease [3]. CKB is a protein found as a dimer in the tissues and serum. It is involved in the rejuvenation of ATP and is therefore found in skeletal muscles, myocardium and smooth muscles as they have high energy concentration cells. Certain factors while keeping in mind to consider CKB as a biomarker are age and muscle mass [39].

2.8. Mesothelin

Mesothelin is a glycoprotein which is expressed on the cell surfaces of mesothelial cells that usually line the pleura, peritoneum and pericardium. The 40 kDa glycoprotein is linked to glycosylphosphatidylinositol and is a differentiation antigen present abundantly in mesothelial cells. It is anchored to the membrane of cells and is found in the body fluids of tumour patients. MSLN can be considered as a tumour marker because of its high expression in 70% of ovarian cancers and also due to its role in progression of tumour and chemoresistance [40, 41]. According to many studies it has been clear that the expression and activity of proteolytic enzymes MMPs [matrix metalloproteinases] are enhanced in many cancers. In a study it has been observed that the overexpression of MSLN caused rapid movement and invasiveness of cancer done by the induction of MMPs. The ERK, Akt and JNK signalling pathways were involved in this mechanism [42]. The activation of MMP-7 by MSLN is responsible for tumour progression in ovarian cancers. This study also shows a promising role of specific and sensitive inhibitors of ERK and JNK pathways in halting the tumour progression. Since mesothelin was secreted by the affected peritoneum cells into the circulation, the elevated levels can also be indicative of disease severity and tumour progression in case of ovarian cancer patients [40].

Since MSLN is heavily expressed in many human cancers and is observable in the sera of ovarian cancer patients, many anti-cancer and immune therapies are targeting mesothelin. Its role in metastasis is still unclear but many suspect that it is involved in adhesion mechanism. The interaction of MSLN with the glycoprotein MUC16 can lead to development of malignant phenotype and also the metastasis of ovarian cancer [43].

2.9. Human epidermal growth factor 2 [HER2]

HER2 is a tyrosine receptor kinase having 3 domains which are an extracellular binding domain, a transmembrane domain and an intracellular domain. They are mainly involved in cellular growth, proliferation and survival. In a study it has been calculated that HER is highly expressed in almost 60%–98% of epithelial ovarian cancers [44]. In a trial study performed by Slamon et al., the expression of HER2 was correlated with ovarian cancer and its role as a prognostic biomarker in tumour progression [45]. The HER2 is believed to dimerise with HER3 receptor to activate certain downstream signaling pathways. The end result of such dimerisation causes a mitogenic effect. Currently there is less clarity on the fact whether HER2 expression is associated with the poor prognosis of ovarian cancer or there is no correlation between the two.

2.10. MCF

Macrophage colony stimulating factor is another biomarker whose expression is indicative of poor prognosis and diagnosis of many cancer types including the ovarian cancer. The MCF growth factor helps in proliferation of monocytes into macrophages and they also secrete certain cytokines and proteases during inflammatory reactions [46]. It is a serum biomarker with 61 % sensitivity and a specificity of around 93%. When tested on both normal and ovarian cancer patients using ELISA and MEIA, it was observed that MCF can be useful in diagnosing OC clinically and better results could be obtained if used in combination with CA125. The diagnostic power also increased when the expression of other cytokines like MMPs and TIMPs was checked in combination [47].

2.11. VEGF

Angiogenesis is a major feature of tumour development and invasiveness and many investigations have revealed that vascularisation can predict the treatment outcomes in cancer patients. Most recently, the vascular endothelial growth factor has been found to be a potent regulator of angiogenesis in many model systems [48]. Many experimental setups have shown the involvement of VEGF in ovarian carcinogenesis. In many patients at early stages of ovarian cancer VEGF was reported to be a prognostic marker as higher serum levels were associated with lower survival rates [13]. High levels of serum VEGF are also indicative of remaining tumour and its metastatic potential after a primary cytoreductive surgery and thus can help clinicians in making a decision between neoadjuvant therapy and cytoreductive surgery [49]. An association has also been found between serum VEGF and melanoma, colon cancer, small cell lung cancer and esophageal cancer prognosis. In many ovarian carcinoma cell lines VEGF receptor overexpression has been detected in the tumour cells, thus promoting their growth, proliferation, survival and migration. In an ELISA study VEGF expression was studied in serum of ovarian cancer patients, in those with benign ovarian tumour and in healthy controls. According to the findings, serum VEGF levels in ovarian cancer patients

were significantly higher than in benign tumour and healthy monitors [50]. As a biomarker for benign tumours, it was less accurate. The implantation of tumour cells in the peritoneum and the deposition of ascites are two characteristics of ovarian cancer. Clinically, there is evidence that by increasing vascular permeability of micro vessels in the peritoneal walls, VEGF might be inducing tumour growth in that region [51]. Moreover, stimulation of macrophage and monocyte infiltration by VEGF overexpression has also been noticed in OVCAR-3 tumour cells. By acting as a chemotactic agent VEGF might attract monocytes from peripheral circulation and lead to their accumulation in tumour cells [52]. As a highly specific endothelial mitogen, VEGF can cause metastasis of malignant ovarian carcinomas, increase in ascitic fluid volume and advancement in tumour stages. A regular checkup to detect the serum VEGF levels can thereby help in detection of tumours and their metastatic potential, which can enhance the survival chances [50]. The two cell surface receptors which promote the activity of VEGF are flt-1 and KDR. A dominant negative mutation in the KDR receptor blocked the downstream signaling of VEGF, thus disrupting its activity. In many animal models, tumour growth and vascular permeability were inhibited by antibodies blocking the function of VEGF or VEGF silencing by antisense mRNA [51]. This can also be considered as a therapeutic option since anti-VEGF therapy can be performed. Bevacizumab which is a monoclonal antibody can be used to target VEGF when it is overexpressed in cancer patients, such as those with primary ovarian tumours that have the potential to spread. Bevacizumab can reduce tumour vasculature, thus killing the endothelial cells within them and also reducing the interstitial fluid. This will also help the chemotherapeutic drugs to reach their target locations. As a result, the extent of VEGF gene expression in EOC patients with residual tumours after the initial surgery could theoretically guide the use of bevacizumab [53].

2.12. FOLR1

Many cancer therapies target the folate receptor alpha which is expressed frequently in ovarian cancers and is responsible for transport of folate into the cells. The receptor occupies the position 11 q on the chromosome. The high protein expression of FOLR1 can be correlated with platinum therapy resistance, progression of tumour and low prognosis. It can also help in predicting the response of an ovarian cancer patient to the therapy and chances of survival. In a study conducted on patients with certain residual tumour after they had their surgery, it was observed that the survival rate was only 30% at 3 years for those with high FOLR1 as compared to 62% in those with low FOLR1 levels [54]. A bioinformatic study revealed that there was higher expression of FOL1 mRNA in tumour tissues of ovarian carcinomas than in the normal tissues, which might signify their role in development of tumour [55].

2.13. IL-8

Another potent and proangiogenic cytokine which is overexpressed in ovarian carcinomas is IL-8. The overexpression is associated with high tumour grade, poor survival rates and clinical outcomes in patients. IL-8 accelerates ovarian cancer metastasis and its abnormal and increased expression can be observed throughout the disease course. Therefore, IL-8 can be considered as a biomarker for predicting treatment response [44]. IL-8 was also found out to be one of the most expressed chemokines in ovarian carcinomas along with GRO-alpha and CXCL-7. They promote metastasis and cancer aggressiveness mainly through NFkB signaling pathway [56]. When samples of high grade ovarian serous cancer tissue were studied, it was observed that there is higher tumoural expression of IL-8 compared in benign neoplasms. High expression of IL-8 was correlated with poor prognosis, thus demonstrating the value of IL-8 as a prognostic biomarker. The response to anti-cancer drugs was strengthened when IL-8 signaling was blocked. Exogenous IL-8 facilitated cancer cell migration, while reparixin, its receptor, had the opposite effect. The epithelial-mesenchymal transition, which was accomplished by decreased expression of E-cadherins, was the main mechanism by which IL-8 caused tumour metastasis [57]. Serum expression levels of IL-8 can be considered as prognostic marker for knowing the stage of cancer [58]. Several other cytokines like IFN-gamma, TNF-alpha, CSF-1 and TGF-beta1 also caused poor prognosis in epithelial ovarian cancer according to a study but the evidence was limited to prove them as efficient prognostic biomarkers [59].

2.14. Protein Z

Another proteomic biomarker which can be utilised in the early detection of ovarian cancer is protein Z [60]. It is a serine protease and has the ability to differentiate between ovarian cancer samples from normal or control samples. Protein Z could also diagnose cancer cases which are at stage 1 or 2 and can be invasive. The elevated levels were indicative for patient to be at the risk of epithelial ovarian cancer [61].

2.15. Exosomes

Exosomes are extracellular vesicles that range from 30 to 150 nanometers in diameter. After the fusion of late endosomes or multivesicular bodies with the plasma membrane, these are one of three extracellular vesicles produced [12, 62]. Exosomes were thought to be produced by cells only to get rid of excess material, but current research is focusing on understanding how they function as promoters of metastasis. It has been reported that exosomes constitute bioactive molecules like nucleic acids, proteins, enzymes and lipids which can modify the functions of the cell receiving them. Exosomes are involved in tumour progression, angiogenesis, immune system modulation and in formation of a pre-metastatic niche [63]. The biogenesis of exosomes and their diagnostic and prognostic applications are shown in figure 2. These functions were observed by isolation of exosomes from tumours and therefore these circulating extracellular vesicles are now being studied to monitor the progression of a cancer. Different methods for isolation of exosomes have been highlighted in figure 3. These EVs after being released can be prominently

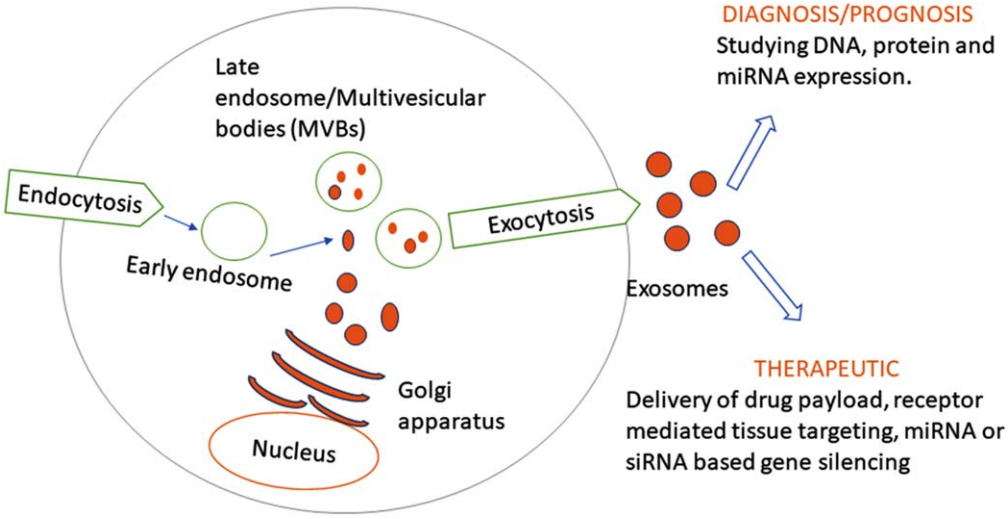


Figure 2. Biogenesis of exosomes. The development of intracellular multivesicular bodies is caused because of the double invagination of the plasma membrane. The first invagination produces an early endosome containing cell surface proteins. They develop into late endosomes after interacting with the trans-Golgi and endoplasmic reticulum networks, forming MVBs by another invagination of the plasma membrane. Intraluminal vesicles [ILVs] are found in MVBs and are released as exosomes after merging with the plasma membrane. After proper separation, these exosomes may be used for therapeutic and diagnostic purposes [64].

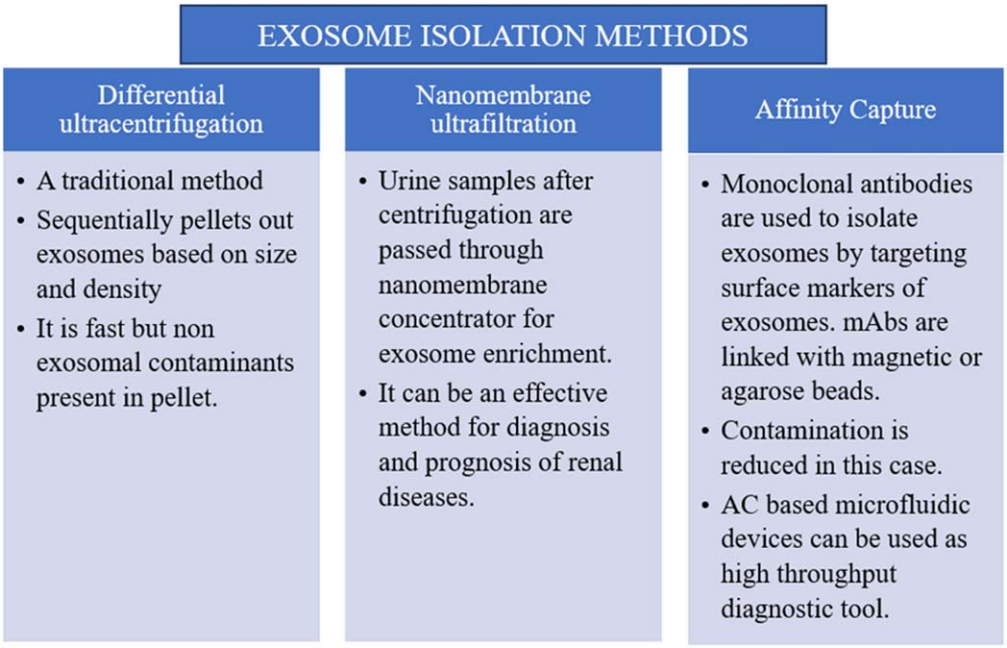


Figure 3. Exosome isolation methods. Differential ultracentrifugation, nanomembrane ultrafiltration and affinity capture [65, 66].

found in body fluids like urine, sweat, saliva, blood and tears and hence can be isolated from there. The molecular cargo present in the exosome is indicative of its parent environment and such cargo can hence be representative of the tumour microenvironment also when isolated from cancer cell lines [12].

In different studies it has been seen that certain known cancer associated miRNAs were present in tumour derived exosomes. A tumor cell derived exosome with its surface signatures and cargo is depicted in figure 4. The miRNA signature in cancer cell lines for ovarian, lung cancer, pancreatic cancer among many others has been of prognostic importance. The small non protein coding miRNAs regulate the post transcriptional gene expression by specific suppression of a target RNA. According to various expression profile studies, it has been reported that there is a linkage between cancer prognosis and specific miRNA expression. The miRNAs are protected from cellular environment within the exosomes and their levels are dysregulated in circulating EVs isolated from patients with ovarian cancer [12].The miRNA is linked to cancer since they promote oncogenesis by either acting as tumour suppressors or as oncogenes. miR-15a and miR-155 are known to act as tumour suppressor and oncogene respectively [67]. By imperfectly pairing with the target

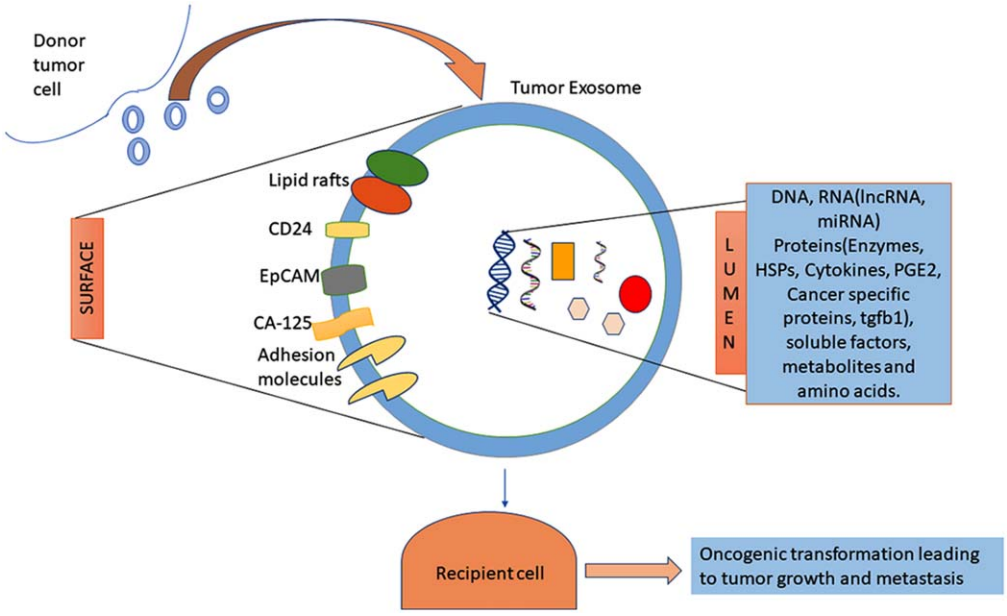


Figure 4. Tumour cell derived exosomes. Exosomes are released from a stressed tumour cell into tumour microenvironment. Such exosomes carry nucleic acids, proteins, soluble factors, metabolites and amino acids in their lumen. The ovarian cancer specific proteins include heat shock proteins like HSP70 and HSP90, TGFbeta1, MAGE3/6 and various cytokines. The surface protein biomarkers include CD24, EpCAM and other adhesion molecules. Lipids rafts are present in the exosomal membrane along with cytoskeletal proteins. On uptake of these exosomes by recipient cells, such cells can undergo oncogenic transformation which further enhances the tumour growth and spread of the cancer [65, 70].

mRNA of protein coding genes at 3' untranslated region, the non-coding RNAs can suppress the translation of target mRNAs. In many expression profile studies of ovarian cancer cell lines, aberrant expression of miRNA was reported and also 46% of micro RNAs were dysregulated both in cellular and exosomal compartments. miRNAs are also known to promote resistance to chemotherapy in ovarian cancer patients, thus making them a significant biomarker to track the response of patients to treatment [68]. Since the exosomes are involved in cell-cell communication, the exosomes from tumour cells can pass on their oncogenic activity to the normal cells which can be via delivery of the tumour-miRNAs [69].

3. Current isolation and detection methods for biomarkers

Double determinant immunoassays are used for isolation and detection of CA125, the most reliable biomarker in ovarian cancer diagnosis. The two antibodies used in this assay are OC125 and M11. The other isolation technique used is the antibody-based affinity chromatography method [71]. As discussed in previous section of biomarkers, certain miRNAs and other cancer specific biomarkers are either packaged in the tumour-secreted exosomes or found on their surface. Their isolation methods like differential ultracentrifugation, filtration and immune-based affinity capture were also highlighted in figure 3 in the same section. A tumour tissue biopsy is an invasive method for isolation of biomarkers which requires a proper clinical investigation of the area, accurate incision, sample handling and interpretation of results. A larger sample can help in studying tumour biomarkers better due to non-homogenous environment of the tumour [72]. Table 2 lists different technical challenges faced in isolation of biomarkers and the existing methodologies that can overcome them.

A clinically significant biomarker that has an ability to provide minimally invasive diagnosis for ovarian cancer is small, non-coding RNA i.e., mi-RNA. The reports of nucleic acids and miRNA- based alterations related to cancer development have led to growing interest in their application as biomarkers. Some of the mi-RNA detection methods are northern blotting, miRNA microarray, RNA sequencing and quantitative real time PCR. A northern blotting technique first separates the small RNA on a gel electrophoresis which is then allowed to hybridise with a detection probe after being transferred on a membrane. A signal is produced when the probe hybridises with the target miRNA. Working on similar grounds, a miRNA microarray focusses on hybridisation of labelled miRNA with oligonucleotide probes which then provides quantitative information on miRNA based on the fluorescent signals. The techniques present challenges like low sensitivity, lack of robustness and complex instrumentation [77]. The most commonly used method of detection for exosome and other biomarkers is ELISA. In this method, the isolated exosomes are immobilised in the microwell plates using blocking agents. The recognition or capture antibody is then used to detect surface antigens on exosomes followed by linking of an enzyme labelled secondary antibody to capture antibody. Addition of colorimetric substrate to the medium provides an estimation of the biomarker. The major challenge faced by ELISA is non-specific adsorption of biomolecules present in the sample fluids. Other

Table 2. Challenges faced in biomarker isolation and alternate approaches.

Isolation methods	Biomarker	Challenges	Alternate approaches	References
Double determinant immunoassay	CA125	Non-uniform and non-specific epitope binding.	A bottom-up proteomics approach in which the ascites sample first undergoes an enrichment process by filtration and chromatography followed by mass spectrometry	[71]
Antibody-based affinity chromatography method	CA125	Underassessment of CA125 due to non-accessible epitopes on proteoforms.		[71]
Differential ultracentrifugation	Exosomes	Time-consuming, expensive instrumentation and possible contamination.	• Multiple resuspensions in a buffer. • Using a sucrose density gradient for avoiding aggregation.	[73]
Filtration	Exosomes	Loss or damage to exosome vesicles due to their entrapment in the filters or due to the force applied for filtration.	• Centrifugation to overcome the damage done by force application. • Using membranes with less affinity for exosome protein binding.	[73]
Immune-based affinity capture	Exosomes, HER2	• Not applicable to samples present in large volumes. • Expensive combinations of antigen and antibody. • Poor yield. • Isolation possible only from cell-free samples.	An alternating current electrokinetic microarray chip that bypasses the cell free isolation step and provides an on-chip isolation and detection of exosome and its cargo.	[74]
Microfluidics lab-on-chip	Exosomes	• Unautomated • Manual sample preparations	A fully automatic microfluidic bases point-of-care device is required.	[73]
Tumour tissue biopsy	Circulating biomarkers like CTCs, exosomes, miRNAs and lnRNAs,	• Invasive • Inadequate amount of tissue cannot represent tumour heterogeneity clearly. • Preservation methods can give false positives.	A liquid biopsy is a minimally invasive isolation method having the following benefits • Repeated sample collection is feasible. • Accessible biomarkers. • A clear picture of tumour, genomic alterations. • Possibility of personalised treatment.	[75, 76]

Table 3. Analytical performance of some nanotechnology based diagnostic platforms.

Diagnostic platform	Method	Nanomaterial	Target biomarker	LOD	Linear range	Sensitivity	References
Immunosensor	Electrochemical [SWV]	3DrGO-MWCNT – PAMAM/AuNPs Suc-CS@MNP-TB	CA125	6 μU ml^{-1}	0.0005–10 and 10–75 U ml^{-1}	High	[87]
Immunosensor	Electrochemical	PANI/Gr-SPE	CA125	0.923 ng μl^{-1}	0.92 pg μl^{-1}– 15 .20 ng μl^{-1}	Medium	[80]
Immunosensor	Optical [chemiluminescence]	GQD	CA125	0.05 U ml^{-1}	0.1 U ml^{-1} to 600 U ml^{-1}	—	[81]
Hybridisation chain reaction based sensor	Electrochemical [Chronocoulometry]	Au Wire	miRNA 21	11 pM	30 pM to 7 nM	High	[88]
Nucleic acid based	Electrochemical	Au@MNPs	mi-RNA 21	1.0 pM	10 aM-1nM	High	[89]
Genosensor	Electrochemical [EIS]	Cysteine–ZnS QD	miR-200a	1.0×10^{-14} to 1.0×10^{-6} M	8.4 fM	-	[90]
Exonuclease III-assisted target recycling and HCR	Photoelectrochemical	$AgVO_3$	VEGF	10 fM to 10 nM	3 fM	High	[91]
Aptasensor	Optical	Lanthanide-doped upconverting nanoparticles	HE 4	0.4 ng ml^{-1} to 7.0 ng ml^{-1}	0.021 ng ml^{-1}	High	[92]
Immunosensor	Electrochemical	CdSeQDs	Exosomes	—	—	100 exosomes/microlitres	[93]
Immunoanalyser	Programmable microfluidic based bio-nanochip	Agarose beads	CA125, HE4, MMP-7, and CA72–4	—	—	68.7%	[94]
Immunoanalyser	Microfluidics Chip	Graphene oxide	CA125, HE4, AFP, CA153, OPN, MSLN HSP70, IL6 and IL8	0.01 U ml^{-1}	1–103 pg ml^{-1} [HE4], 0.1–100 U ml^{-1} [CA125, CA153]	High	[84]

Abbreviations:
SWV: square wave voltammetry, EIS: electrochemical impedance spectroscopy, 3DrGO: three-dimensional reduced graphene oxide, MWCNT: multiwall carbon nanotube, PAMAM: polyamidoamine, Au: gold ,NPs: nanoparticles, Suc-CS: succinyl-chitosan, MNP: magnetic nanoparticles, TB: toluidine blue, PANI: polyaniline, Gr-SPE: graphene-screen printed electrode, GQD: graphene quantum dots ZnS: zinc sulphide, QD: quantum dots, $AgVO_3$: silver vanadate, CdSe: cadmium selenide, CA: cancer antigen, HE 4: human epididymis protein 4, miRNA: microRNA, VEGF: vascular endothelial growth factor, MMP 7: matrix metalloproteinase-7, AFP: alpha-fetoprotein, OPN: osteopontin, MSLN: mesothelin, HSP70: heat shock protein 70, IL: interleukin.

methodology used in detection of ovarian cancer biomarkers is flow cytometry and more specifically FACS. In these methods, the fluorescence and the light scattered by every single exosome are recorded, amplified and digitalised in a readable form. Moreover, a much complex FACS method employs antibodies tagged with fluorescent dyes. The drawbacks of flow cytometry are complex and costly instrumentation along with inefficient capture of nanovesicles. Better optimisation, microfluidics approaches and electrochemical detection techniques are being used to address these challenges [73]. The demand for ultrasensitive, highly efficient and clinically viable diagnostic methods has increased the research interests in the field of nanotechnology. Nanodiagnostic platforms have an ability to provide high-performance, timely, cost-effective and sensitive detection. Such methods are further discussed in the next section. Most ovarian cancers have a 90% chance of getting cured, if detected at initial stages. An improved screening at early stages involving analysis of biomarkers CA125 and HE4 in combination with other markers like E-cadherin, IL-6 can provide better results [78].

4. Prospects of nanotechnology

The goal of this review was to compile a list of biomarkers that can be used for early detection and selective treatment of ovarian cancer. It is becoming increasingly evident that combining the traditional surgical and chemotherapeutic approaches with targeted and personalised therapy can enhance the survival rates in patients. They can also help in overcoming recurrence, relapse and resistance in patients. One of the recent approaches in both diagnosis and treatment of ovarian cancer has been the use of nanotechnology. Nanotechnology has enabled the use of point-of-care platforms to detect the biomarkers with high specificity and sensitivity at the earliest stage. A programmable bio-nanochip system can be used as a multi-marker analyser. It comprises of an immunoanalyser which helps to diagnose the ovarian cancer by detecting the combination of markers like CA 125, HE4, MMP-7 and CA72.4 with high specificity and low cross reactivity in the reaction. The analysis time is only 43 min and the screening can be done at very low cost [60].

Nanotechnology provides both innovation and miniaturisation in cancer diagnosis. The methods are cost effective, rapid and easy to perform. Nanosensors are most recent diagnostic tool that can convert the chemical information into an electric signal which can thereafter be assessed. Measurement of important ovarian cancer biomarkers like CA125, HE4 and MUC1 using these sensors can help in diagnosing the cancer at early stages, thus increasing the chances of survival and directing the treatment regimen afterwards. Some of the advantages of biosensors are real time and remote testing, portability, label free detection, use of small volume of samples and high specificity and sensitivity in detection [79].

In an experiment by Gazze et al., a graphene based electrochemical biosensor was developed on which an anti-CA125 antibody was conjugated. The sensor had a lower limit of detection of 0.923 ng ul^{-1} which is lower than traditional sensing techniques and it also showed a good sensitivity to detect the OC at earlier stages [80].

An optical biosensor that works on the principle of chemiluminescence was developed by Israel et al. in the form of an immunoassay to detect the presence of CA125. In this graphene, quantum dots were used as a nanosystem and chemiluminescence resonance energy transfer was principally employed. The sensor provided a simplistic and sensitive approach towards CA125 diagnosis by using graphene quantum dots as energy acceptors. They were immobilised on a glass chip and conjugated with a capture antibody which formed a complex with antigen CA125 when it is present in the immunoassay. The sensor had a low detection limit of 0.05 U ml^{-1}, a wide linear range and also there was no need of an external excitation source. This can prove to be high throughput device for multiplex detection in future for CA125 and even other such antigenic biomarkers [81]. Another approach used in cancer diagnosis is a nanozyme based biosensor. It provides high-performance and highly sensitive platforms and the sensing materials used are metal NPs, metal oxide NPs, carbon nanomaterials and graphene oxide. These nanozymes mimic the natural enzyme activities and can simultaneously provide analytical results based on their optical, magnetic and electrochemical properties [82].

There have also been flourishing interests in the utility of superparamagnetic nanostructure based optical and electrochemical biosensors. Such magnetic nanoparticles can be functionalised easily, thereby providing sensitive detection. Recently, nanoporous iron nanocubes loaded with gold were used for detection of miRNA. This was possible due to magnetic and electrochemical properties of the nanostructures [83].

The most popular and efficient miniaturised sensing system nowadays is microfluidic chip sensor which can integrate multiple biomarkers on a single chip, thus reducing the sampling time and rate and reagent volume. It can make screening much more effective and can provide a rapid point-of-care diagnostic device [79]. Yu Wu et al. in their research study developed a multi biomarker panel using microfluidics to diagnose the presence of 9 biomarkers in the sera of ovarian cancer patients. The device consisted of substrate which was graphene oxide, a capture antibody immobilised on the substrate, microchannels and sample loading chip. The biomarkers were detected simultaneously in 30 min with an ultra-low detection limit. The marker panel included a combination of biomarkers like CA125, HE4, AFP, CA153, OPN, MSLN HSP70 and inflammatory factors like IL6 and IL8. The integrated panel not only increased the sensitivity and specificity of assay but also helped the researchers to sort ovarian cancer patients in different subgroups [84].

Microfluidic lab on chip nanosensor has also been utilised to capture exosomes by functionalising the platform with membrane biomarkers of cancer and normal exosomes. It helped in isolating exosomes from high grade serous ovarian cancer serum which were then characterised by fluorescence microscopy, flow cytometry and nanoparticle

tracking analysis. The results showed that the amount of total and EpCAM + exosomes increased as the high grade serous ovarian cancer advanced to higher stages [85].

Another diagnostic technology used is the microarray technology which can differentiate gene expression in normal and ovarian cancer patients. It can look for genes whose expression is different in two states, thus giving a slight picture of the respective proteins in the serum of individuals. New molecular biomarkers can be discovered by this technology and also gene expression of some thousand genes can be studied by DNA microarray technology. HE4 marker was discovered with the help of this technology only [86]. The sensitivity data of above mentioned diagnostic platforms has been listed in table 3.

The biosensors which have been investigated until recently for ovarian cancer detection are based on determination of CA125 serum levels in patients. Hence more work needs to be done in fabrication of biosensors that can detect other sensitive biomarkers like HE4 and prostasin. Even though optical biosensors have provided promising analytical and diagnostic results, they have an intricate and sophisticated arrangement which can consume a lot of time and may not be always specific for CA125 detection. This interferes with their potential to be commercialised for the time-being. The electrochemical and magneto resistive based sensors have the ability to tackle many challenges that are faced by other sensors because they have shown excellent sensitivity and biomolecule loading capacity in various experimental analyses. However, the reports currently present for portable sensors that can be employed in clinical settings are very few. Such a challenge can be overcome by fabrication of paper-based sensors and microfluidic lab-on-chip devices. The development of miniaturised biosensors into home-based testing kits is a promising future approach that will allow screening of health of ovarian cancer patients at homes. The factors that still limit the application of nanotechnology based diagnostic approaches are low specificity towards a biomarker, stability of a sensor, its redeemability over the operational time period and an inability to detect biomarker concentrations when expected levels are quite low [15, 95, 96]. Another approach can be colorimetric and electrochemical based detection of nucleic acid biomarkers but without amplification, as amplification might create errors due to certain biases [97]. Addressing these problems can provide great avenues for opening these diagnostic platforms to the commercial setup.

Nanotechnology has also enabled the biomarker targeted therapy of tumours. The nanoparticles can be used as drug carriers and also targeted to their specific distant sites by functionalising them. A theranostic nanosystem was developed by using magnetic iron oxide nanoparticles. The nanoparticles were loaded with drug cisplatin and also imaging ability was provided by near infrared dye labelled with HER2 antibody. This approach helped in imaging the drug delivery and also allowed researchers to receive information on why tumours could be resistant to drugs. It also enables to study heterogeneous expression of HER2 in ovarian cancer tumours [98].

5. Conclusion

The conventional and current methods in place for detection of ovarian cancer including pelvic examination, sonography, invasive biopsies and radiography have not helped much in reducing the mortality rate of one of the deadliest gynecological malignancies. The minimally invasive, highly specific and sensitive biomarkers can change the way such cancers are diagnosed and treated. Such biomarkers need to be identified, evaluated and analysed for better detection. The analysis of cell lines and patient samples and application of proteomic have helped the researchers and clinicians to identify biomarkers which can give good clinical outcomes. These need to be validated by further research so that they can proceed from lab to clinical settings. Different protein, nucleic acids and immune biomarkers have been discussed in this review which have the potential to be used in screening of ovarian cancer patients. There is a need to develop better diagnostic platforms that can assess the expression of these markers in the serum, blood or urine sample of the patients and give rapid results. The most recent approach which can provide point-of-care diagnostic is nanotechnology. The microfluidic chip can integrate multiple markers, thus providing faster results with small sample volumes. The biomarker knowledge has also helped researchers to combine diagnosis using nanotechnology with therapy which can be a promising approach in future in tackling ovarian cancer.

References

[1] Rakoff-Nahoum S 2006 Why cancer and inflammation ? *Yale J. Biol. Med.* **79** 123–30 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1994795/)
[2] Sporn M B 1997 The war on cancer: a review *Ann. N. Y. Acad. Sci.* **833** 137–46
[3] Suh K S *et al* 2010 Ovarian cancer biomarkers for molecular biosensors and translational medicine *Expert Rev. Mol. Diagn.* **10** 1069–83
[4] Matulonis U A, Sood A K, Fallowfield L, Howitt B E, Sehouli J and Karlan B Y 2016 Ovarian cancer *Nat. Rev. Dis. Prim.* **2** 1–22
[5] Lengyel E 2010 Ovarian cancer development and metastasis *Am. J. Pathol. (Internet)* **177** 1053–64
[6] Jayson G C, Kohn E C, Kitchener H C and Ledermann J A 2014 Ovarian cancer *Lancet (Internet)* **384** 1376–88
[7] Sueblinvong T and Carney M E 2009 Current understanding of risk factors for ovarian cancer *Curr. Treat. Options Oncol.* **10** 67–81
[8] Cramer D W and Welch W R 1983 Determinants of ovarian cancer risk. II. Inferences regarding pathogenesis *J. Natl. Cancer Inst.* **71** 717–21
[9] Antoniou A C, Gayther S A, Stratton J F, Ponder B A J and Easton D F 2000 Risk models for familial ovarian and breast cancer *Genet. Epidemiol.* **18** 173–90
[10] Cragun J M 2011 Screening for ovarian cancer *Cancer Control* **18** 16–21
[11] Van Nagell J R and Hoff J T 2014 Transvaginal ultrasonography in ovarian cancer screening: current perspectives *Int. J. Womens Health* **6** 25–33

[12] Carollo E *et al* 2019 Detecting ovarian cancer using extracellular vesicles: progress and possibilities *Biochem. Soc. Trans.* **47** 295–304

[13] Coticchia C M, Yang J and Moses M A 2008 Ovarian cancer biomarkers: current options and future promise *JNCCN J. Natl. Compr. Cancer Netw.* **6** 795–802

[14] Henry N L and Hayes D F 2012 Cancer biomarkers *Mol. Oncol. (Internet)* **6** 140–6

[15] La F B D and Thompson M 2019 Early stage detection and screening of ovarian cancer: a research opportunity and significant challenge for biosensor technology *Biosens. Bioelectron. (Internet)* **135** 71–81

[16] Mahato K, Kumar A, Maurya P K and Chandra P 2018 Shifting paradigm of cancer diagnoses in clinically relevant samples based on miniaturized electrochemical nanobiosensors and microfluidic devices *Biosens. Bioelectron. (Internet)* **100** 411–28

[17] Yuan J, Duan R, Yang H, Luo X and Xi M 2012 Detection of serum human epididymis secretory protein 4 in patients with ovarian cancer using a label-free biosensor based on localized surface plasmon resonance *Int. J. Nanomedicine* **7** 2921–8

[18] Strimbu K and Tavel J A 2010 What are biomarkers ? *Curr. Opin. HIV AIDS* **5** 463–6

[19] Ludwig J A and Weinstein J N 2005 Biomarkers in cancer staging, prognosis and treatment selection *Nat. Rev. Cancer* **5** 845–56

[20] Leon S A, Shapiro B, Sklaroff D M and Yaros M J 1977 Free DNA in the serum of cancer patients and the effect of therapy *Cancer Res.* **37** 646–50 (https://cancerres.aacrjournals.org/content/37/3/646.long)

[21] Sarojini S *et al* 2012 Early detection biomarkers for ovarian cancer *J. Oncol.* **2012** 709049

[22] Capriglione S *et al* 2017 Ovarian cancer recurrence and early detection: may HE4 play a key role in this open challenge? A systematic review of literature *Med. Oncol.* **34** 164

[23] Grayson K, Gregory E, Khan G and Guinn B 2019 Urine biomarkers for the early detection of ovarian cancer—are we there Yet ? *Biomarkers in Cancer* **11** 1–8

[24] Bast R C and Spriggs D R 2011 More than a biomarker: CA125 may contribute to ovarian cancer pathogenesis *Gynecol. Oncol. (Internet)* **121** 429–30

[25] Moore R G, MacLaughlan S and Bast R C 2010 Current state of biomarker development for clinical application in epithelial ovarian cancer *Gynecol. Oncol. (Internet)* **116** 240–5

[26] Drapkin R *et al* 2005 Human epididymis protein 4 (HE4) is a secreted glycoprotein that is overexpressed by serous and endometrioid ovarian carcinomas *Cancer Res.* **65** 2162–9

[27] Naz Z *et al* 2018 Alpha-fetoprotein : a fabulous biomarker in hepatocellular, gastric and rectal cancer diagnosis *Biomedical Research* **29** 2478–83

[28] Klein T and Bischoff R 2011 Physiology and pathophysiology of matrix metalloproteases *Amino Acids* **41** 271–90

[29] Li L, Zhou X, Gu Y and Yan J 2013 Prognostic value of MMP-9 in ovarian cancer : a meta-analysis *Asian Pac. J. Cancer Prev.* **14** 4107–13

[30] Malemud C J 2006 Matrix metalloproteinases (MMPs) in health and disease: an overview *Front. Biosci.* **11** 1696–701

[31] Shay G, Lynch C C and Fingleton B 2015 Moving targets: emerging roles for MMPs in cancer progression and metastasis *Matrix Biol. (Internet)* **44–46** 200–6

[32] Huang H 2018 Matrix metalloproteinase-9 (MMP-9) as a cancer biomarker and MMP-9 Biosensors: recent advances *Sensors* **18** 3249

[33] Hu X *et al* 2012 Matrix metalloproteinase-9 expression correlates with prognosis and involved in ovarian cancer cell invasion *Archives of Gynecology and Obstetrics* **286** 1537

[34] Zeng L, Qian J, Zhu F, Wu F, Zhao H and Zhu H 2019 The prognostic values of matrix metalloproteinases in ovarian cancer *J. Int. Med. Res.* **48** 1–9

[35] Al-alem L 2016 *HHS Public Access.* **150** 2

[36] Liu X, Ye H, Wang P, Li L, Zhang Y and Zhang J 2017 Proteomic-based identification of HSP70 as a tumor-associated antigen in ovarian cancer *Oncology Reports* **37** 2771–8

[37] Chanteloup G *et al* 2020 Membrane-bound exosomal HSP70 as a biomarker for detection and monitoring of malignant solid tumours : a pilot study *Pilot and Feasibility Studies* **6** 35

[38] Mok S C *et al* 2001 Prostasin, a potential serum marker for ovarian cancer: Identification through microarray technology *J. Natl. Cancer Inst.* **93** 1458–64

[39] Huddleston H G, Wong K K, Welch W R, Berkowitz R S and Mok S C 2005 Clinical applications of microarray technology: creatine kinase B is an up-regulated gene in epithelial ovarian cancer and shows promise as a serum marker *Gynecol. Oncol.* **96** 77–83

[40] Huang C Y *et al* 2006 Serum mesothelin in epithelial ovarian carcinoma: a new screening marker and prognostic factor *Anticancer Res.* **26** 4721–8 (https://ar.iiarjournals.org/content/26/6C/4721.long)

[41] Hilliard T S 2018 The impact of mesothelin in the ovarian cancer tumor microenvironment *Cancers (Basel)* **10** 277

[42] Chang M C *et al* 2012 Mesothelin enhances invasion of ovarian cancer by inducing MMP-7 through MAPK/ERK and JNK pathways *Biochem. J* **442** 293–302

[43] Coelho R, Marcos-Silva L, Ricardo S, Ponte F, Costa A, Lopes J M and David L 2018 Peritoneal dissemination of ovarian cancer: role of MUC16-mesothelin interaction and implications for treatment *Expert Review of Anticancer Therapy* **18** 177–86

[44] Huang J, Hu W and Sood A K 2011 Prognostic biomarkers in ovarian cancer *Cancer Biomarkers* **8** 231–51

[45] Slamon D J *et al* 1989 HER-2/neu proto-oncogene in human bresat and ovarian cancer *Science (80-)* **244** 707–12

[46] Lubowicka E *et al* 2020 Diagnostic power of cytokine M-CSF, Metalloproteinase 2 (MMP-2) and tissue inhibitor-2 (TIMP-2) in cervical cancer patients based on ROC analysis *Pathol. Oncol. Res.* **26** 791–800

[47] Będkowska G E *et al* 2019 Diagnostic power of selected cytokines, MMPs and TIMPs in ovarian cancer patients—ROC analysis *Anticancer Res.* **39** 2575–82

[48] Nash A D, Baca M, Wright C and Scotney P D 2006 The biology of vascular endothelial growth factor-B (VEGF-B) *Pulm. Pharmacol. Ther.* **19** 61–9

[49] Hefler L A *et al* 2006 Preoperative serum vascular endothelial growth factor as a prognostic parameter in ovarian cancer *Gynecol. Oncol.* **103** 512–7

[50] Li L *et al* 2004 Correlation of serum VEGF levels with clinical stage, therapy efficacy, tumor metastasis and patient survival in ovarian cancer *Anticancer Res.* **24** 1973–9 (https://ar.iiarjournals.org/content/24/3B/1973.long)

[51] Mesiano S, Ferrara N and Jaffe R B 1998 Role of vascular endothelial growth factor in ovarian cancer: inhibition of ascites formation by immunoneutralization *Am. J. Pathol.* **153** 1249–56

[52] Duyndam M C A *et al* 2002 Vascular endothelial growth factor-165 overexpression stimulates angiogenesis and induces cyst formation and macrophage infiltration in human ovarian cancer xenografts *Am. J. Pathol.* **160** 537–48

[53] Hata K, Watanabe Y, Nakai H, Hata T and Hoshiai H 2011 Expression of the vascular endothelial growth factor (VEGF) gene in epithelial ovarian cancer: an approach to anti-VEGF therapy *Anticancer Res.* **31** 731–7 (https://ar.iiarjournals.org/content/31/2/731)

[54] Kelemen L E 2006 The role of folate receptor α in cancer development, progression and treatment: cause, consequence or innocent bystander ? *Int. J. Cancer.* **119** 243–50

[55] Wang Y, Li X and Qu P 2020 Bioinformatics analysis of FOLR1 expression, functional enrichment, related signaling pathways and relationship with prognosis in ovarian cancer *Pteridines* **31** 46–54

[56] Yung M M-H, Tang H W-M, Cai P C-H, Leung T H-Y, Ngu S-F, Chan K K-L, Xu D, Yang H, Ngan H Y-S and Chan D W 2018 GRO-α and IL-8 enhance ovarian cancer metastatic potential via the CXCR2-mediated TAK1/NFκB signaling cascade *Theranostics* **8** 1270–85

[57] Wen J, Zhao Z, Huang L, Wang L, Miao Y and Wu J 2020 IL-8 promotes cell migration through regulating EMT by activating the Wnt/β-catenin pathway in ovarian cancer *J. Cell. Mol. Med.* **24** 1588–98

[58] Zhang L, Liu W, Wang X, Wang X and Sun H 2019 Prognostic value of serum IL-8 and IL-10 in patients with ovarian cancer undergoing chemotherapy *Oncol. Lett.* **17** 2365–9

[59] Fahmi M, Pradjatmo H, Astuti I and Nindrea R 2021 Cytokines as prognostic biomarkers of epithelial ovarian cancer (EOC): a systematic review and meta-analysis *Asian Pac. J. Cancer Prev.* **22** 315–23

[60] El Bairi K, Kandhro A H, Gouri A, Mahfoud W, Louanjli N, Saadani B, Afqir S and Amrani M 2017 Emerging diagnostic, prognostic and therapeutic biomarkers for ovarian cancer *Cell Oncol.* **40** 105–18

[61] Russell M R *et al* 2016 Protein Z : a putative novel biomarker for early detection of ovarian cancer *International Journal of Cancer* **138** 2984–92

[62] Hessvik N P and Llorente A 2018 Current knowledge on exosome biogenesis and release *Cell. Mol. Life Sci.* **75** 193–208

[63] Becker A, Thakur B K, Weiss J M, Kim H S, Peinado H and Lyden D 2016 Extracellular vesicles in cancer: cell-to-cell mediators of metastasis *Cancer Cell* **30** 836–48

[64] Kalluri R and LeBleu V S 2020 The biology, function, and biomedical applications of exosomes *Science* **367** eaau6977

[65] Soung Y, Ford S, Zhang V and Chung J 2017 Exosomes in cancer diagnostics *Cancers* **9** 8

[66] Cheruvanky A, Zhou H, Pisitkun T, Kopp J B, Knepper M A, Yuen P S T and Star R A 2007 Rapid isolation of urinary exosomal biomarkers using a nanomembrane ultrafiltration concentrator *Am. J. Physiol. Renal Physiol.* **292** F1657–61

[67] Zaman M S, Maher D M, Khan S, Jaggi M and Chauhan S C 2012 Current status and implications of microRNAs in ovarian cancer diagnosis and therapy *J. Ovarian Res.* **5** 1–11

[68] Alharbi M *et al* 2020 miRNa signature in small extracellular vesicles and their association with platinum resistance and cancer recurrence in ovarian cancer *Nanomedicine: Nanotechnology, Biology and Medicine (Internet)* **28** 102207

[69] Taylor D D and Gercel-Taylor C 2008 MicroRNA signatures of tumor-derived exosomes as diagnostic biomarkers of ovarian cancer *Gynecol. Oncol.* **110** 13–21

[70] Huang T and Deng C X 2019 Current progresses of exosomes as cancer diagnostic and prognostic biomarkers *Int. J. Biol. Sci.* **15** 1–11

[71] Schuster-Little N, Fritz-Klaus R, Etzel M, Patankar N, Javeri S, Patankar M S and Whelan R J 2021 Affinity-free enrichment and mass spectrometry analysis of the ovarian cancer biomarker CA125 (MUC16) from patient-derived ascites *Analyst* **146** 85–94

[72] Ehrhart N 1998 Principles of tumor biopsy *Clinical Techniques in Small Animal Practice* **13** 10

[73] Boriachek K *et al* 2018 Biological functions and current advances in isolation and detection strategies for exosome nanovesicles *Small* **14** 1702153

[74] Ibsen S D *et al* 2017 Rapid isolation and detection of exosomes and associated biomarkers from plasma *ACS Nano.* **11** 6641–51

[75] Soda N, Rehm B H A, Sonar P, Nguyen N T and Shiddiky M J A 2019 Advanced liquid biopsy technologies for circulating biomarker detection *J. Mater. Chem. B* **7** 6670–704

[76] Ilié M and Hofman P 2016 Pros: can tissue biopsy be replaced by liquid biopsy? *Transl. Lung Cancer Res.* **5** 420–3

[77] Aziz NB *et al* 2021 MicroRNAs in ovarian cancer and recent advances in the development of microRNA-based biosensors *Analyst* **145** 2038–57

[78] Bast Jr, Lu, Han *et al* 2020 *American Association for Cancer Research.* **29** 2504–12 (https://cebp.aacrjournals.org/content/29/12/2504)

[79] Barani M, Bilal M, Sabir F, Rahdar A and Kyzas G Z 2021 Nanotechnology in ovarian cancer: diagnosis and treatment *Life Sci.* **266** 118914

[80] Gazze A, Ademefun R, Conlan R S and Teixeira S R 2018 Electrochemical impedence spectroscopy enabled CA125 detection; toward early ovarian cancer diagnosis using graphene biosensors *Journal of Interdisciplinary Nanomedicine* **3** 82–8

[81] Al-Ogaidi I, Gou H, Aguilar Z P, Guo S, Melconian A K, Al-kazaz A K A, Meng F and Wu N 2014 Detection of the ovarian cancer biomarker CA-125 using chemiluminescence resonance energy transfer to graphene quantum dots *Chem. Commun.* **50** 1344–6

[82] Mahmudunnabi R G, Farhana F Z, Kashaninejad N, Firoz S H, Shim Y B and Shiddiky M J A 2020 Nanozyme-based electrochemical biosensors for disease biomarker detection *Analyst* **145** 4398–420

[83] Masud M K *et al* 2019 Superparamagnetic nanoarchitectures for disease-specific biomarker detection *Chem. Soc. Rev.* **48** 5717–51

[84] Wu Y, Wang C, Wang P, Wang C, Zhang Y and Han L 2021 A high-performance microfluidic detection platform to conduct a novel multiple-biomarker panel for ovarian cancer screening *RSC Adv.* **11** 8124–33

[85] Hisey C L, Dorayappan K D P, Cohn D E., Selvendiran K and Hansford D J 2018 Microfluidic affinity separation chip for selective capture and release of label-free ovarian cancer exosomes *Lab on a Chip* **18** 3144–53

[86] Kulasingam V, Pavlou M P and Diamandis E P 2010 Integrating high-throughput technologies in the quest for effective biomarkers for ovarian cancer *Nat. Rev. Cancer* **10** 371–8

[87] Samadi Pakchin P, Fathi M, Ghanbari H, Saber R and Omidi Y 2020 A novel electrochemical immunosensor for ultrasensitive detection of CA125 in ovarian cancer *Biosens. Bioelectron.* **153** 112029

[88] Liang M, Pan M, Hu J, Wang F and Liu X 2018 Electrochemical biosensor for MicroRNA detection based on cascade hybridization chain reaction *ChemElectroChem* **5** 1380–6

[89] Boriachek K *et al* 2018 An amplification-free electrochemical detection of exosomal miRNA-21 in serum samples *Analyst* **143** 1662–9

[90] Moazampour M, Zare H R and Shekari Z 2021 Femtomolar determination of an ovarian cancer biomarker (miR-200a) in blood plasma using a label free electrochemical biosensor based on l-cysteine functionalized ZnS quantum dots *Anal. Methods* **13** 2021–9

[91] Da H, Liu Y, Li M, Yuan R, Liu H and Chai Y 2019 A highly sensitive photoelectrochemical VEGF165 biosensor with a dual signal amplification strategy by using $AgVO_3$ as a photoactive material *Chem. Commun.* **55** 8076–8

[92] Ma X, Huang G, Ye M, Zhang X, Wang Y, Liang T, Deng H and Li C 2021 A homogeneous biosensor for human epididymis protein 4 based on upconversion luminescence resonance energy transfer *Microchem. J.* **164** 106083

[93] Boriachek K, Islam M N, Gopalan V, Lam A K, Nguyen N T and Shiddiky M J A 2017 Quantum dot-based sensitive detection of disease specific exosome in serum *Analyst* **142** 2211–9

[94] Bunney P E, Zink A N, Holm A A, Billington C J and Kotz C M 2017 Orexin activation counteracts decreases in nonexercise activity thermogenesis (NEAT) caused by high-fat diet *Physiology & Behavior* **176** 139–48

[95] Razmi N and Hasanzadeh M 2018 Current advancement on diagnosis of ovarian cancer using biosensing of CA 125 biomarker: analytical approaches *TrAC, Trends Anal. Chem.* **108** 1–12

[96] Sha R and Badhulika S 2020 Recent advancements in fabrication of nanomaterial based biosensors for diagnosis of ovarian cancer: a comprehensive review *Microchim. Acta* **187** 181

[97] Soda N, Umer M, Kashaninejad N, Kasetsirikul S and Shiddiky M J A 2020 PCR-Free Detection of Long Non-Coding HOTAIR RNA in Ovarian Cancer Cell Lines and Plasma Samples *Cancers* **12** 2233

[98] Satpathy M *et al* 2019 Targeted drug delivery and image-guided therapy of heterogeneous ovarian cancer using HER2-targeted theranostic nanoparticles *Theranostics* **9** 778–95

Clean Energy, 2021, 476–491

https://doi.org/10.1093/ce/zkab028
Homepage: https://academic.oup.com/ce

Research Article

Performance comparison of the solar-driven supercritical organic Rankine cycle coupled with the vapour-compression refrigeration cycle

Yunis Khan*[*] and Radhey Shyam Mishra

Delhi Technological University, Department of Mechanical Engineering, Bawana Road, New Delhi, 110042, India

*Corresponding author. E-mail: yuniskhan21@gmail.com

Abstract

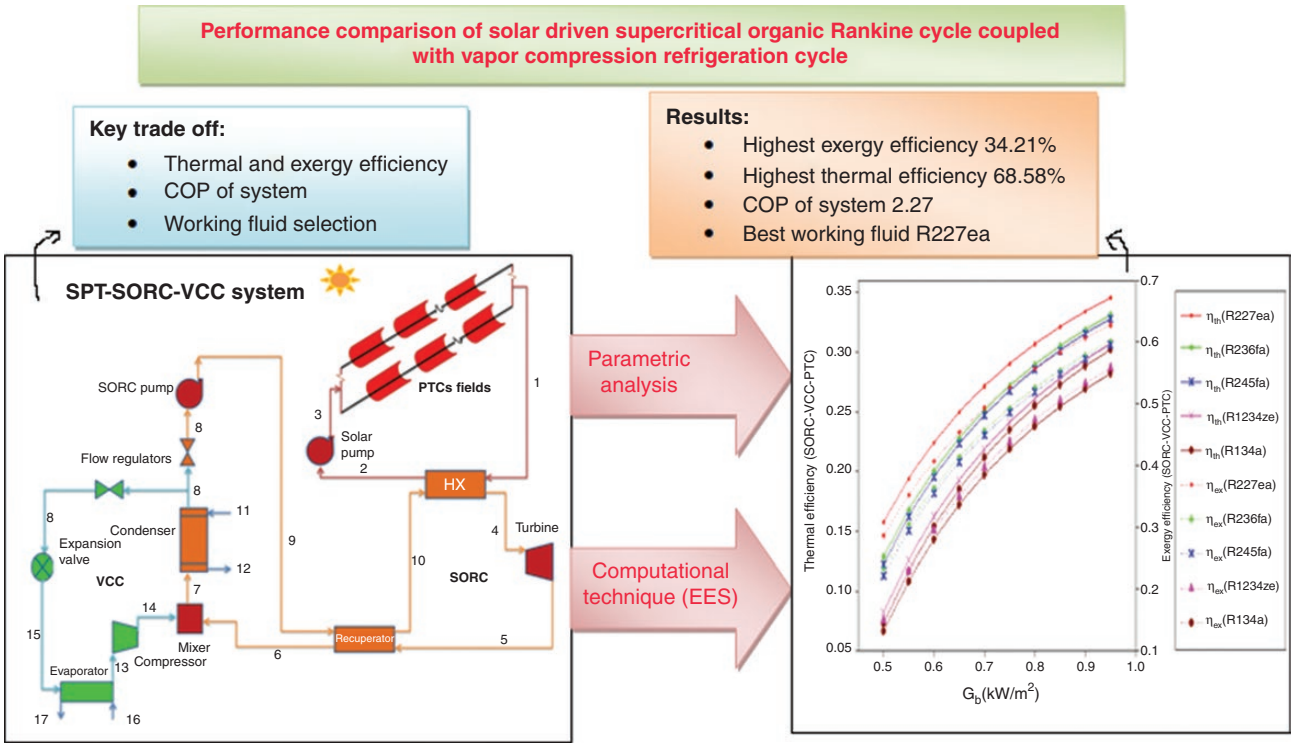
In this study, a parametric analysis was performed of a supercritical organic Rankine cycle driven by solar parabolic trough collectors (PTCs) coupled with a vapour-compression refrigeration cycle simultaneously for cooling and power production. Thermal efficiency, exergy efficiency, exergy destruction and the coefficient of performance of the cogeneration system were considered to be performance parameters. A computer program was developed in engineering equation-solver software for analysis. Influences of the PTC design parameters (solar irradiation, solar-beam incidence angle and velocity of the heat-transfer fluid in the absorber tube), turbine inlet pressure, condenser and evaporator temperature on system performance were discussed. Furthermore, the performance of the cogeneration system was also compared with and without PTCs. It was concluded that it was necessary to design the PTCs carefully in order to achieve better cogeneration performance. The highest values of exergy efficiency, thermal efficiency and exergy destruction of the cogeneration system were 92.9%, 51.13% and 1437 kW, respectively, at 0.95 kW/m^2 of solar irradiation based on working fluid R227ea, but the highest coefficient of performance was found to be 2.278 on the basis of working fluid R134a. It was also obtained from the results that PTCs accounted for 76.32% of the total exergy destruction of the overall system and the cogeneration system performed well without considering solar performance.

Received: 16 May 2021; **Accepted:** 28 July 2021

© The Author(s) 2021. Published by Oxford University Press on behalf of National Institute of Clean-and-Low-Carbon Energy
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/ce/article/5/3/476/6375889 by Delhi Technological University user on 06 October 2021

Graphical Abstract



Keywords: parametric analysis; solar parabolic trough collector; vapour-compression refrigeration cycle supercritical organic Rankine cycle; cogeneration; cooling; power generation

Introduction

Energy consumption used for cooling and power generation has increased drastically in recent years [1]. The energy consumed in air-conditioning and cooling equipment is large [2].The vapour-compression refrigeration system is a widely used cooling system in which the compressor is the highest energy-consuming component because this energy is used to increase the coolant pressure. However, more energy use in the cooling system increases carbon-dioxide emissions, leading to higher global temperature and greenhouse effects. Therefore, renewable-energy technology is now used for refrigeration purposes. Such renewable energy sources are geothermal, solar, biomass, etc. [3]. The justification for using renewable energy for cooling is to reduce emissions of carbon dioxide and power consumption. Moreover, the cogeneration system of the organic Rankine cycle and vapour-compression cycle (ORC–VCC) is a heat-driven power and refrigeration system. The heat-driven cooling system is a clean technology and free from pollution. Solar energy is one of the most promising renewable heat sources because of its low costs, noise-free operation and abundance in nature [4]. Solar energy is the most suitable for cooling, heating and power generating, among other renewable energy sources. Solar collectors are being used to harvest solar energy where heat-transfer fluid (HTF) circulates for absorbing the heat of the Sun, which is used to drive the different types of thermodynamic cycles [5].

Nowadays, solar parabolic trough collectors (PTCs) are being widely used as heat sources to run combined as well as simple cycle electricity-generation systems. In this direction, a few studies were conducted using PTCs for trigeneration and cogeneration applications, e.g. Al-Sulaiman *et al.* [6] carried out a study on the PTC-driven ORC for utilizing waste heat for a cogeneration process. They found that during trigeneration, the energy efficiency was increased from 15% to 94%. Al-Sulaiman [7] carried out a thermodynamic analysis of the PTC-integrated steam Rankine cycle with ORC and R134a as the best working fluid among the other selected working fluids because it provided the highest electrical efficiency of 26%. The exergy and energy analysis of the PTC-integrated supercritical organic Rankine cycle (SORC) system was examined by Singh and Mishra [8]. Exergetic metrics including the fuel-depletion ratio, improvement potential and irreversibility ratio were found to be 0.579, 11 859 kW and 0.9296, respectively. Among the other working fluids tested, R600a was the best. Singh and Mishra [9] carried out a study on the PTC-driven supercritical carbon-dioxide cycle (sCO$_2$)–ORC. They discovered that solar irradiation increased the thermal and exergy efficiency of the combined cycle. R407c was chosen as the best working fluid, with a combined system thermal

Downloaded from https://academic.oup.com/ce/article/5/3/476/6375889 by Delhi Technological University user on 06 October 2021

and exergy efficiency of 43.49% and 78.07%, respectively, at 0.95 kW/m^2. A PTC-driven partial heating sCO_2 cycle combined with ORC was studied by Khan and Mishra [10]. They discovered that solar irradiation enhanced the thermal and exergy efficiency, while the solar incidence angle reduced performance. Al-Zahrani and Dincer [11] performed exergy and energy analysis of the PTC-driven sCO_2 cycle to convert solar heat into power. They considered the parameters of the PTC as design parameters such as solar irradiation, receiver emittance and solar-beam incidence angle. The exergy and energy efficiency of the PTC were observed as 66.35% and 38.51%, respectively.

The SORC is better technology for the conversion of waste heat to power. For example, Song *et al.* [12] analysed the SORC to recover waste heat at low temperature. They concluded that the SORC was an appropriate method for industrial waste-heat recovery and the R152a was observed as the best-performing working fluid and produced the maximum net power and thermal efficiency. Kalra *et al.* [13] demonstrated that the operating fluid pressure was above its critical pressure in the SORC and found that the SORC had various benefits over its subcritical ORC because the heat-source cooling curve in the SORC matched the working-fluid heating curve. A parametric analysis of the SORC system for the medium-temperature geothermal heat source was undertaken by Moloney *et al.* [14]. They came to the conclusion that the SORC was more efficient than the ORC for low-temperature heat sources. Saadon and Islam [15] performed a thermal analysis of the SORC and discovered that it had a higher thermal efficiency and power output than the conventional ORC with a preheater.

Also several studies were conducted on the combined ORC–VCC system in recent years. Pektezel and Acar [16] analysed the combined ORC and single and double evaporator VCC systems for exergy and energy. They discovered that R600a was the most suitable fluid for the combined system. They also discovered that, with a single evaporator, the coefficient of performance (COP) and exergy efficiency of the combined cycle were higher than with a dual evaporator. The ORC-integrated VCC system was subjected to an exergy and energy study by Saleh [17]. He came to the conclusion that R602 was a good working fluid for system performance and environmental concerns. Finally, he concluded that the highest COP, exergy efficiency and pressure ratio of the turbine of the system and the corresponding total mass flow rate of the working fluid using R602 were 0.99, 53.8%, 12.2 and 0.005 kg/s-kW, respectively, at 25°C condenser temperature and maintaining other parameters as constant. Javanshir *et al.* [18] performed thermal and exergoeconomic analyses on the regenerative ORC–VCC system. Their results indicated that among other selected working fluids, R134a showed the lowest exergy and thermal efficiencies while R143a and R22 showed the highest exergy and energy efficiency. They also observed that the total unit cost of the product was found to be $60.7/GJ. The maximum exergy destruction was found in the boiler followed by the turbine. Hu *et al.* [19] investigated the performance of an ORC–VCC system operated by solar power. They concluded that the ice production and the cooling power depended on the condensation and generation temperature. R245fa was selected as an appropriate working fluid for this purpose. Using R245fa, cooling power and ice production per unit metre square collector area were 126.44 W/m^2 and 7.61 kg/m^2-day, respectively. Moles *et al.* [20] proposed a model of the low-temperature heat-activated combined ORC–VCC system. They concluded that the computed thermal and electrical COP of the ORC–VCC system varied between 0.30–1.10 and 15–110, respectively. Additionally, R1234ze(E) was considered an acceptable working fluid for enhanced efficiencies. Li *et al.* [21] performed a working-fluid selection study for the combined ORC–VCC system. They concluded that butane was chosen as the better working fluid for the ORC–VCC system, with temperature variations ranging from 60°C to 90°C, –15°C to 15°C and 30°C to 55°C at boiler exit, evaporation and condensation temperatures varying, respectively. The COP of the complete system was obtained as 0.47 based on butane at a boiler exit temperature of 90°C.

It has been observed from the above-mentioned literature survey that several studies have been conducted on the ORC–VCC cogeneration system. However, no study was conducted on the SORC–VCC cogeneration system driven by PTCs. The impacts of the PTC design parameters on the cogeneration system have also not been observed in previous research. This demonstrates the novelty of the current study. The main objective of the present study is to examine the exergetic and energetic performance of the cogeneration SORC–VCC system simultaneously for the cooling and power generation driven by PTCs. The system-performance parameters were considered to be exergy and thermal efficiency, the COP of the cogeneration system and exergy destruction. The effects on the system performance of different parameters such as solar irradiation, the solar-beam incidence angle and the HTF velocity in the absorber tube, condenser temperature and evaporator temperature, and turbine inlet pressure were examined. Finally, the performance of the cogeneration system was compared with and without PTCs.

1 System description

Fig. 1 shows the schematic diagram of the SORC–VCC system driven by the PTC. A PTC field, a regenerating SORC and a VCC are all part of the system. Because there is a lot of heat energy available at the turbine outlet, regeneration of the SORC is used in this study. With the help of a recuperator, this heat is utilized to warm the incoming stream to the heat exchanger. Also, the SORC is beneficial for maximum power production [13]. This system uses two fluids: a single HTF in the solar field and another fluid in both the SORC and VCC subsystems. However, the SORC and VCC subsystems have different mass flow rates. There is a common condenser for both subsystems. For controlling the mass flow rate to both of the subsystems, two flow regulators and a common condenser for both cycles are used. To control the mass flow rate to both cycles, two flow

Downloaded from https://academic.oup.com/ce/article/5/3/476/6375889 by Delhi Technological University user on 06 October 2021

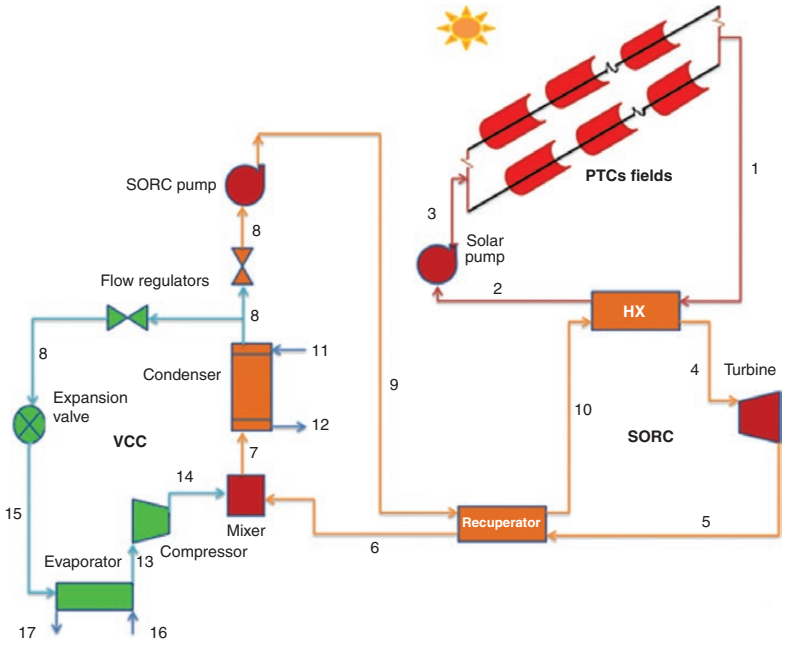


Fig. 1: Schematic of the PTC-driven SORC–VCC cogeneration system

Table 1: Input conditions for simulation of the proposed model

Parabolic trough collector	LS-3 [22]
Maximum exit temperature of HTF	390°C [7]
Maximum exit pressure of HTF	10 MPa
HTF mass flow rate	0.575 (kg/s)
Atmospheric pressure	1.013 bar
Atmospheric temperature	25°C
Effectiveness heat exchanger	0.95 [7]
SORC turbine's isentropic efficiency	0.87 [18]
Compressor's isentropic efficiency	0.8 [18]
SORC pump's isentropic efficiency	0.8 [18]
SORC's mass flow rate	1.5 kg/s
VCC's mass flow rate	1 kg/s
Turbine inlet pressure	50 bar
Outlet pressure to the compressor/turbine	6.018 bar
Condenser temperature	35°C [17]
Effectiveness of recuperator	0.95 [7]
Evaporator temperature	0°C [17]

regulators are used. The fluid flow directions are shown by the arrows. The main advantage of this combined SORC–VCC system is that it produces cooling when power is not needed or vice versa. The advantage of using a combined SORC–VCC system produces cooling when refrigeration is not needed. Therefore, this system is beneficial throughout the year [17]. A solar field with a few hundred PTCs is used. The type of collector and other input data are given in Table 1. The T-s diagram has been displayed in Fig. 2.

2 Thermodynamic analysis

The mathematical modelling of the SORC–VCC system has been conducted in this section. Modelling of the PTC system has already been conducted by Al-Sulaiman [7]. Modelling of the system was performed using the exergy and energy balance in each component. Results are calculated with help of the Engineering Equation Solver (EES) [23] software.

2.1 Thermal modelling of the SORC–VCC system

While solving equations, the following assumptions were considered: (i) steady-state conditions are assumed for all thermodynamic processes; (ii) friction and pressure losses are negligible in each component and pipes; (iii) heat loss to the surroundings is negligible in each component except the condenser unit; (iv) the working-fluid conditions at the entry to the compressor and exit to the condenser in the VCC subsystem are saturated vapour and saturated liquid, respectively; (v) the PTCs have 50 collectors in series in a single row and there are a total of seven rows and the length of each collector is 12.27 m [7].

Modelling equations for the cogeneration (SORC–VCC) system are adapted from the previous studies [13, 18].

The total solar heat available in the SORC–VCC cogeneration system is determined as:

$$\dot{Q}_{HX} = \dot{m}_{SORC} \cdot (h_4 - h_{10}) \quad (1)$$

where $\dot{m}_{SORC}$ is the working-fluid mass flow rate in the SORC subsystem.

The work done by the turbine is given as:

$$\dot{W}_{Turbine} = \dot{m}_{SORC} \cdot (h_4 - h_{5s}) \cdot \eta_{turbine} \quad (2)$$

The heat recovered in the in the recuperator is given by the heat balance equation:

$$\dot{m}_{SORC} \cdot (h_6 - h_5) = \dot{m}_{SORC} \cdot (h_{10} - h_9) \quad (3)$$

Downloaded from https://academic.oup.com/ce/article/5/3/476/6375889 by Delhi Technological University user on 06 October 2021

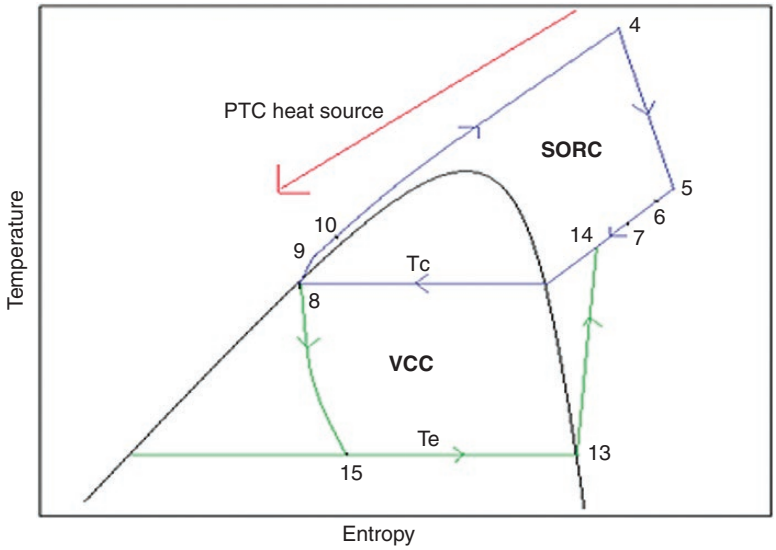


Fig. 2: T-s diagram for the PTC-driven SORC–VCC cogeneration system

If the same working fluid is circulating having the same mass flow rate on the hot and cold sides of the recuperator, then the effectiveness of the recuperator is given as:

$$\varepsilon = \frac{(T_5 - T_6)}{(T_5 - T_9)} = \frac{\text{Actual heat transfer}}{\text{Maximum heat transfer}} \quad (4)$$

The rejection of heat by the condenser unit is given as:

$$\dot{Q}_{cond} = (\dot{m}_{SORC} + \dot{m}_{VCC}) \cdot (h_8 - h_7) \quad (5)$$

where $\dot{m}_{VCC}$ is the mass flow rate of the working fluid in the VCC subsystem.

Applying the energy-balance equation in the mixer:

$$(\dot{m}_{SORC} + \dot{m}_{VCC}) \cdot h_7 = \dot{m}_{SORC} \cdot h_6 + \dot{m}_{VCC} \cdot h_{14} \quad (6)$$

The compressor work is given by:

$$\dot{W}_{comp} = \frac{\dot{m}_{VCC} \cdot (h_{14s} - h_{13})}{\eta_{comp}} \quad (7)$$

where h_{14s} is the enthalpy at state 14 when the isentropic work is to be done.

Further, the heat absorbed by the evaporation or cooling effect can be given as:

$$\dot{Q}_{evap} = \dot{m}_{VCC} \cdot (h_{13} - h_{15}) \quad (8)$$

The SORC pump work can be written as:

$$\dot{W}_{SORC\ \ pump} = \frac{\dot{m}_{SORC} \cdot (h_{9s} - h_8)}{\eta_{SORC\ \ pump}} \quad (9)$$

The thermal efficiency of the cogeneration system (SORC–VCC) is determined as [18]:

$$\eta_{th} = \frac{\dot{W}_{net} + \dot{Q}_{evap}}{\dot{Q}_{HX}} \quad (10)$$

$$\dot{W}_{net} = \dot{W}_{turbine} - \dot{W}_{compressor} - \dot{W}_{SORC\ \ pump} \quad (11)$$

The overall thermal efficiency of the system is calculated as [24]:

$$\eta_{th,overall} = \eta_{ex,overall} \cdot \eta_{Carnot} \quad (12)$$

Furthermore, when the system is brought to dead conditions (298.15 K temperature, 1.013 bar pressure), the maximum theoretical work obtained is termed as exergy. On the basis of the control volume, the exergy balance equation is determined as:

$$\sum \left(1 - \frac{T_0}{T_Q}\right) Q_Q - W_{c.v} - \sum (m_i Ex_i) - \sum (m_e Ex_e) - Ed = 0 \quad (13)$$

where subscript Q denotes the thermal property numerically available at a given stage and Ed refers to the exergy-destruction rate. Furthermore, Ex signifies the specific system exergy. Without considering the chemical exergy, potential energy and kinetic energy, the physical exergy can then be assessed as [7, 25]:

$$Ex_{ph} = (h - h_0) - T_0(s - s_0) \quad (14)$$

The exergy inlet to the system and obtained from solar radiation by Petela's formula [26] is described as:

$$Ex_s = A_p \cdot G_b \cdot \left[1 + \frac{1}{3}\left(\frac{T_0}{T_{su}}\right)^4 - \frac{4}{3}\left(\frac{T_0}{T_{su}}\right)\right] \quad (15)$$

where T_{su} is the temperature of the Sun, i.e. 5800 K [7].

Ex_{evap} can be defined as [18]:

$$Ex_{evap} = \dot{Q}_{evap} \cdot \left(\frac{T_0}{T_e} - 1\right) \quad (16)$$

where T_e is the evaporator temperature.

The exergy efficiency of the SORC–VCC congregation system is defined as [18]:

$$\eta_{ex} = \frac{\dot{W}_{net} + Ex_{evap}}{Ex_S} \quad (17)$$

The exergy efficiency of the overall system (SORC–VCC–PTC) is shown as [24]:

$$\eta_{ex,overall} = 1 - \frac{Ed_{Total}}{Ex_s} \quad (18)$$

Downloaded from https://academic.oup.com/ce/article/5/3/476/6375889 by Delhi Technological University user on 06 October 2021

where Ed_{Total} is the total exergy destruction of the overall system.

Another performance parameter to be defined is the COP of the SORC–VCC cogeneration system (COP_S).This can be defined as [17]:

$$COP_S = \eta_{SORC} \cdot COP_{vcc} = \frac{\dot{Q}_{evap}}{\dot{Q}_{HX} + \dot{W}_{SORC\ \ pump}} \quad (19)$$

where η_{SORC} and COP_{vcc} are the efficiency of the SORC and the COP of the vapour-compression subsystem, respectively.

2.2 Working-fluid selection

Working-fluid selection is challenging because it affects the performance and economic feasibility of the system. It is good if the critical temperatures of the working fluid are near the temperature of the heat source and the temperature of the waste [17]. The HTF selected in the current study is Therminol-VP1 in the absorber tube of the solar collector to absorb the heat of the Sun because this synthetic oil has excellent heat-transfer properties as well as good temperature stability compared with other HTFs such as molten salt and a maximum working temperature of 400°C [27]. Its thermophysical properties were discussed in the study of Mwesigye *et al.* [27]. The working fluids in the SORC system should be non-toxic, environmentally friendly, economical, able to use heat energy from the source of waste heat and also, at optimum pressure and temperature, the thermophysical and thermodynamic properties should be optimal [28]. Due to these reasons, fluids such as R227ea, R236fa, R245fa, R1234ze and R134a were selected. The thermophysical properties of the considered working fluids are shown in Table 2.

2.3 Verification of the model

The current model was validated using previous studies. The PTC as well as the cogeneration SORC–VCC system has been separately verified. Verification with the change in heat loss in the absorber tube with the temperature change above the ambient temperature and exit temperature of the PTC have been shown in Tables 3 and 4, respectively, at the same baseline situations. There is <1% deviation from the experimental results. It is better agreement than in the previous theoretical study by Al-Sulaiman [7] and the experimental study by Dudley *et al.* [22]. Apart from this, the thermal efficiency of the SORC system has been validated by the study by Le *et al.* [29] under the same baseline conditions as given in Table 5. Table 6 shows the COP variation of

Table 2: Thermophysical properties of the working fluids [16, 17]

Working fluids	Weights (g/mole)	Critical temperature (°C)	Critical pressure (MPa)	ODP	GWP
R227ea	170.03	101.8	2.925	0	3580
R236fa	152.04	124.92	3.2	0	680
R245fa	134.05	154.1	3.65	0	1050
R1234ze	104.04	109.36	3.635	0	<1
R134a	102.03	101	4.059	0	1430

Table 3: Heat loss (W/m^2) variation with temperature difference

Temperature difference (°C)	Current model	Dudley *et al.* [22]	Al-Sulaiman [7]
100.6	9.78	10.6	8.719
149.1	16.24	19.3	19.3
196.7	31.15	30.6	34.2
245.8	53.48	45.4	53.0
293.3	66.69	62.9	75.5

Table 4: Verification of the exit temperature of the PTC with experimental results

					T_e (K)		
Cases	G_b (W/m^2)	Discharge (m^3/s)	T_0 (K)	T_i (K)	Dudley *et al.* [22]	Current model	Error (%)
1	933.7	47.7	294.35	375.35	397.15	397.56	0.10
2	968.2	47.8	295.55	424.15	446.45	446.72	0.06
3	982.3	49.1	297.45	470.65	492.65	492.02	0.12
4	909.5	54.7	299.45	523.85	542.55	541.99	0.10
5	937.9	55.5	299.35	570.95	589.55	590.11	0.09
6	880.6	55.6	301.95	572.15	590.35	590.26	0.01
7	903.2	56.3	300.65	529.05	647.15	647.35	0.03
8	920.9	56.8	304.25	652.65	671.15	671.51	0.05

Downloaded from https://academic.oup.com/ce/article/5/3/476/6375889 by Delhi Technological University user on 06 October 2021

Table 5: Verification of the SORC system

		Thermal efficiency (%)		
Baseline conditions	Literature	Literature	Current model	Error (%)
R134a, heat-source temperature = 139°C	Le *et al.* [29]	12	11.99	–1

Table 6: COP variation of the VCC system with the evaporator average temperature

	COP	
Evaporator average temperature (°C)	Current model	Krishnan *et al.* [30]
13	2.85	2.99
13.5	3.22	3.12
14	3.33	3.45
14.5	3.82	3.95
15	3.98	4.02
15.5	4.10	4.21
16	4.42	4.38

the VCC system with the evaporator average temperature and validated by the experimental study by Krishnan *et al.* [30]. Table 7 shows that the thermal efficiency of the SORC–VCC cogeneration system matches that from the study by Javanshir *et al.* [18]. Therefore, the current model is ready for analysis.

3 Results and discussion

In the present study, a thermal performance analysis of the PTC-driven SORC–VCC cogeneration system was conducted. The effects of solar irradiation (G_b), solar-beam incidence angle, HTF velocity in the absorber tube, turbine inlet pressure (TIP), condenser temperature (T_c) and evaporator temperature (T_e) on thermal performance were examined. While investigating the effects of the parameters, other parameters were kept constant as listed in Table 1. A computer program was made using EES software to solve the modelling equations.

3.1 Performance evaluation with solar irradiation

Solar irradiation, also termed as the direct normal irradiation, is a primary parameter to be examined. The variation in solar irradiation was taken as 0.5–0.95 kW/m^2 as per the climate of Mumbai, India. Variation in solar irradiation has a large impact on the system performance ranging from 0.5 to 0.95 kW/m^2. The thermal performance of the SORC–VCC cogeneration system increases with solar irradiation due to the effective utilization of collector rows as shown in Fig. 3. Among the selected working fluids, R227ea gives the highest thermal efficiency that varies from 47.55% to 51.13% at 0.5–0.95 kW/m^2, respectively, because R227ea has the lowest critical pressure compared to other selected working fluids. Near the critical conditions, the working fluid performs better. The exergy efficiency of the cogeneration system also improved with the solar irradiation. The reason behind this is that, as the irradiation increases, the inlet exergy to the cogeneration system increases according to Equation (20) and simultaneously the exergy destruction decreases. R227ea has the highest exergy efficiency ranging from 86.38% to 92.9% at 0.5–0.95 kW/m^2, respectively.

The overall system (PTC-SORC–VCC) exergy and thermal efficiency were also examined. It is observed from Fig. 4 that the overall system thermal efficiency increases with the solar irradiation due to the same reason as above. R227ea and R134a, among the other selected working fluids, gave respectively the highest and lowest thermal efficiency. Using R227ea, the solar irradiation increases from 0.5 to 0.95 kW/m^2, which results in a nearly 58.88% increase in the thermal efficiency. The rise in solar irradiation results in an increase in the overall system exergy efficiency. As the values of the solar irradiation vary from 0.5 to 0.95 kW/m^2, the exergy efficiency based on R227ea improves by almost 58.90%, as illustrated in Fig. 4.

The variation in solar irradiation also has a significant effect on the COP_S, as denoted in Fig. 5. Fig. 5 shows that the COP_S is a function of the solar irradiation. As the solar irradiation increases, this leads to an increase in the COP_S. Increased solar irradiation has no effect on the VCC subsystem, which means no effect on the COP_{VCC}. However, an increase in solar irradiation leads to an increase in η_{SORC} [8]. The COP_S increases with solar irradiation based on Equation (19). R134a has the highest COP_S and R227ea the lowest among the considered working fluids. This means that if this system only works for cooling purposes, R134a is more effective than R277ea. Using the working fluid R134a, as the solar irradiation varied from 0.5 to 0.95 kW/m^2, the COP_S increased by nearly 7.69%. It is also shown that the exergy destruction decreases with the solar irradiation. It has the reverse trend to that of the exergy efficiency, as explained above. R134a and R227ea gave the highest and lowest exergy destruction as shown in Fig. 5. R134a gave exergy-destruction decreases from 3966 to 2060 kW when the solar irradiation varied from 0.5 to 0.95 kW/m^2, respectively. Furthermore, without the solar SORC–VCC, the system is more thermodynamically efficient than with solar.

3.2 Performance evaluation with solar incidence angle

Fig. 6 shows that the thermal and exergy efficiency of the cogeneration system decreases with the increase in the

Downloaded from https://academic.oup.com/ce/article/5/3/476/6375889 by Delhi Technological University user on 06 October 2021

Table 7: Verification of the cogeneration SORC–VCC system

Baseline conditions	Literature	Thermal efficiency (%)		
		Literature	Current model	Error (%)
R143a, TIP = 3600 kPa, heat-source temperature = 150°C	Javanshir *et al.* [18]	27.21	27.98	2.49

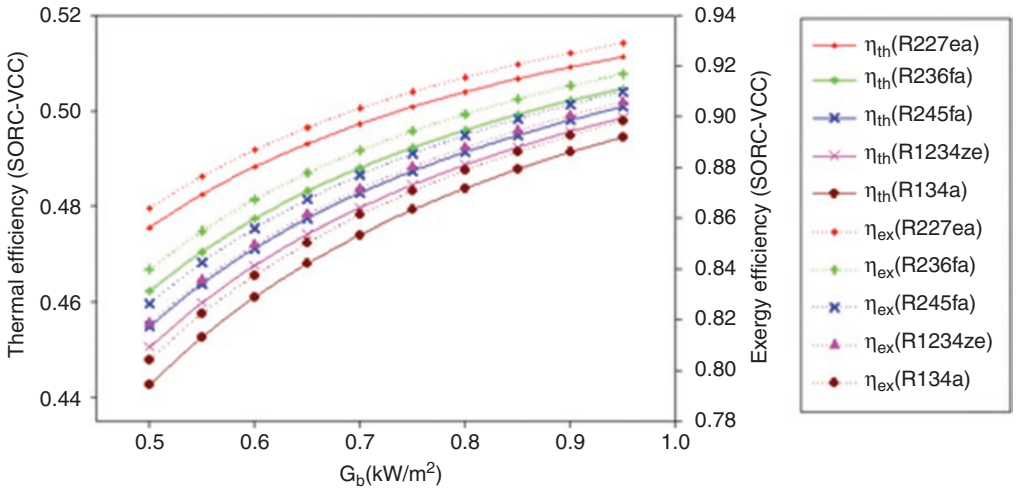


Fig. 3: Variation in the efficiency of the SORC–VCC cogeneration system with solar irradiation

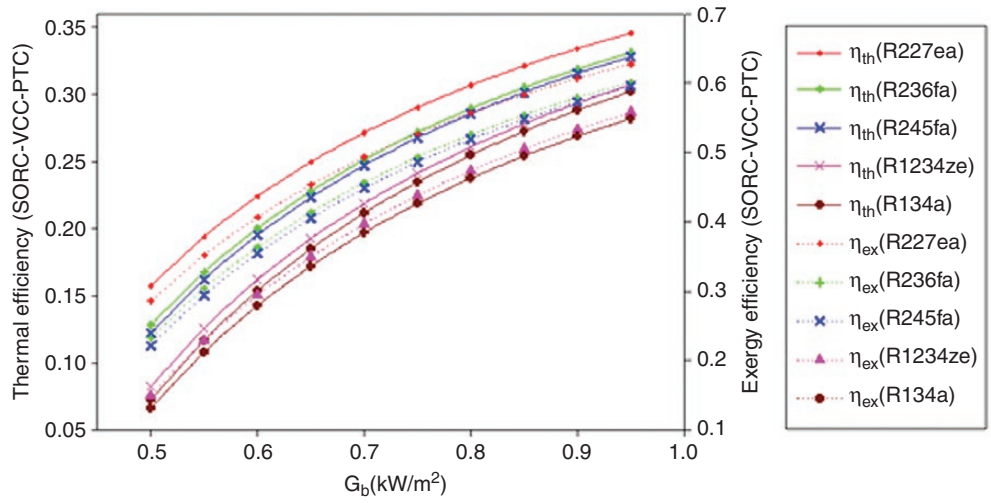


Fig. 4: Variation in the efficiency of the overall system (SORC–VCC–PTC) with solar irradiation

solar incidence angle. An increase in the incidence angle reduces the exergetic, energetic and optical efficiencies of the PTCs [11] and leads to system-performance reduction. The maximum and minimum thermal efficiency and exergy efficiency were found for R227ea and R134a, respectively. The thermal and exergy efficiency decreased by 51.02–50.57% and 92.70–91.88% with increases in the solar incidence angle from 3° to 30°, respectively, based on R227ea. Therefore, it is important to minimize the solar incidence angle to obtain better performance of the system. Also, the COP_s decreased with the solar incidence angle as shown in Fig. 7. The solar incidence angle does not affect the COP_{vcc}. However, it does affect the efficiency of the SORC. Thus, according to Equation (19), the COP decreases with the solar incidence angle. The highest and lowest COP_s were found with R134a and R227ea, respectively. The COP_s decreased from 2.6 to 1.3 as the incidence angle increased from 3° to 30°, respectively, whereas the exergy destruction increased with the solar incidence angle, as shown in Fig. 7. It has been discussed above that the exergy efficiency decreased with the solar incidence angle, which is why the exergy destruction showed a reverse trend. The exergy destruction for all other selected working fluids was found to be between these of the fluids R227ea and R134a. As the solar incidence angle increased from 3° to 30°, the exergy destruction increased from 3966 to 2060 kW on the basis of R134a, as seen in Fig. 7.

Downloaded from https://academic.oup.com/ce/article/5/3/476/6375889 by Delhi Technological University user on 06 October 2021

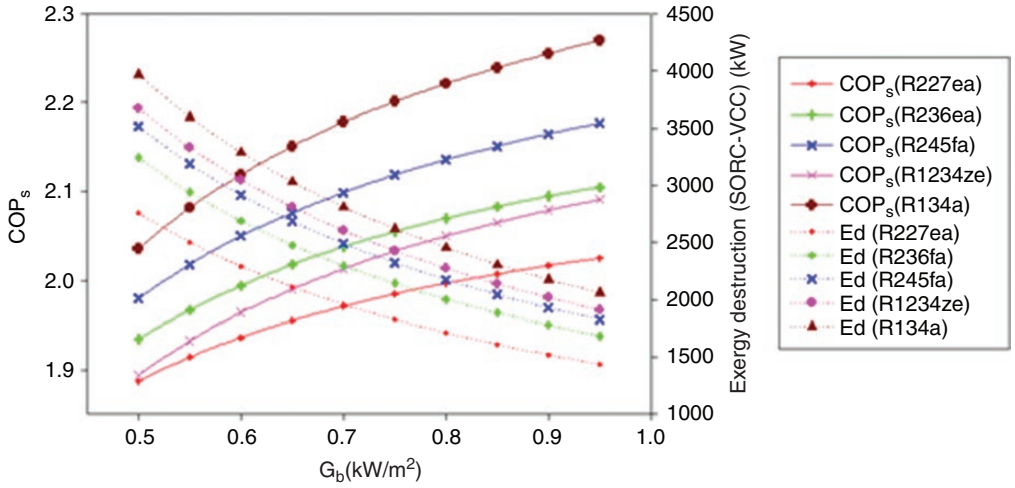


Fig. 5: Variation in the COP_s with solar irradiation

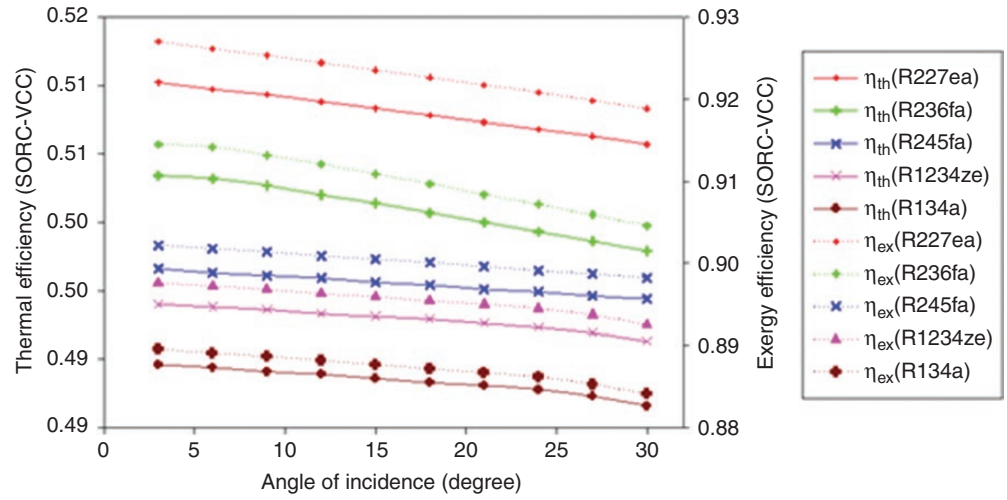


Fig. 6: Variation in the efficiency of the SORC–VCC cogeneration system with solar incidence angle

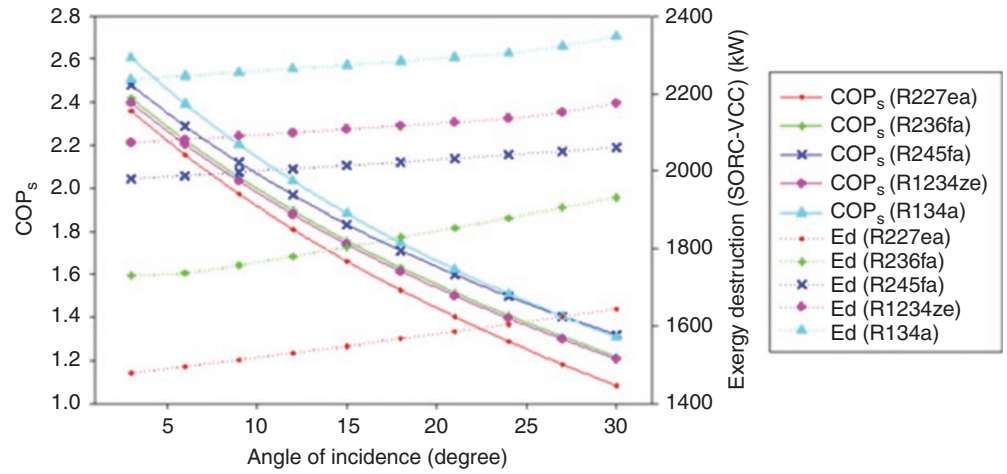


Fig. 7: Variation in the COP_s with the solar incidence angle

3.3 Effects of HTF velocity in the absorber tube on system performance

Fig. 8 displays that the thermal and exergy efficiency of the SORC–VCC cogeneration system has increased with the HTF velocity in the absorber tube. The increase in the efficiency with the velocity is due to the increase in the velocity of the fluid. The Reynolds number is increased as a result of the increase in the convective heat-transfer

Downloaded from https://academic.oup.com/ce/article/5/3/476/6375889 by Delhi Technological University user on 06 October 2021

coefficient; therefore, much heat is carried by the HTF. This leads to an increase in efficiency. Among the operating fluids considered, R227ea and R134a were reported to have the greatest and smallest thermal and energy efficiency. Thermal and exergy efficiency increased from 47.55% to 51.13% and from 86.38% to 92.9% when the velocity increased from 0.01 to 0.1 m/s on the basis of R227ea. The COP_s also increased with the velocity of the HTF. As the velocity increased, the efficiency of the SORC was enhanced. However, there was no effect on the COP_{vcc}. Therefore, according to Equation (19), the COP_s increased with the HTF velocity as seen in Fig. 9. The highest value of the COP_s was found with R134a followed by R245fa, R236fa, R1234ze and R227ea. The exergy destruction of the SORC–VCC cogeneration system was also investigated with the HTF velocity in the absorber tube. The exergy destruction decreased with the HTF velocity. The highest exergy destruction was identified with R134a followed by R1234ze, R245fa, R236fa and R227ea with values of 3966, 3673, 3515, 3241 and 2757 kW, respectively, at an HTF velocity of 0.01 m/s, as shown in Fig. 9.

3.4 Effects of the TIP on system performance

The system performance also depends on the TIP. As shown in Fig. 10, the efficiency of the cogeneration system increased slightly with the TIP. The TIP variation had no impact on the VCC subsystem performance. The VCC subsystem does not contribute to the efficiency of the SORC–VCC cogeneration system. Therefore, the efficiency of the cogeneration system depends only on the performance of the SORC subsystem. It is due only to the η_{SORC} because the η_{SORC} increases with the TIP [8]. It is observed that R227ea and R134a provided the maximum and minimum thermal efficiency of the cogeneration system. In the case of the R227ea, the TIP increase from 40 to 55 bar improved the cogeneration system thermal efficiency by almost 0.85%. From Fig. 10, it can be seen that as the TIP increased, the exergy efficiency of the cogeneration system also increased. The destruction of exergy decreases with the TIP led to exergy-efficiency improvement [8]. Using R227ea, the increase in the TIP from 40 to 55 bar improved the exergy efficiency from 91.82% to 92.60%.

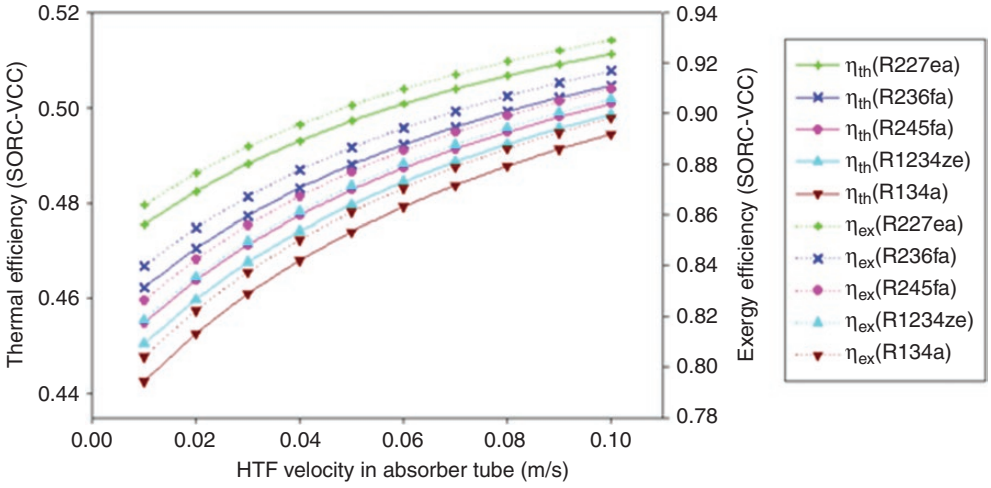


Fig. 8: Variation in the efficiency of the SORC–VCC cogeneration system with HTF velocity

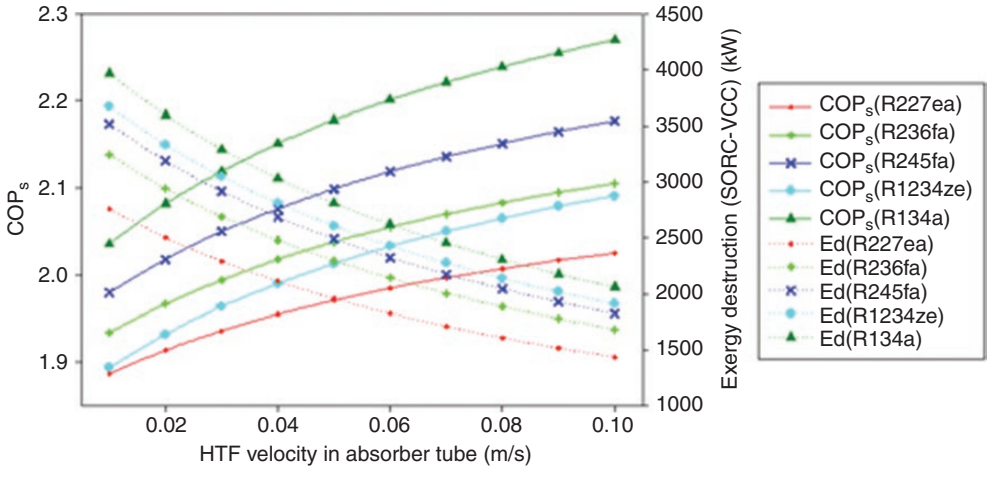


Fig. 9: Variation in the COP_s with the HTF velocity in the absorber tube

Downloaded from https://academic.oup.com/ce/article/5/3/476/6375889 by Delhi Technological University user on 06 October 2021

In addition to the cogeneration system performance, the performance of the overall system was also investigated and the overall system (SORC–VCC–PTC) thermal efficiency was found to increase with the TIP, as seen in Fig. 11. The exergy efficiency of the overall system also depended on the TIP. The exergy efficiency of the overall system improved with the TIP. The TIP did not affect the PTC and VCC subsystems. As a result, the exergy efficiency was only enhanced by the SORC subsystem. In the case of the working fluid R227ea, the TIP increases from 40 to 55 bar improved the exergy efficiency by 5.5%, as shown in the right axis of Fig. 11.

The COP_S is also a function of the TIP. The COP_S slightly increased with the TIP. The reason behind this is that the variation in the TIP had no effect on the COP_{VCC}, although η_{SORC} increased with the TIP [8, 18], from Equation (19), the COP_S increased. R134a and R227ea show the highest and lowest COP_S, respectively. If the TIP increased from 40 to 55 bars, the COP_S increased by nearly 1.43%, as shown in Fig. 12. Fig. 12 also shows that the exergy destruction decreased with the TIP. As already discussed, the exergy efficiency increased with the TIP and, consequently, the exergy destruction decreased with the TIP. The highest and lowest exergy were obtained with the working fluids R134a and R227ea, respectively, although the exergy destruction for fluids R236fa, R245fa and R1234ze was found between these two. Further, comparing Fig. 10 with Fig. 11, the performance of the PTCs combined system performance was reduced. It does not mean that the solar system is worse overall, but only from the efficiency point of view.

3.5 Performance evaluation with condenser temperature

The condenser temperature also has an impact on both subsystems. As the condenser temperature increased, this led to a slight decrease in the exergy and thermal efficiency of the cogeneration system (SORC–VCC), as shown in Fig. 13. Improvement in the condenser temperature reduced the enthalpy drop in the turbine leading to the lower output work. Therefore, the thermal efficiency of the SORC and COP_{vcc} system was reduced. Consequently, the thermal

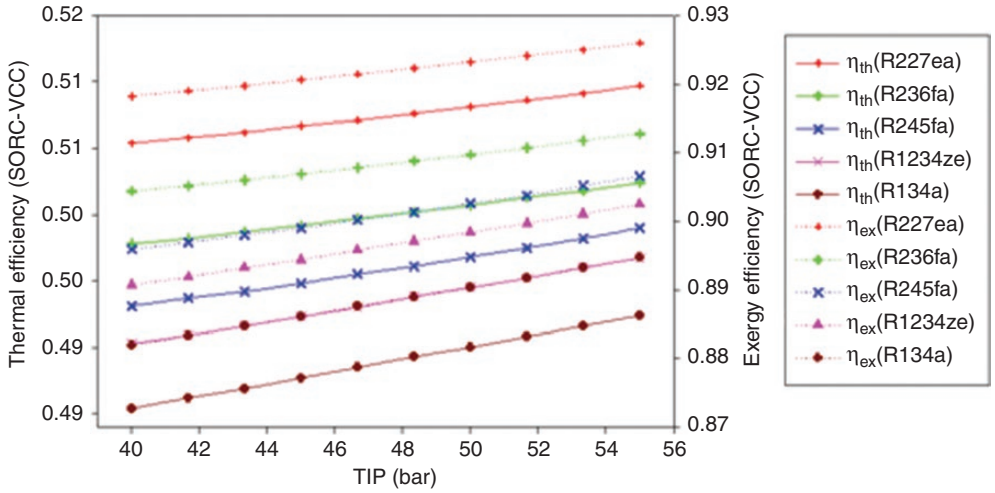


Fig. 10: Variation in the efficiency of the SORC–VCC cogeneration system with the TIP

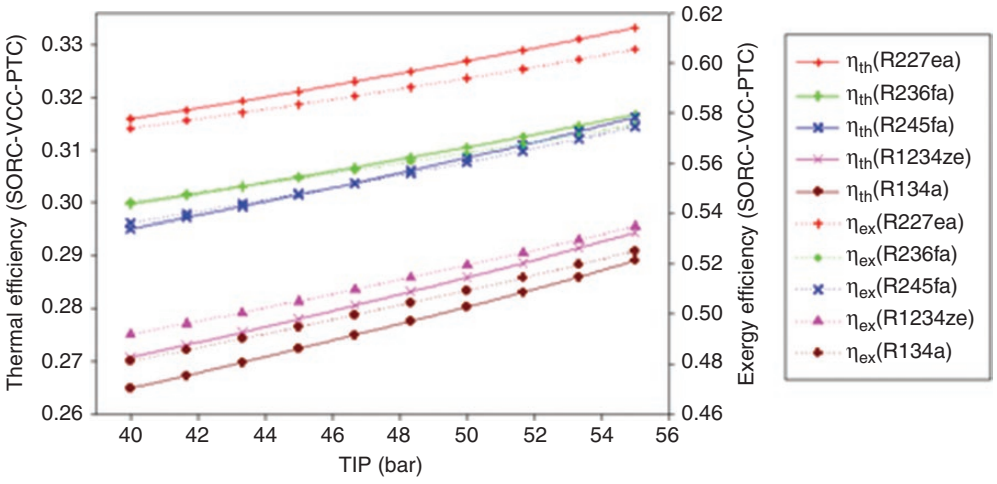


Fig. 11: Variation in the efficiency of the overall system (SORC–VCC–PTC) with the TIP

Downloaded from https://academic.oup.com/ce/article/5/3/476/6375889 by Delhi Technological University user on 06 October 2021

performance of the cogeneration system was reduced. The exergy efficiency of the cogeneration system also decreased with the condenser temperature because, according to Equation (12), the exergy efficiency is directly related to the thermal efficiency, as revealed in Fig. 13. Furthermore, the thermal efficiency of the overall system also depends on the condenser temperature. As the temperature of the condenser increased from 303 K to 328 K, the thermal efficiency decreased by almost 14.07%, as shown in Fig. 14. This was due only to the performance of the SORC–VCC system, not to that of the PTC system. The effect of the condenser temperature on the exergy efficiency of the overall system (SORC–VCC–PTC) was discussed. The exergy efficiency also decreased with the condenser temperature. It is already known that thermal efficiency decreases with the temperature of the condenser. Hence the exergy efficiency also decreased with the condenser temperature by Equation (12). The efficiencies of the other fluids are between those of the R227ea and R134a fluids, as shown in Fig. 14.

Finally, in this section, the impact of the condenser temperature variation on the system COP_S was also discussed, as shown in Fig. 15. In the case of R134a, as the condenser temperature varied from 303 K to 328 K, the COP_S decreased by 49.8%. It is already known that the η_{SORC} decreases with the condenser temperature. However, the temperature, pressure and enthalpy increase with the condenser temperature at the outlet of the compressor. This results in a reduction in the COP_{vcc}. As a result, the COP_S decreases according to Equation (19). R134a and R227ea show the highest and lowest COP_S, respectively. It can also be seen in Fig. 15 that if this system only works for the purpose of cooling, then R134a is selected as the appropriate working fluid among other selected working fluids. The exergy destruction increases with the condenser temperature. The exergy efficiency slightly increases with condenser temperature, therefore the exergy decreases with the condenser temperature. The highest exergy destruction was found with R134a, followed by R1234ze, R245fa, R236fa and R227ea of 2347, 2176, 2061, 1932 and 1644 kW, respectively, as displayed in right axis of Fig. 15.

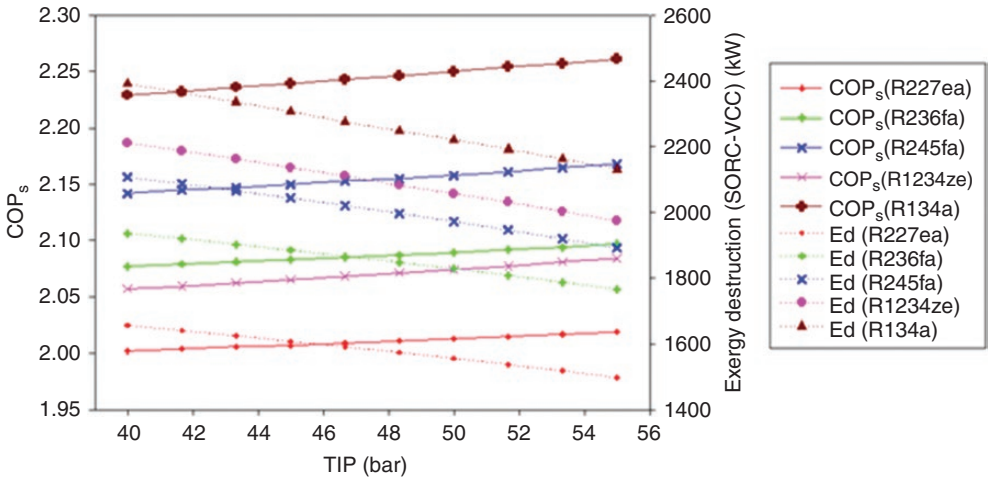


Fig. 12: Variation in the COP_s with the TIP

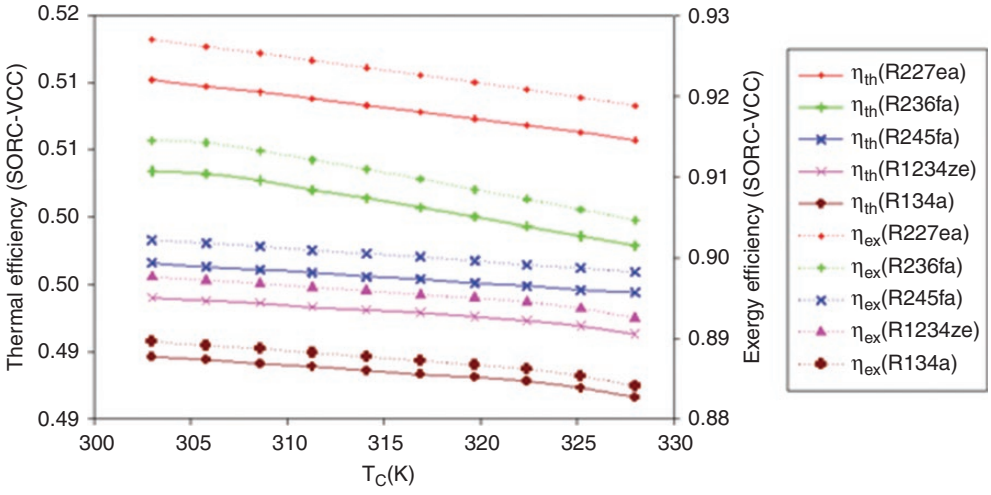


Fig. 13: Variation in the efficiency of the SORC–VCC cogeneration system with the condenser temperature

Downloaded from https://academic.oup.com/ce/article/5/3/476/6375889 by Delhi Technological University user on 06 October 2021

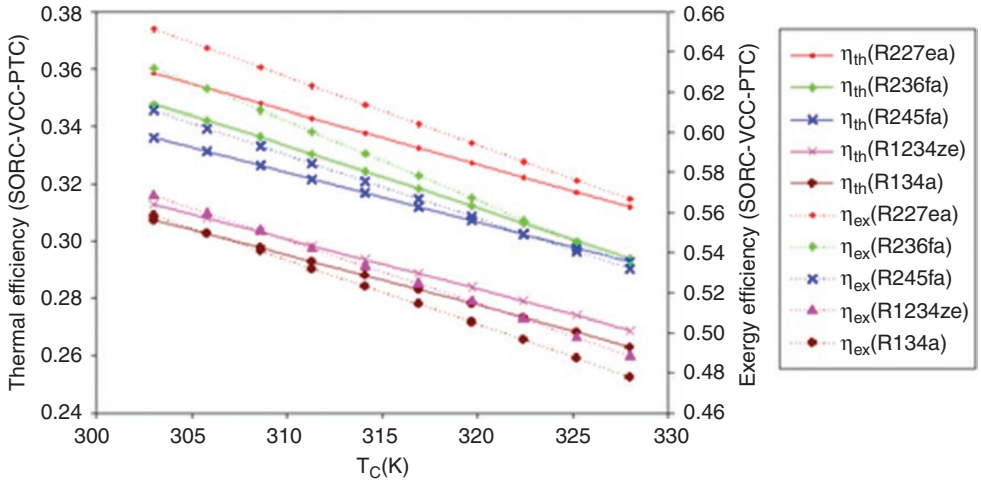


Fig. 14: Variation in the efficiency of the overall system (SORC–VCC–PTC) with the condenser temperature

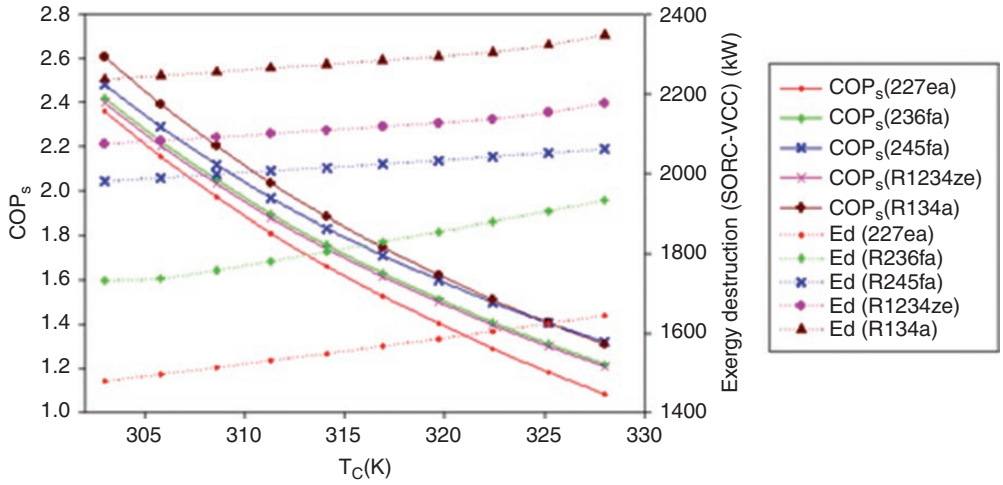


Fig. 15: Variation in the COP_s with the condenser temperature

3.6 Performance evaluation with evaporator temperature

Furthermore, the effect of the evaporator temperature was also examined. The cooling capacity is directly affected by the evaporator temperature. During examining the evaporator-temperature effect on the system performance, all other assumed parameters are kept constant, as listed in Table 1 such as the condenser temperature and pressure, evaporator pressure, etc. Fig. 16 reveals that the efficiency is not significantly affected by the evaporator temperature because there was no impact of the evaporator temperature on the SORC subsystem performance. The working fluids R227ea and R134a gave the highest and lowest efficiency, respectively. Further to Fig. 16, it is noticed that the evaporator temperature also did not significantly affect the exergy efficiency of the cogeneration system because the exergy efficiency has as direct relation with the thermal efficiency [31]. However, the COP_S was affected by the evaporator temperature significantly and it increased with the evaporator temperature. The reason behind this is that the COP_{vcc} increases with the evaporator temperature [17]. Also, the saturation pressure of the evaporator improves with the improvement of the evaporator temperature for all working fluids leading to a decline in the compressor work with constant condenser temperature. Conversely, the refrigeration effect improves with the evaporator temperature. Both impacts enhance the COP_{vcc}. However, simultaneously, the η_{SORC} is not affected by the evaporator temperature. As per Equation (19), this resulted in improvement in the COP_S, as shown in Fig. 17. Also, R134a and R227ea showed the highest and lowest values, respectively. Regarding the exergy destruction of the cogeneration cycle, the exergy destruction also was not affected much by the evaporator temperature like the efficiencies, as discussed above. At a constant evaporator temperature, the highest and lowest exergy destruction was for R134a and R227ea, respectively, whereas the exergy destruction for the R245fa was significantly affected due to different thermophysical properties of R245fa, as shown in the right axis of Fig. 17.

Downloaded from https://academic.oup.com/ce/article/5/3/476/6375889 by Delhi Technological University user on 06 October 2021

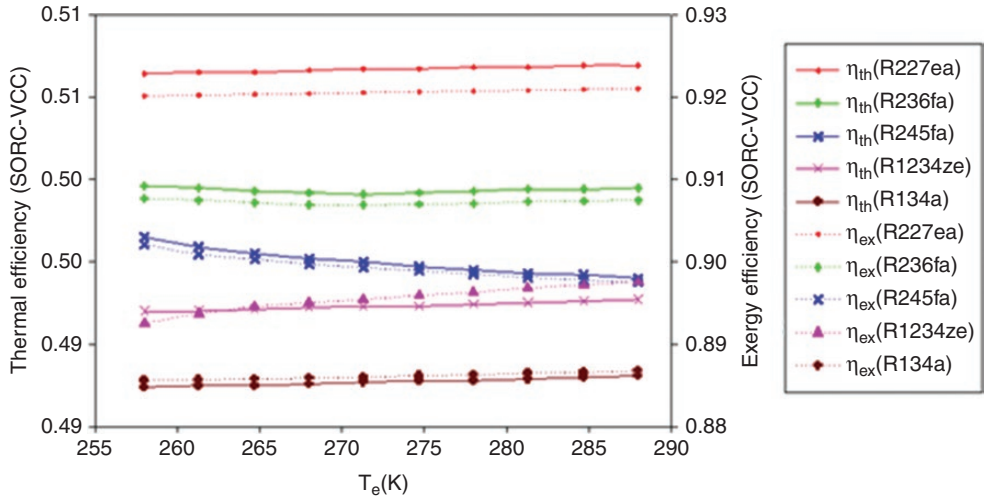


Fig. 16: Variation in the efficiency of the SORC–VCC cogeneration system with the evaporator temperature

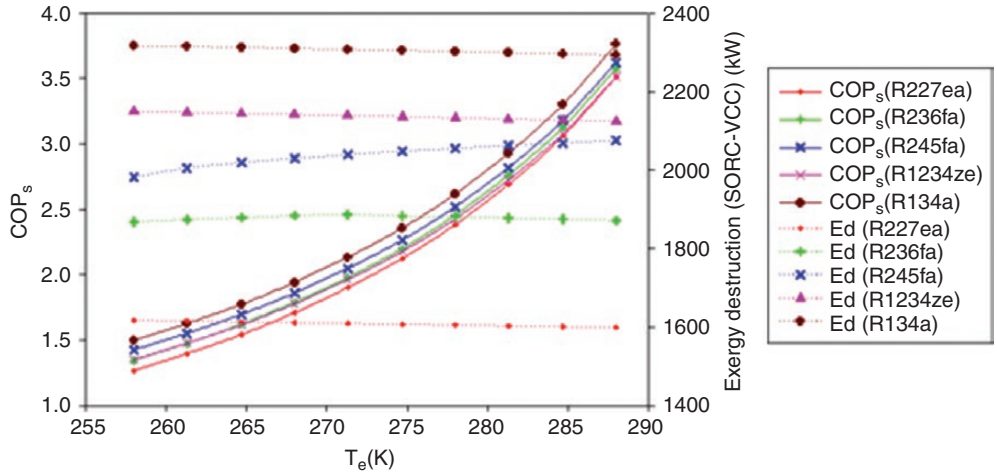


Fig. 17: Variation in the COP_s with the evaporator temperature

3.7 Exergy destruction in each component

In addition to the above parameters, the exergy destruction for each component and for each working fluid has been also evaluated, as revealed in Table 8. On the basis of R134a, a major part of the energy destruction rate was found in the PTCs followed by the turbine and heat exchanger with values of 8067, 1787 and 312.8 kW, respectively. It can be seen that PTCs alone accounted for 76.32% of the total exergy destruction of the overall system. The total exergy input into the cogeneration system was calculated at ~20 250 kW. It is therefore shown that 39.83% of the total inlet exergy was destroyed in the solar field only. This is due to the high exergy losses in the PTCs as well as the high outlet and inlet temperature difference of the HTF stream of the collector. Therefore, the PTC is the worst subsystem in terms of exergetic performance in the current model. Therefore, it is necessary to design the PTC system carefully. It can be also observed from Table 8 that the expansion valve was responsible for the lowest exergy-destruction rate among the other components and that it was ~3.695 kW only. Among the other considered working fluids, R227ea showed the lowest exergy-destruction rate. R227ea is superior because this has a more supercritical operating range due to its low critical values compared with other selected fluids, as listed in Table 2. The optimal parameters are summarized in Table 9.

Table 8: Exergy-destruction rate (kW) for each component

Components	R227ea	R236fa	R245fa	R1234ze	R134a
Solar field	6894	7265	7253	8006	8067
HX	218.3	309.1	311.8	288.5	312.8
Turbine	1092	1216	1361	1596	1787
Recuperator	46.3	52.62	26.99	24.21	17.24
Condenser	23.03	64.64	130.5	15.51	22.01
Mixer	181.7	191.9	172.5	175.9	167.6
Compressor	170.3	186.1	153.9	159.5	147.2
EX valve	2.937	3.088	3.109	3.509	3.695
Evaporator	15.36	21.41	7.783	24.9	27.54
ORC pump	16.47	10.87	18.31	19.19	17.61

Downloaded from https://academic.oup.com/ce/article/5/3/476/6375889 by Delhi Technological University user on 06 October 2021

Table 9: Summary of values of optimal parameters at 0.95 kW/m² solar irradiation

Parameters	Values
Optimum thermal efficiency	51.13%
Optimum exergy efficiency	92.9%
Optimum COP of system	2.27
Best working fluid for power generation	R227ea
Best working fluid for cooling	R134a

4 Conclusions

In the present study, an exergy and energy analysis of the PTC-driven SORC–VCC cogeneration system was conducted. The following conclusions were made from the results:

- In Mumbai, India, the exergy efficiency, thermal efficiency and COP of the SORC–VCC cogeneration system increased with solar irradiation and HTF velocity in the absorber tube, but declined with the solar-beam incidence angle. As a result, in order to increase the performance of the cogeneration system, the PTCs must be carefully designed.
- The maximum values of exergy efficiency, thermal efficiency and COP of the SORC–VCC cogeneration system were 92.9%, 51.13% and 2.27, respectively, using R227ea and R134a at 0.95 kW/m² of solar irradiation. At 0.95 kW/m², the overall system exergy efficiency and thermal efficiency were 62.78% and 34.5%, respectively.
- The exergy efficiency, thermal efficiency and COP_s decreased with the condenser temperature and increased with the TIP. The COP_s also increased continuously with the evaporator temperature and reached a maximum value of 3.764 at 328 K for R134a.
- The exergy destruction decreased with the solar irradiation and HTF velocity in the absorber tube, although it increased with the solar-beam incidence angle.
- For the working fluid R134a, the maximum total exergy-destruction rate was found to be ~8067 kW in a solar field. It contributed 76.32% of the total exergy-destruction rate of the system. It was also discovered that the solar field alone destroyed 39.83% of the input exergy.
- The overall system performance was reduced by considering the performance of the PTCs from an efficiency point of view. If the performance of the PTCs was not considered, then the cogeneration system performed well with solar energy.
- It was not necessarily true that a working fluid that was better for power generation was also better for cooling. Among the other working fluids considered, R227ea and R134a fared better for power generation and cooling, respectively.
- In future research, an exergoeconomic and thermoeconomic analysis should be performed for the current model. Practical applications of the current model will be also possible in those areas for simultaneous cooling and power generation where conventional heat sources are not present.

Nomenclature

$\dot{Q}$	Heat transfer (kW)	T_e	Evaporator temperature (K)
s	Specific entropy (kJ/kg-K)	η_{th}	Thermal efficiency
Ed	Exergy-destruction rate (kW)	$\dot{m}_{HTF}$	Mass flow rate of HTF (kg/s)
G_b	Solar irradiation (kW/m²)	T_c	Condenser temperature (K)
A_p	Aperture area (m²)	H	Specific enthalpy (kJ/kg)
T_0	Atmospheric temperatures	η_{ex}	Exergy efficiency
T	Temperature (K)	T_{su}	Temperature of the Sun (K)
$\dot{W}$	Work (kW)		
Subscripts			
am	ambient	i	Inlet
cond	condenser	o	Outside
comp	compressor	rec	Recuperator
evap	evaporator	e	Exit
s	solar	0	environmental conditions
VCC	vapour compression cycle		

Acknowledgements

Y.K. acknowledges the support of Department of Mechanical, Industrial & Production, Automobile Engineering of the Delhi Technological University, New Delhi, India.

Conflict of interest statement

None declared.

References

[1] Toujeni N, Bouaziz N, Kairaouani L. Energetic investigation of a new combined ORC-VCC system for cogeneration. *Energy Procedia*, 2017, 139:670–675.

[2] Bao J, Zhang L, Song C, *et al*. Comparative study of combined organic Rankine cycle and vapor compression cycle for refrigeration: single fluid or dual fluid? *Sustainable Energy Technologies and Assessments*, 2020, 37:100595.

[3] Khan Y, Mishra RS. Performance analysis of solar driven combined recompression main compressor intercooling supercritical CO_2 cycle and organic Rankine cycle using low GWP fluids. *Energy and Built Environment*, 2021. doi: 10.1016/j.enbenv.2021.05.004.

[4] Khan Y, Mishra RS. Thermo-economic analysis of the combined solar based pre-compression supercritical CO2 cycle and organic Rankine cycle using ultra low GWP fluids. *Thermal Science and Engineering Progress*, 2021, 23:100925.

[5] Kalogirou SA. A detailed thermal model of a parabolic trough collector receiver. *Energy*, 2012, 48:298–306.

Downloaded from https://academic.oup.com/ce/article/5/3/476/6375889 by Delhi Technological University user on 06 October 2021

[6] Al-Sulaiman FA, Hamdullahpur F, Dincer I. Performance assessment of a novel system using parabolic trough solar collectors for combined cooling, heating, and power production. *Renewable Energy*, 2012, 48:161–172.
[7] Al-Sulaiman FA. Exergy analysis of parabolic trough solar collectors integrated with combined steam and organic Rankine cycles. *Energy Conversion Management*, 2014, 77:441–449.
[8] Singh H, Mishra RS. Performance evaluation of the supercritical organic Rankine cycle (SORC) integrated with large scale solar parabolic trough collector (SPTC) system: an exergy energy analysis. *Environmental Progress & Sustainable Energy*, 2018, 37:891–899.
[9] Singh H, Mishra RS. Performance analysis of solar parabolic trough collectors driven combined supercritical CO_2 and organic Rankine cycle. *Engineering Science and Technology, an International Journal*, 2018, 21:451–464.
[10] Khan Y, Mishra RS. Parametric (exergy-energy) analysis of parabolic trough solar collector-driven combined partial heating supercritical CO_2 cycle and organic Rankine cycle. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 2020. doi: 10.1080/15567036.2020.1788676.
[11] Al-Zahrani AA, Dincer I. Energy and exergy analyses of a parabolic trough solar power plant using carbon dioxide power cycle. *Energy Conversion and Management*, 2018, 158:476–488.
[12] Song X, Chen X, Qi L, *et al*. Analysis of a supercritical organic Rankine cycle for low grade waste heat recovery. *Proceeding of the Institution of Civil Engineering-Energy*, 2020, 173:3–12.
[13] Kalra C, Becquin G, Jackson J, *et al*. High-potential power cycles & working fluids for next generation binary supercritical organic Rankine cycle for enhanced geothermal systems. In: *Thirty-Seventh Workshop on Geothermal Reservoir Engineering*, 30 January–1 February 2012, Stanford, California, USA.
[14] Moloney F, Almatrafi E, Goswami DY. Working fluid parametric analysis for recuperative super critical organic Rankine cycle for medium geothermal reservoir temperatures. *Renewable Energy*, 2020, 147:2874–2881.
[15] Saadon S, Islam SMS. Performance analysis of supercritical organic Rankine cycle system with different heat exchanger design configuration. *Journal of Advance Research in Fluid Mechanics and Thermal Sciences*, 2020, 65:324–333.
[16] Pektezel O, Acar HI. Energy and exergy analysis of combined organic rankine cycle-single and dual evaporator vapor compression refrigeration cycle. *Applied Science*, 2019, 9:5028. doi: 10.3390/app9235028.
[17] Saleh B. Energy and exergy analysis of an integrated organic Rankine cycle-vapor compression refrigeration system. *Applied Thermal Engineering*, 2018, 141:697–710.
[18] Javanshir N, Mahmoudi SMS, Rosen MA. Thermodynamic and exergoeconomic analyses of a novel combined cycle comprised of vapor-compression refrigeration and organic Rankine cycles. *Sustainability*, 2019, 11:3374.
[19] Hu B, Bu X, Ma W. Thermodynamic analysis of a Rankine cycle powered vapor compression ice maker using solar energy. *ScientificWorldJournal*, 2014;2014:742606.
[20] Moles F, Navarro-Esbrí J, Peris B, *et al*. Thermodynamic analysis of a combined organic Rankine cycle and vapor compression cycle system activated with low temperature heat sources using low GWP fluids. *Applied Thermal Engineering*, 2015, 87:444–453.
[21] Li H, Bua X, Wang L, *et al*. Hydrocarbon working fluids for a Rankine cycle powered vapor compression refrigeration system using low-grade thermal energy. *Energy and Buildings*, 2013, 65:167–172.
[22] Dudley VE, Koib GJ, Mahoney AR, *et al*. *Test Results: SEGS LS-2 Solar Collector. SANDIA94-1884*. Albuquerque, NM: Sandia National Laboratories, 1994 .
[23] Klein SA. Engineering Equation Solver (EES), Academic Commercial V7.714. F-Chart Software, www.fChart.com, 2020. https://fchartsoftware.com/ees/index.php/ (26 July 2021, date last accessed).
[24] Cengel YA, Boles MA. *Thermodynamics: An Engineering Approach*, 5th edition. New York, NY: McGraw-Hill, 2005.
[25] Khan Y, Mishra RS. Performance evaluation of solar based combined pre-compression supercritical CO_2 cycle and organic Rankine cycle. *International Journal of Green Energy*, 2021, 18:172–186. doi: 10.1080/15435075.2020.1847115.
[26] Petela R. Exergy analysis of the solar cylindrical-parabolic cooker. *Solar Energy*, 2005, 79:221–233.
[27] Mwesigye A, Yılmaz IH, Meyer JP. Numerical analysis of the thermal and thermodynamic performance of a parabolic trough solar collector using SWCNTs-Therminol®VP-1 nano fluid. *Renewable Energy*, 2018, 119:844–862. doi: 10.1016/j.renene.2017.10.047.
[28] Kose O, Koc Y, Yagli H. Performance improvement of the bottoming steam Rankine cycle (SRC) and organic Rankine cycle (SORC) systems for a triple combined system using gas turbine (GT) as topping cycle. *Energy Conversion and Management*, 2020, 211:112745.
[29] Feidt M, Kheiri A, Pelloux-Prayer S. Performance optimization of low-temperature power generation by supercritical ORCs (organic Rankine cycles) using low GWP (global warming potential) working fluids. *Energy*, 2014, 67:513–526.
[30] Krishnan BP, Vijayan R, Gokulnath K, *et al*. Experimental analysis of a vapour compression refrigeration system by using nano refrigerant (R290/R600a/Al_2O_3). *AIP Conference Proceedings*, 2019, 2128:050023. doi: 10.1063/1.5117995.
[31] Duffie J, Beckman W. *Solar Engineering of Thermal Processes*. 4th edn. Hoboken, NJ: John Wiley & Sons Inc., 2013.

Downloaded from https://academic.oup.com/ce/article/5/3/476/6375889 by Delhi Technological University user on 06 October 2021

2021 8th International Conference on Signal Processing and Integrated Networks (SPIN) | 978-1-6654-3564-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SPIN52536.2021.9566049

PM-BLDC Motor Drive for Solar Water Pumping with Modified Boost Converter

Shivangi Yadav
Dept. of Electrical Engineering
Delhi Technological University
Delhi, India
yadavshivangi26@gmail.com

Saurabh Mishra
Dept. of Electrical Engineering
Delhi Technological University
Delhi, India
saurabh.mishra1410@gmail.com

Garima
Dept. of Electrical Engineering
Delhi Technological University
Delhi, India
garima.dtu@gmail.com

***Abstract*— Solar Photovoltaic (SPV) systems are being used nowadays for a number of applications and one such application which is attracting attention is solar water pumping. Solar water pumping has become popular in rural areas for the purpose of farm irrigation, supplying drinking water to households, fish farms etc. A solar photovoltaic water pumping system with a modified boost DC-DC converter has been demonstrated here. The DC-DC converter serves as a link between the load and the PV array. Maximum Power Point Tracking (MPPT), which tracks maximum PV power, causes the SPV system to operate efficiently even in varying atmospheric conditions. In order to supply switching pulse for DC-DC converter, Perturb and Observe algorithm is used. The permanent magnet brushless DC motor drives the centrifugal pump linked to the shaft in the SPV water pumping system. PM-BLDC's Electronic Commutation is implemented to provide switching pulses for the operation of the voltage source inverter (VSI). The modelling of BLDC motor drive for solar PV powered system for water pumping, is done in MATLAB/Simulink and has been examined under dynamic conditions.**

Keywords— SPV Water Pumping System, Modified Boost DC-DC Converter, Perturb and Observe algorithm(P&O), PM-BLDC Motor, Centrifugal Pump

I. INTRODUCTION

The prediction of energy crisis in the future caused as a result of excessive utilization of fossil fuels for energy generation has led to a global shift towards the renewable sources of energy. Renewable energy offers a more efficient and cost-effective procedure for the generation of electricity. Countries all across the world are focusing on making the most use out of solar, wind, tidal, geothermal and other available renewable sources of energy. The benefits of solar energy over any other renewable source are innumerable. The reduction in the photovoltaic modules prices, as well as the power electronic converters and the microprocessors offers an added advantage for the solar PV systems. Solar PV systems are being utilized for a large number of applications including on-grid systems, standalone systems and also the solar water pumping systems which has lately gained attention and popularity. Solar water pumping seems to be a promising option as it is widely useful for farm irrigation, supplying drinking water to households, street watering in urban regions.

A SPV based system used for pumping water typically constitutes of a solar array, DC-DC converter, DC-AC converter, motor and a water pump. A wide variety of DC-DC converter are being employed in solar PV systems, some of which have been discussed in [1,2]. Apart from typical boost, buck-boost DC-DC converters, some of its topologies are also being developed. Several High-Gain converters are also under study for different solar applications because of their advantages over conventional step-up converters [3-7]. In this study one such modified form of boost DC-to-DC converter is interfaced to the solar fed water pumping system. This converter unlike their conventional counterparts, enable high efficiency operation of the switches without putting them under high current and voltage stresses. Motors are another very essential components of a water pumping system. In our system we have used a permanent magnet brushless DC motor. The DC motor with brushes need more maintenance than BLDC due to the existence of brushes and commutator. Therefore, PM-BLDC motor has replaced other motors owing to its lower maintenance requirement, reliability, efficiency and simpler control mechanism [8].

The PV cell is a non-linear source of energy therefore, tracking of power to its maximum value is essential to derive the utmost power under any environmental condition from the array. The power being produced by the PV array depends on elements like the irradiance and temperatures. So, to build on the efficiency of the system it is imperative to maintain the power extracted irrespective of the environmental conditions. This purpose is served by the MPP tracking algorithm. The converter's duty cycle is tuned using this MPPT algorithm that enables a constant amount of power to get extracted from the solar array. The tracking of PV array's point of maximum power by MPPT, improves its efficiency and thereby of the entire system. Some methods for tracking the maximum power by (MPPT) are discussed in [9,10].

Centrifugal water pump which is the load for PV system is driven with the help of BLDC. The motor shaft is connected to the centrifugal pump. The supply to the BLDC comes from the inverter. The BLDC motor's electronic commutation, functions to provide switching sequence to the VSI [11,12]. The BLDC motor is different from the other DC motors as it does not have commutator and brushes. This is the reason they are said to be electronically commutated motor. Electronic commutation is basically an encoding and decoding process. The encoder, based on the rotor position provides signals which are then decoded to supply switching for the inverter switches. The response of the BLDC motor drive powered by solar PV, is simulated in MATLAB/ Simulink.

The following is the breakdown of paper's structure. Section II explains the system structure. Section III elaborates the design procedure of different components of SPV water

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:41:34 UTC from IEEE Xplore. Restrictions apply.

pumping followed by section IV showing the control strategy for the two stages of the system. Section V discusses the simulation results for the water pumping system proceeded by the conclusion in section VI.

II. SYSTEM STRUCTURE

The system under our study which is the solar fed water pumping system comprises of four important components. These components include the PV array which provide supply to the DC-DC converter. The converter serves two purposes, one is to step-up the voltage level as per requirement of the application and second is to operate with the MPPT algorithm to improve the overall efficacy. Adjustments in the duty ratio of the DC-DC converter is carried out by P&O algorithm, with changes in irradiance. For our system we will be considering P&O algorithm. The next stage in the system includes the DC-AC voltage source inverter (VSI). The inverter is fed with dc link voltage which is controlled by DC-DC converter. The switching pulse for turning ON/OFF the VSI switches is obtained from the process of electronic commutation occurring in the BLDC motor, which is the load to the system. In electronic commutation, the encoder provides the signals that are decoded to obtain the switching of the inverter. At last is the BLDC motor which is connected to the centrifugal pump through its shaft for pumping water. Fig.1 demonstrates the block representation of the system proposed.

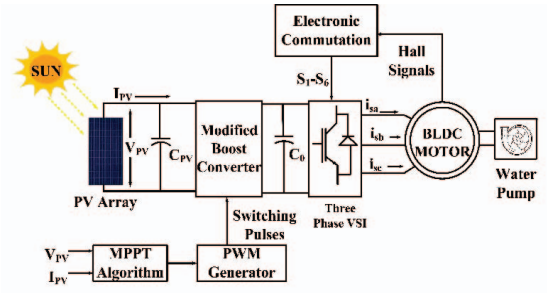


Fig.1. Block Representation [7]

III. DESIGN PROCEDURE

The array rating for the system considered is taken to be 3 kW and that of BLDC motor is 2.5 kW. Solar array's rating is chosen at a value slightly higher to compensate for the different losses linked to DC-DC conversion stage. The designing of various parts of the solar PV water pumping system have been discussed below.

A. PV array

As mentioned earlier, the DC-DC converter stage involves losses that have been taken into account while considering the solar array ratings. The power capacity of solar module is taken as 3 KW. Solar PV Array's other specifications are mentioned in table I.

TABLE I. PV ARRAY PARAMETERS

DESCRIPTION	RATING
Power at MPP [Pmax]	250 Watts
Short-Circuit Current [I_{SC}]	8.8 Amps
Open-Circuit Voltage [V_{OC}]	36.5 V
Current at MPP [I_{MP}]	8.2 Amps
Voltage at MPP [V_{MP}]	30 V
Cells in series	4
Cells in parallel	3

For conventional test circumstances (1000 W/m^2 and 25^0 C), the solar arrays P-V and I-V waveforms are presented in Fig. 2.

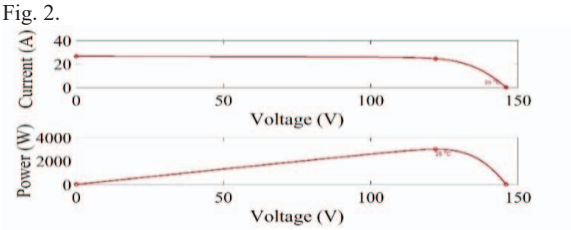


Fig.2 I-V and P-V Curves

B. DC-DC Converter

This stage of conversion, which is DC-DC receives the electricity generated by the solar PV array. Converter (DC-DC) considered for our system is a modification of boost converter which can provide a gain value as high as ten times to that of the input voltage. It comprises of two inductors, two capacitors, three diodes and three switches. Fig.3 below depicts the converter structure.

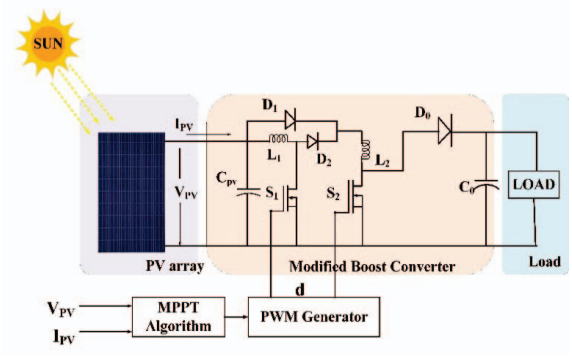


Fig.3. Topology of DC-DC Converter [5]

The switches in the converter are turned ON and OFF together. The modes of operation for the suggested DC-DC converter are explained in detail below.

Mode I: The two switches present in the converter are in the ON state during this mode. Whereas the diodes D_0, D_2 are in the OFF state. Therefore, the inductors L_1 and L_2 get charged during this mode by the array voltage.Fig.4 shows the operation of mode I.

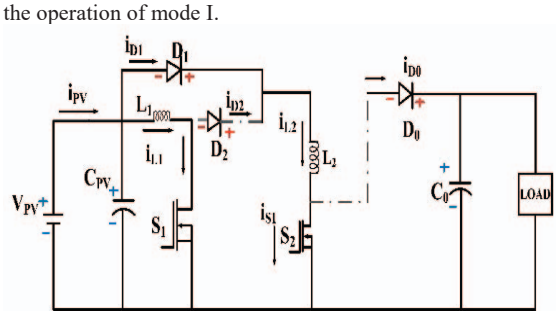


Fig. 4 Mode 1 of Operation [5]

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:41:34 UTC from IEEE Xplore. Restrictions apply.

Mode II: The two switches are switched OFF in this mode along with diode D_1. The diode D_2 and D_0 are in the ON state and they provide path to the inductor currents. The stored energy in the inductors during Mode I is discharged in this mode through capacitor C_0. Fig.5 presents the operation of mode II.

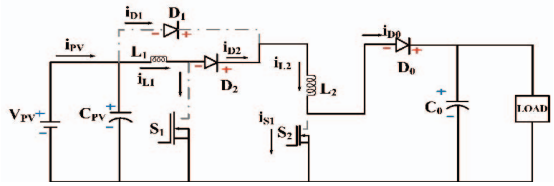


Fig.5 Mode II of operation [5]

The design equations for the calculating various components of the converter are given as follows:

$$V_{PV} = 120\ V \quad (1)$$

$$V_{dc} = 150\ V \quad (2)$$

$$\frac{V_{dc}}{V_{PV}} = \frac{1+d}{1-d} \text{ so }, d = 0.11 \quad (3)$$

$$f = 5\ \text{kHz} \quad (4)$$

$$\Delta i_L = 3\% \quad (5)$$

$$\Delta V_0 = 5\% \quad (6)$$

$$L_1 = L_2 = \frac{V_{PV}}{2*\Delta i_L} * \frac{d}{f} = \frac{120*0.11}{2*0.03*5000} = 44\ mH \quad (7)$$

$$C_{PV} = C_0 = \frac{V_{dc}}{2*R*\Delta V_0} * \frac{d}{f} = \frac{150*0.11}{2*7.5*0.05*5000} = 4000\ \mu F \quad (8)$$

Where V_{PV} is voltage of array, V_{dc} is voltage at the DC output, d is duty ratio, f is frequency, Δi_L is allowed ripple in inductor current, ΔV_0 is voltage ripple, L_1 and L_2 are the inductance values, C_{PV} and C_0 are the capacitance values.

C. Permananet- Magnet Brushless DC Motor(PM-BLDC)

Brushless DC (BLDC) motor is now being broadly operated for solar fed water pumping requirements due to its advantage over the AC motor and DC motor with brushes. It requires lesser maintenance as the brushes and commutator are eliminated. The BLDC motor used for our system has a power rating of 2.5 kW. Other BLDC motor parameters are mentioned in table II.

TABLE II. BLDC MOTOR SPECIFICATIONS

Description	Values
Power (P_m)	2.5 kW
Poles	6
Moment of Inertia (J)	8 kg.m^2
Speed (N)	3000 rpm
Resistance of stator (R_s)	2.3 ohms
Inductance of stator (L_s)	0.05 mH
Torque Constant (K_t)	0.17 N-m/A
Voltage Constant (K_e)	17.5 V/krpm

D. Centrifugal Water Pump

This pump is connected to the system through BLDC motor's shaft. The power developed by the BLDC motor is 2.5 kW. In order to find out the proportionality constant for the centrifugal pump we equation given below.

$$K = \frac{P}{\omega_r} = \frac{2.5*1000}{\left(2*\pi*\frac{3000}{60}\right)^3} = 8.06*10^{-5} \quad (9)$$

Where P is power rating for the BLDC motor, ω_r is the rotor speed in rpm.

IV. Control Strategy

Two different control techniques are used in a PV fed water pumping system. The first stage involves controlling the DC-DC converter by regulating its duty cycle and the second stage control technique is implemented for obtaining Voltage Source Inverter's switching signals. The two controlling techniques have been discussed in detail below.

A. Perturb and Observe Algorithm

For the efficient operation of the solar PV system, it is essential to derive maximum possible power irrespective of the environmental conditions. Therefore, continuous studies have been carried to make the tracking of MPP more and more reliable and efficient. The simple nature and easy to implement process for P&O algorithm makes it a popular control method for solar PV systems. P&O algorithm largely acts on the hit and trial technique to find the maximum power value. The voltage output of PV array is sensed and is varied by a small value. The algorithm then compares the power at consecutive steps. In case, if the power rises, the perturbations are continued in direction same as in previous step. Contrary to that, if the power decreases the direction of perturbations is made opposite, to track the maximum power point. The equations given below summarize the principle of operation of the Perturb and Observe algorithm.

$$\frac{dP_{PV}}{dV_{PV}} = 0 \quad \text{at MPP} \quad (10)$$

$$\frac{dP_{PV}}{dV_{PV}} > 0 \quad \text{MPP's Left Side} \quad (11)$$

$$\frac{dP_{PV}}{dV_{PV}} < 0 \quad \text{MPP's Right Side} \quad (12)$$

Fig.6 demonstrates the P&O algorithm flowchart.

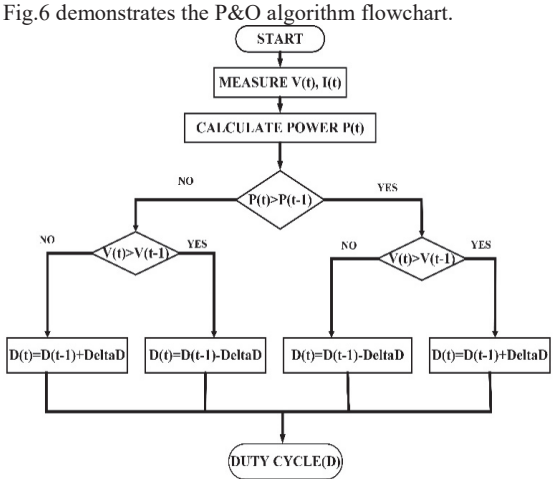


Fig.6. Flowchart of P&O Algorithm [10]

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:41:34 UTC from IEEE Xplore. Restrictions apply.

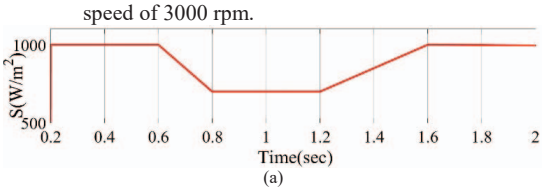
B. Electronic Commutation

The inverter stage utilizes BLDC motor's electronic commutation as the control technique. The function of Electronic Commutation is to provide the pulse sequence for the Voltage Source Inverter (VSI). In this technique, the encoder built in the BLDC motor produces hall effect signals depending on the position of the rotor. These encoded signals are then decoded in order to provide the controlling signals to turn ON the switches of the VSI. The position of rotor is sensed by the encoder in every 60º span thereby producing a specific set of hall effect signals. Accordingly, these signals are decoded and VSI's switching pulses are obtained. Therefore, this kind of commutation that occurs is an encoding and decoding process. Hall effect sensors utilize the position of rotor and encoder encodes them accordingly. To provide the pulses for the VSI switches, the signals are then decoded. This type of commutation reduces losses, as there are no moving parts involved.

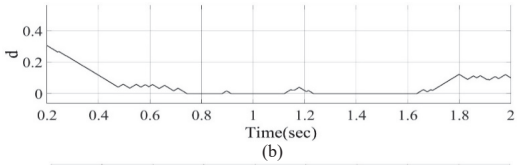
V. SIMULATED RESULTS

The water pumping system fed by solar PV depicted in Fig. 1 is implemented with MATLAB/Simulink. The system has been studied under dynamic conditions with varying irradiance levels within a time span of 2 s. The irradiance values are varied at time t=0.6s and t=1.2s, as presented in Fig.7(a). Following are some observations from the simulated results:

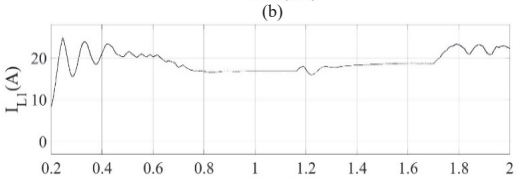
- Fig.7(a) shows that the irradiance value is constant at 1000W/m^2 till time t =0.6 s. After time t=0.6 s there is decline in irradiance value up to 700W/m^2. After time instance of t=1.2 s, the irradiance value follows a ramp function and rises to 1000 W/m^2.
- Fig.7(b) shows the d value of the converter (DC-DC) that is adjusted by the control algorithm(P&O) to draw out maximum power from the solar array.
- Fig.7(c)-Fig.7(f) presents the waveforms for the different elements of the converter (DC-DC). All these waveforms for current as well as voltage for different components can be seen to follow the variation in irradiance instantly.
- Fig.7(h) and Fig.7(i) shows the power curve for the array and the DC-DC converter. MPPT algorithm works efficiently to maintain power from the array. Rated power of approximately 3 kW can be seen to be drawn at maximum irradiance value of 1000 W/m^2.
- Pulsations are seen in the torque (T_e) waveform in Fig.7(m). These pulsations are caused as a result of commutating electronically.
- The speed of the motor also follows the irradiance curve, as it changes. With maximum irradiance value of 1000 W/m^2, it is able to achieve its rated speed of 3000 rpm.



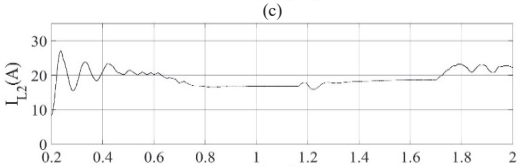
(a)



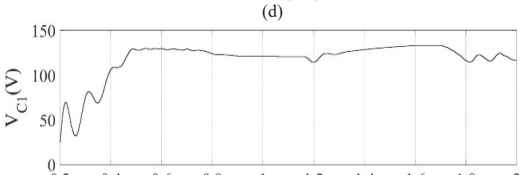
(b)



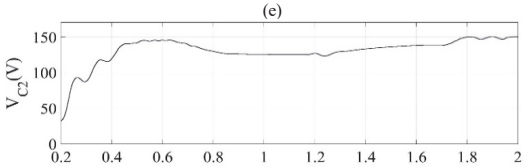
(c)



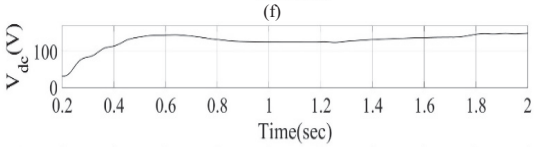
(d)

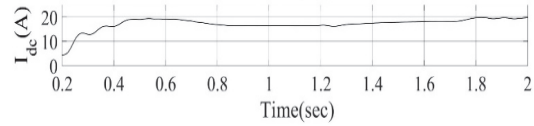


(e)

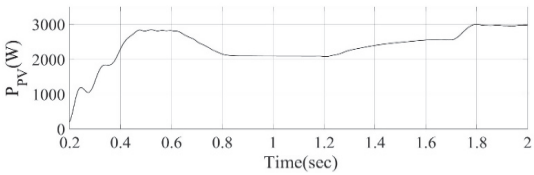


(f)

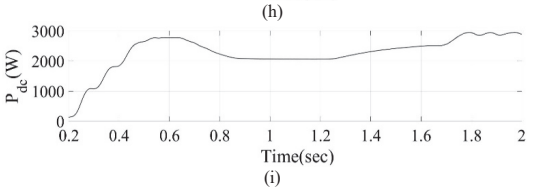




(g)



(h)



(i)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:41:34 UTC from IEEE Xplore. Restrictions apply.

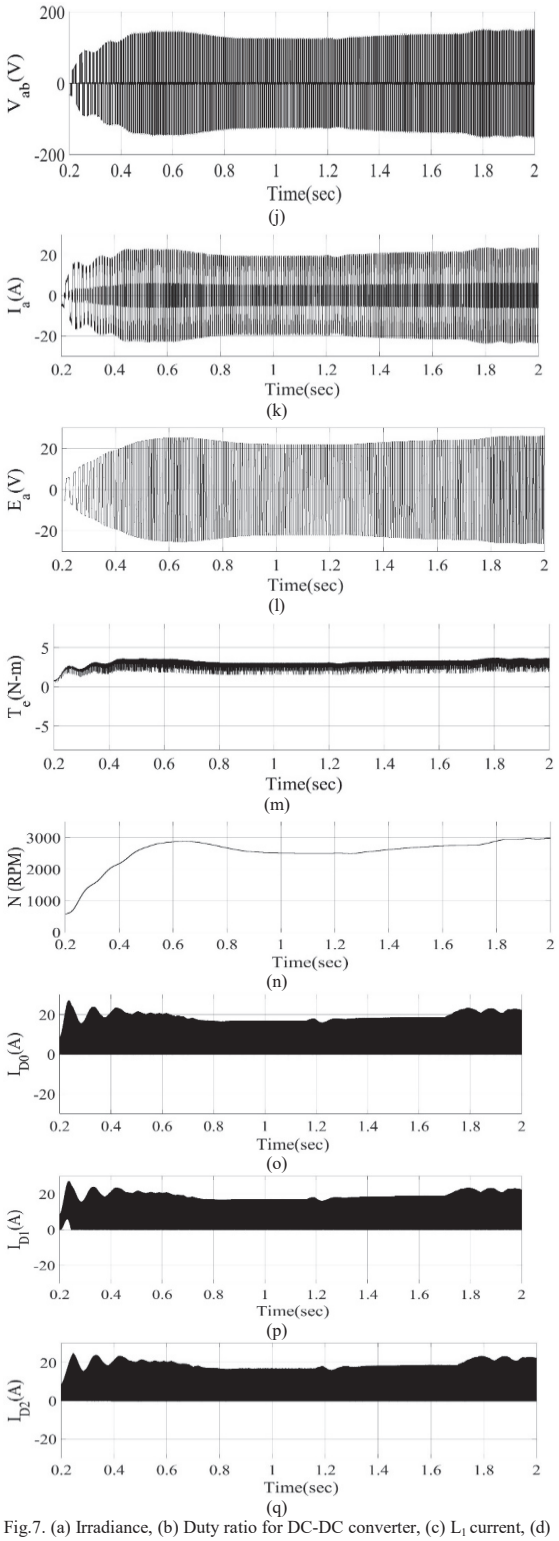
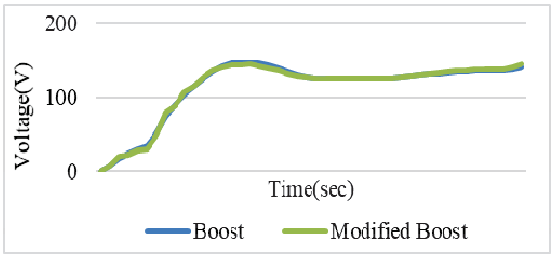
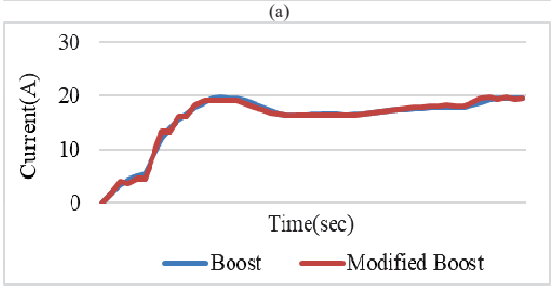


Fig.7. (a) Irradiance, (b) Duty ratio for DC-DC converter, (c) L_1 current, (d) L_2 current, (e) C_1 voltage, (f) C_2 voltage(g) Current and Voltage of the DC-DC conversion stage, (h) PV array power output, (i) Power output of DC-DC converter, (j) VSI output Voltage (k) Stator current, (l) Back-Emf, (m) Electromagnetic Torque, (n) Speed, (o)Diode D_0 current, (p) Diode D_1 current, (q) Diode D_2 current.

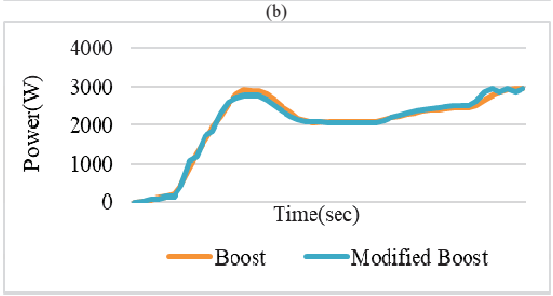
Fig.8(a) to Fig.8(g) represents the comparison of performance of boost as well as modified boost converter for solar PV powered water pumping system having equal rated values.



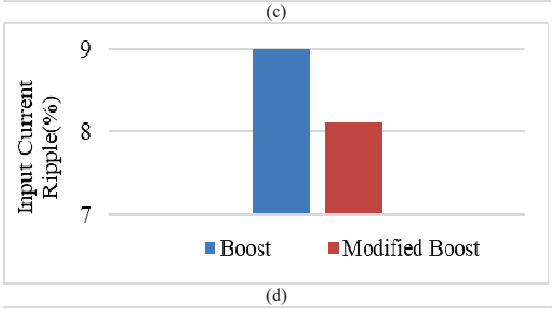
(a)



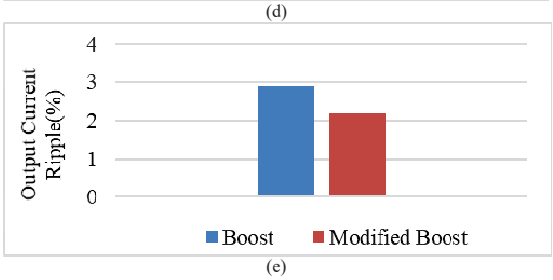
(b)



(c)

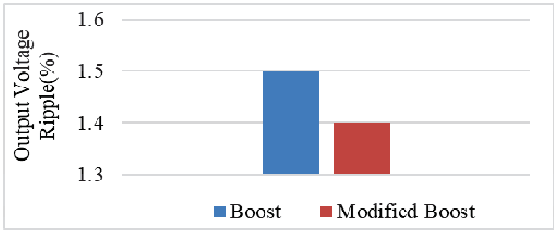


(d)

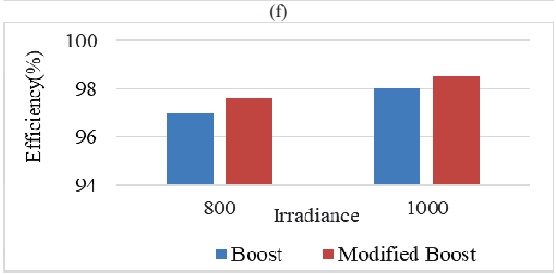


(e)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:41:34 UTC from IEEE Xplore. Restrictions apply.



(f)



(g)

Fig.8(a) Output Current curve for boost and modified boost converter, (b) Output Voltage curve for boost and modified boost converter, (c) Output power curve for boost and modified boost converter, (d) Input Current ripple comparison, (e) Output Current Ripple Comparison, (f) Output Voltage Ripple Comparison, (g)Efficiency Comparison at irradiance 800 W/m^2 and 1000 W/m^2.

The comparison between the Boost and Modified Boost converter shown in Fig.8(d)-Fig.8(g) proves that modified boost converter is a potential substitute for boost converter in solar fed water pumping systems. It has considerably good efficiency even under reduced levels of irradiance. The voltage and current ripples have also been reduced with the use of Modified Boost converter with the solar PV array.

VI. CONCLUSION

The modelling of the water pumping system fed by SPV consisting of Modified Boost converter is done in MATLAB/Simulink tool. The results validate the potential of the SPV water pumping structure. Modification of traditional converters can be used in place of conventional DC-DC converter for such requirements. Comparison has been made between the modified boost and conventional boost converter which shows that modified boost performs satisfactorily when interfaced to the SPV water pumping system. Boost converter shows a smoother transition with variation in irradiance than modified Boost. But, modified boost converter efficiently extracts the maximum power from the array and helps in minimizing the current and voltage stress on the switch unlike the boost converter. It also reduces the ripples in voltage and current. The waveforms in the results have proven that the P&O algorithm along with the DC-DC converter is capable of making the system efficient by drawing the peak power from the array. Inverter (VSI) is set to operate in 120^0 conduction mode which has helped in reducing the switching losses caused as a result of high frequency switching. The system responds well to dynamic environmental conditions like the variance in irradiance levels. Simulation results justify the working efficacy of the combination of the PM-BLDC motor solar water pumping system using modified boost converter.

ACKNOWLEDGMENT

Authors are thankful to the Electrical Engineering Department (EED) of Delhi Technological University (DTU). We are also thankful to SPIN 2021 for providing a platform to share our research study.

REFERENCES

[1] Dileep. G, S.N. Singh, “Selection of non-isolated DC-DC converters for solar photovoltaic system”,*Renewable and Sustainable Energy Reviews*,Volume 76,Pages 1230-1247,2017.

[2] V. B. Savakhande, C. L. Bhattar and P. L. Bhattar, "Voltage-lift DC-DC converters for photovoltaic application-a review," *International Conference on Data Management, Analytics and Innovation (ICDMAI),* pp. 172-176, 2017.

[3] M. Das, M. Pal and V. Agarwal, "Novel High Gain, High Efficiency DC–DC Converter Suitable for Solar PV Module Integration With Three-Phase Grid Tied Inverters," *in IEEE Journal of Photovoltaics*, vol. 9, no. 2, pp. 528-537, March 2019.

[4] P. K. Maroti, S. Padmanaban, J. B. Holm-Nielsen, M. Sagar Bhaskar, M. Meraj and A. Iqbal, "A New Structure of High Voltage Gain SEPIC Converter for Renewable Energy Applications," *in IEEE Access*, vol. 7, pp. 89857-89868, 2019.

[5] O. Abdel-Rahim and H. Wang, "A new high gain DC-DC converter with model-predictive-control based MPPT technique for photovoltaic systems," *in CPSS Transactions on Power Electronics and Applications*, vol. 5, no. 2, pp. 191-200, June 2020.

[6] A. E. Lucena da Costa and R. L. Andersen, "High-gain Boost-Boost-Flyback converter for renewable energy sources applications," *IEEE 13th Brazilian Power Electronics Conference and 1st Southern Power Electronics Conference (COBEP/SPEC),* 2015, pp. 1-6, 2015.

[7] R. Kumar and B. Singh, "Buck-boost converter fed BLDC motor drive for solar PV array based water pumping," *IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES),* pp. 1-6, 2014.

[8] R. B. Bollipo, S. Mikkili and P. K. Bonthagorla, "Hybrid, optimal, intelligent and classical PV MPPT techniques: A review," *in CSEE Journal of Power and Energy Systems*, vol. 7, no. 1, pp. 9-33, Jan. 2021.

[9] Y. Sun, S. Li, B. Lin, X. Fu, M. Ramezani and I. Jaithwa, "Artificial Neural Network for Control and Grid Integration of Residential Solar Photovoltaic Systems," *in IEEE Transactions on Sustainable Energy*, vol. 8, no. 4, pp. 1484-1495, Oct. 2017.

[10] A. Kchaou, A. Naamane, Y. Koubaa and N. K. M'Sirdi, "Comparative study of different MPPT techniques for a stand-alone PV system," *17th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA),* 2016.

[11] R. Kumar and B. Singh, "Solar PV array fed water pumping system using SEPIC converter based BLDC motor drive,"*Eighteenth National Power Systems Conference (NPSC)*, pp. 1-5,2014.

[12] R. Kumar and B. Singh, "Solar photovoltaic array fed Luo converter based BLDC motor driven water pumping system," *9th International Conference on Industrial and Information Systems (ICIIS)*, pp. 1-5.2014.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:41:34 UTC from IEEE Xplore. Restrictions apply.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/354863157

Realization of IIMC based Higher-Order Floating Fractional-Order Element

Conference Paper · May 2021

CITATIONS
0

READS
49

3 authors:



Garima Varshney
Delhi Technological University
8 PUBLICATIONS **6** CITATIONS
SEE PROFILE



Neeta Pandey
Delhi Technological University
289 PUBLICATIONS **1,678** CITATIONS
SEE PROFILE



Rajeshwari Pandey
Delhi Technological University
99 PUBLICATIONS **752** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Capacitance Multiplier View project

Fractional-Order Signal Processing and Signal Generating Circuits View project

All content following this page was uploaded by Garima Varshney on 27 September 2021.

The user has requested enhancement of the downloaded file.

Realization of IIMC based Higher-Order Floating Fractional-Order Element

Garima Varshney, Neeta Pandey and Rajeshwari Pandey

Department of Electronics & Communication Engineering, Delhi Technological University, Delhi, India
E-mail: gvarshn1@gmail.com, neetapandey@dce.ac.in, rpandey@dce.ac.in

Abstract: **The realization of floating fractional-order element (FOE) based upon operational transconductance amplifier (OTA) based IIMC is presented in this paper. The proposed circuit can be used to realize floating fractional order capacitor (FC) and floating fractional order inductor (FI) of order (n - 1 + α). The magnitude of the FOE can be altered using external bias currents to OTA. Functional verification of proposed circuit to realize FC and FI of order 1.5 is done using PSPICE simulation with 0.18µm CMOS technology parameters. The magnitude responses and phase responses of FC and FI of order 1.5 are presented in the manuscript.**

Index Terms—IIMC, OTA, Fractional-order element, Fractional-order inductor, Fractional-order capacitor.

I. INTRODUCTION

These-days, fractional order calculus is an expanding field for research, by virtue of it's inherit capability to describe a real phenomenon more precisely when compared to conventional integer-order techniques [1]. Fractional order systems are used in many fields of science in numerous applications, for example: control systems [2], electrical engineering [3], analog signal processing [4], biomedical [5], communication [6] and many more. Consequently, at interdisciplinary level fractional-order systems have huge opportunity for research. For a fractional-order system, the elementary component is a FOE. The FOEs with fractional order greater than 1 are beneficial in designing higher order fractional-order systems. A thorough comparison of FOEs realized using various techniques, such as, emulated, on chip fabricated, floating or grounded or electronic tunability of order is presented in Tab. I. From Tab. I it is clear that limited literature is available for floating FOEs with fractional order grater than 1, with electronic tunability. In this manuscript, realization of floating fractional-order element (FOE) based upon OTA based IIMC [16] is presented. The proposed circuit can be used to realize floating FC and floating FI of order (n - 1 + α). Also, the magnitude of the FOE can be altered using external bias currents to OTA.

This manuscript is organised as follows: In Sect. 2, basics of OTA and FOE are given. The analysis of proposed floating FOE is given in Sect. 3. Section 4 talks about PSPICE simulations and lastly conclusion is given in Sect. 5.

TABLE I

DETAILED COMPARISON OF FOES

Ref.	Active block	Order of FOE	Electronic tuning	Floating/ Grounded
[7]	OTA	$0<\alpha<1$	Yes	Floating
[8]	Fabricated	$0<\alpha<2$	No	Floating
[9]	Passive RC and Op-Amp based GIC	$0<\alpha<2$	No	Floating
[10]	Passive RC and Op-Amp based GIC	$0<\alpha<2$	No	Floating
[11]	Passive RC and Op-Amp based GIC	$0<\alpha<2$	No	Floating
[12]	Passive RC	$0<\alpha<1$	No	Floating
[13]	OTA	$-1<\alpha<1$	Yes	Floating
[14]	Fabricated	$0<\alpha<1$	No	Floating
[15]	CFOA	$0<\alpha<1$	Yes	Grounded
[16]	OTA	$0<\alpha<(n-1+\alpha)$	Yes	Grounded
Proposed	OTA	$0<\alpha<(n-1+\alpha)$	Yes	Floating

II. PRELIMINARIES

A. Dual-output OTA

The OTA processes difference of input voltage and provides current at low impedance output node. Fig. 1 shows the block diagram of OTA and the input-output port equation is given by (1).

$$I_{o\pm} = \pm g_m(V_+ - V_-) \quad (1)$$

Here g_m depends upon bias current I_b and called as transconductance gain of OTA. The dependence of gm upon Ib is shown in (2)

$$g_m = \sqrt{\mu_n C_{ox} \frac{W}{L} I_b} \quad (2)$$

The symbols μ_n, C_{ox}, W/L and I_b have their usual meaning.

B. Fractional-order Element (FOE)

FOE is the core component of fractional order circuit design in analog domain. The impedance of a FOE is more generalized version of normal impedances and are

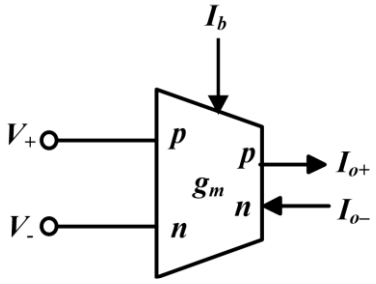


Fig. 1. Block diagram of dual-output OTA

used to model various circuits precisely for diverse applications. The general impedance of FOE is stated as:

$$Z(s) = As^{\alpha} = A(\omega)^{\alpha} e^{j\frac{\alpha\pi}{2}} = A(\omega)^{\alpha}(cos(\frac{\alpha\pi}{2}) + jsin(\frac{\alpha\pi}{2})) = |z|\angle\theta \quad (3)$$

Here θ represents the phase of FOE equals to απ/2 and α denotes the order of the FOE. For a particular value of α, the phase of FOE is constant, thus another name of FOE is constant phase element. For -1<α<0 FOE behaves as a FC and for 0<α<1 FOE acts as a FI. For α=0, 1, -1 FOE acts like pure resistor, inductor and capacitor respectively [17]. The block diagrams for FC and FI appear in Fig. 2.

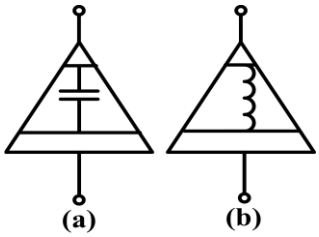


Fig. 2. Block diagram of (a) FC (b) FI

III. PROPOSED IIMC BASED FLOATING FOE

In this work the grounded IIMC [16] is modified to realize a floating IIMC. The modified circuit for generalized floating IIMC is shown in Fig. 3. Taking into account the port relationships of OTA and by performing a routine algebraic analysis the admittance matrix between terminals 1 and 2 of Fig. 3 is given by (4)

$$\begin{bmatrix} I_1 \\ I_2 \end{bmatrix} = g_{m0}(g_{m1}g_{m2}...g_{mn})(Z_1Z_2....Z_n) \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \quad (4)$$

By replacing Z_1 by FC ($C_{1\alpha}$) and other impedances by normal (integer-order) capacitances, the input impedance between terminals 1 and 2 of Fig. 3 modifies to

$$Z_{in(1-2)} = \frac{s^{n-1+\alpha} C_{1\alpha} C_2 ... Cn}{g_{m0}(g_{m1}g_{m2}\cdots g_{mn})} \quad (5)$$

Eq. (5) represents an input impedance of FI with fractional order (n-1+α). By making n=2, FI of fractional order (1+α) can be realized, which is shown in Fig. 4 (a).

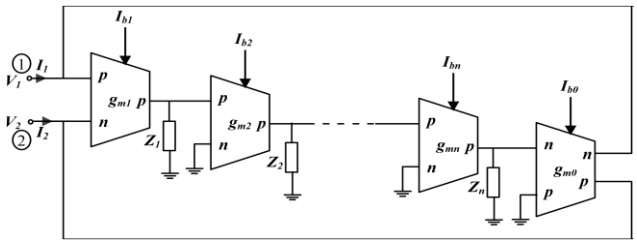
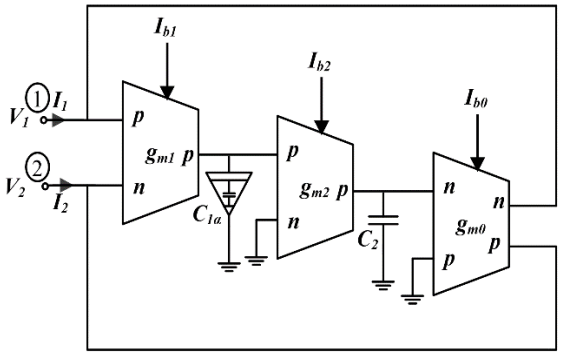


Fig. 3. Generalized floating IIMC

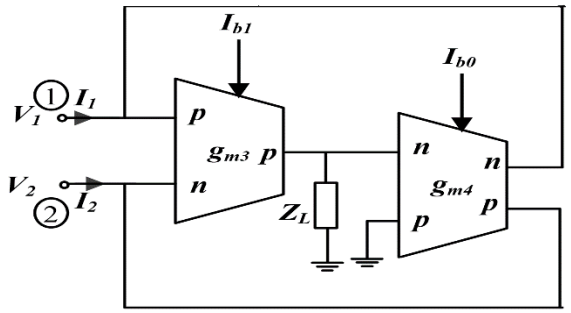
The impedance inverter circuit of Fig. 4(b) is used to realize FC of equal order, where Z_L is replaced with the FI of desired order. The magnitudes of FI and FC presented in Fig. 4 (a) and (b) are given by (6) and (7) respectively.

$$L_{1+\alpha(1-2)} = \frac{C_{1\alpha}C_2}{g_{m0}g_{m1}g_{m2}} \quad (6)$$

$$C_{1+\alpha(1-2)} = \frac{g_{m3}g_{m4}C_{1\alpha}C_2}{g_{m0}g_{m1}g_{m2}} \quad (7)$$



(a)



(b)

Fig. 4. Proposed IIMC based floating FOEs of order (1+α) (a) IIMC circuit realizing FI (b) Impedance inverter circuit realizing FC, with Z_L as FI

IV. Simulation Results

The proposed circuit of Fig. 4 is verified via PSPICE simulation, with 0.18μm CMOS technology parameters. The OTA offered in [18] is used to construct the circuit, Fig. 5 shows CMOS structure of the OTA with VDD as +1.8V and VSS as -1.8V. Tab. II contains aspect ratios (W/L) of the transistors shown in Fig. 5. The $C_{1\alpha}$ is constructed using Foster I network realization based upon CFE approximation, presented in Fig.6. The components' value (R_i and C_i) can be referred from [17].

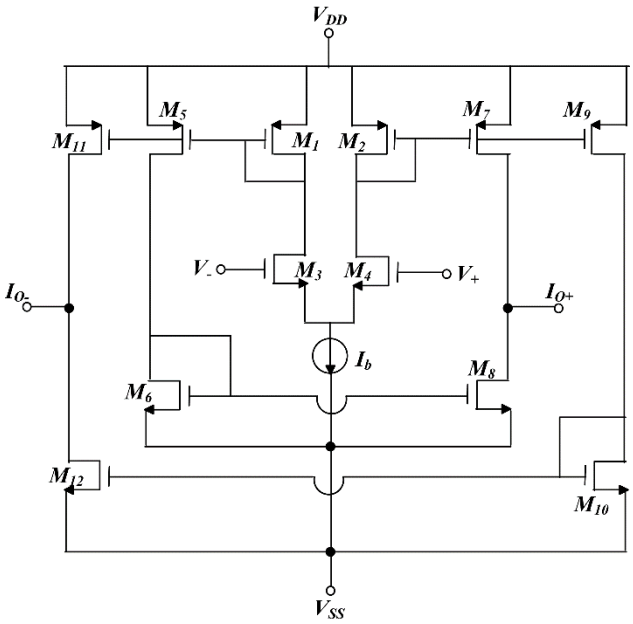


Fig. 5. CMOS realization of OTA [18]

TABLE II
W/L RATIOS OF THE TRANSISTORS IN FIG. 5

	$M_{3,4}$	$M_{1,2,5,7,9,11}$	$M_{6,8,10,12}$
W/L	5.76 μm/0.72 μm	2.16 μm/0.72 μm	1.44 μm/0.72 μm

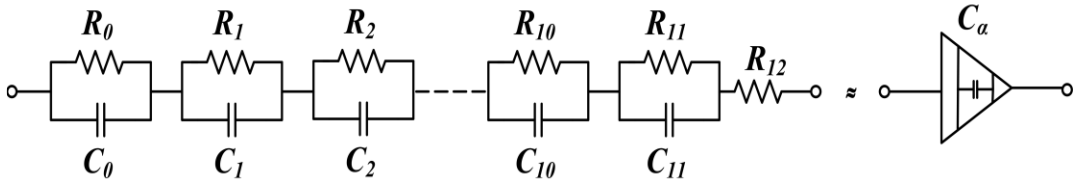


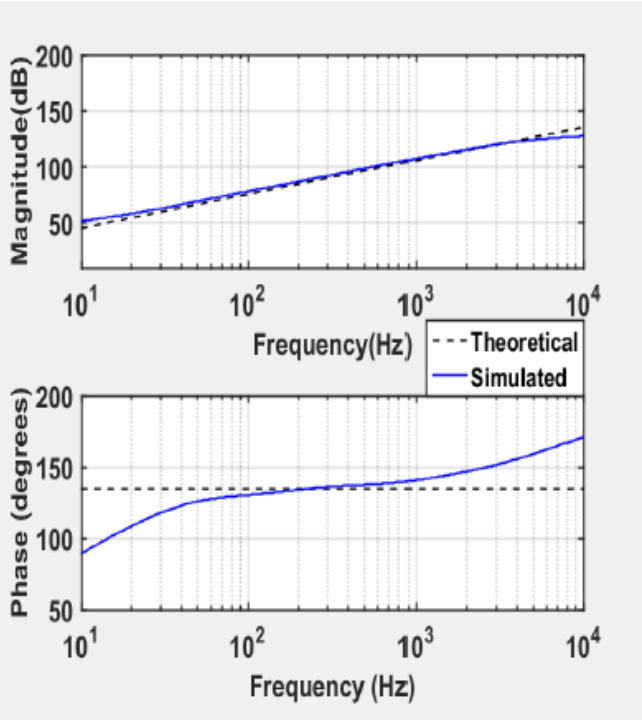
Fig. 6. FC realization using Foster I RC network (12th order approximation)

The order and magnitude of $C_{1\alpha}$ are taken as 0.5 and $3.75\mu\text{℧}/s^{0.5}$ respectively. Bias currents for all the OTAs are taken as equal and the transconductance gain (g_m) is set to 100 μA/V. The value of normal capacitance C_2 is taken as 10nF. Figs. 7(a) and (b) show simulated and theoretical frequency response of (1+α) order FI and FC respectively. The equivalent magnitude of inductance ($L_{1.5}$) and capacitance ($C_{1.5}$) are $37.5m\Omega/s^{1.5}$ and $0.375n\text{℧}/s^{1.5}$. As suggested by eqs. (6) and (7), the magnitudes of FI and FC can be tuned electronically by using external bias currents of OTAs. Tab. III summarizes the frequency range of operation for FI and FC for a maximum deviation of ±5dB in magnitude and $\pm5^0$ phase.

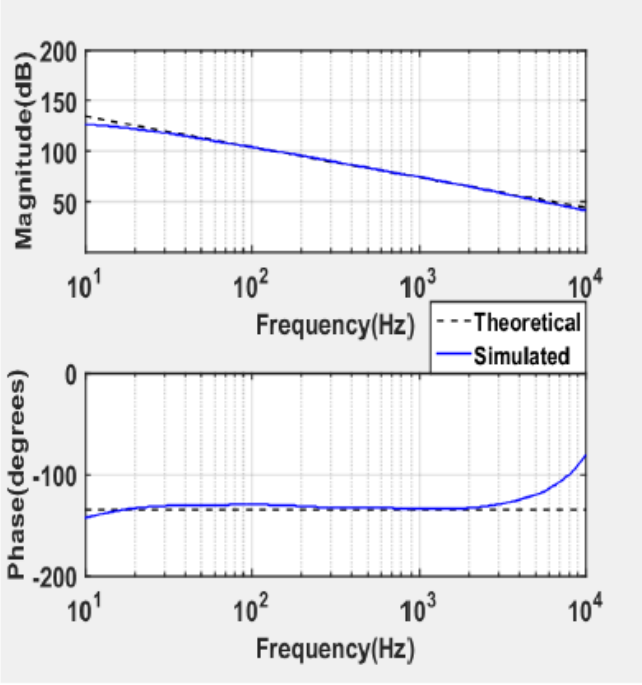
V. Conclusion

In present work, a circuit to realize higher order floating FOE is presented. The proposed circuit can be used to realize floating FI and floating FC. Functional verification of the proposed circuit is accomplished via PSPICE simulations and it is verified that the results are in decent agreement with the theory for a maximum deviation of

±5dB and ±5° in magnitude and phase respectively. The proposed structure utilizes modularity and the magnitude of proposed FI and FC can be controlled electronically by using external bias currents of OTAs.



(a)



(b)

Fig. 7. Impedance magnitude and phase response for floating FOEs of order 1.5 (a) FI (b) FC

TABLE III
FREQUENCY RANGE OF OPERATION FOR PROPOSED FLOATING FOEs

Impedance response	FI	FC
Magnitude	10Hz - 7.7kHz	10Hz - 10kHz
Phase	84Hz - 1.4kHz	11Hz - 2.9kHz

REFERENCES

[1] L. Debnath, "Recent applications of fractional calculus to science and engineering," Int. J. Math. Math. Sci. vol. 54, pp. 3413–3442, 2003.

[2] L. Dorcak, J. Valsa, E. Gonzalez, J. Terpak, I. Petras and L. Pivka, "Analogue realization of fractional-order dynamical systems," Entropy, vol. 15, no. 10, pp. 4199–4214, 2013.

[3] S. Das and I. Pan, "Fractional order signal processing--introductory concepts and applications," In: Springer Briefs in Applied Sciences and Technology, 2012.

[4] G. Varshney, N. Pandey and R. Pandey, "Electronically tunable multifunction transadmittance-mode fractional-order filter," Arab J Sci Eng, vol. 46, pp. 1067–1078, 2021.

[5] L. A. Geddes and L. E. Baker, "Principles of Applied Biomedical Instrumentation," John Wiley and Sons Ltd, Chichester, UK, 3rd ed., 1989

[6] E. Tlelo-Cuautle, A.D. Pano-Azucena, O. F. Guill´en and A. Silva-Ju´arez, "Analog/Digital Implementation of Fractional Order Chaotic Circuits and Applications," Springer, 2020.

[7] G. Tsirimokou, C. Psychalinos, A. S. Elwakil and K. N. Salama, "Experimental behavior evaluation of series and parallel connected constant phase elements," International Journal of Electronics and Communications (AEU), vol. 74, pp. 5-12, 2017.

[8] M. C. Tripathy, D. Mondal, K. Biswas and S. Sen, "Experimental studies on realization of fractional inductors and fractional-order bandpass filters," Int. J. of Circuit Theory and Applications, vol. 43, no. 9, pp. 1183–1196, 2015.

[9] A. S. Ali, A. G. Radwan and A. M. Soliman, "Fractional Order Butterworth Filter: Active and Passive Realizations," in IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 3, no. 3, pp. 346-354, Sept. 2013.

[10] A. Soltan, A. G. Radwan and A.M. Soliman, "Fractional order Sallen–Key and KHN filters: stability and poles allocation," Circuits, Systems, and Signal Processing, vol. 34, no. 5, pp. 1461–1480, 2015.

[11] L. A. Said, S. M. Ismail, A.G. Radwan, A. H. Madian, M. F. A. El-Yazeed and A. M. Soliman, "On the optimization of fractional order low-pass filters," Circuits Systems and Signal Processing, vol. 35, no. 6, pp. 2017–2039, 2016.

[12] M. S. Semary, M. E. Fouda, H. N. Hassan and A. G. Radwan, "Realization of fractional-order capacitor based on passive symmetric network," Journal of Advanced Research, vol. 18, pp. 147–159, 2019.

[13] S. Kapoulea, C. Psychalinos, A. S. Elwakil and A.G. Radwan, "One terminal electronically controlled fractional-order capacitor and inductor emulator," Int. J. of Electronics and Communications (AEUE), vol. 103, pp. 32-45, 2019.

[14] K. Biswas, R. Caponetto, G. D. Pasquale, S. Graziani, A. Pollicino and E. Murgano, "Realization and Characterization carbon black fractional order element," Microelectronics Journal, vol. 82, pp. 22-28, 2018.

[15] R. Sotner, J. Jerabek, J. Petrzela and T. Dostal, "Simple approach for synthesis of fractional-order grounded immittances based on OTAs," 2016 39th International Conference on Telecommunications and Signal Processing (TSP), 2016, pp. 563-568.

[16] R. Verma, N. Pandey and R. Pandey, "Realization of a higher fractional order element based on novel OTA based IIMC and its application in filter," Analog Integrated Circuits and Signal Processing vol. 97, no. 1, pp. 177-191, 2018.

[17] A. Adhikary, P. Sen, S. Sen and K. Biswas, "Design and performance study of dynamic fractors in any of the four quadrants,"

Circuits, Syst. and Signal Proc., vol. 35, no. 6, pp. 1909-1932, 2016.
[18] T. Tsukutani, Y. Sumi and Y. Fukui, “Electronically Controlled Current- Mode Oscillators Using MO-OTAs and Grounded Capacitors,” Frequenz, vol. 60, no. 11-12, pp. 220-223, 2006.

View publication stats

2021 Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 28-29, 2021

Semantic Segmentation using Generative Adversarial Networks with a Feature Reconstruction Loss

Gaurav Sohaliya
Department of Information Technology
Delhi Technological University
Delhi, India
sohaliyagaurav@gmail.com

Kapil Sharma
Department of Information Technology
Delhi Technological University
Delhi, India
kapil@ieee.org

2021 Asian Conference on Innovation in Technology (ASIANCON) | 978-1-7281-8402-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ASIANCON51346.2021.9544954

***Abstract*—The image-to-semantic labels classification is a very challenging task in image processing. Convolutional neural networks (CNN) have managed to achieve the state-of-the-art quality of the segmented image in semantic segmentation tasks. Still, the classification capability of such algorithms is not satisfactory to segment images that contain complex object boundaries and minimal regions. Recently, the Generative Adversarial Networks (GAN) were introduced, which can solve the overfitting of the generator network using the adversarial loss. In this paper, a GAN-based segmentation model is proposed, in which the Conditional GAN (CGAN) model is used as base architecture. Feature reconstruction loss is introduced in this composite model to solve the identification and classification of visually small elements in images. A pre-trained deep convolution neural network is adopted to calculate feature reconstruction loss. The usage of Feature reconstruction loss has demonstrated the high quality of the output labels and improved segmentation masks. The evaluation of the proposed model on the cityscapes dataset has shown the effectiveness of GAN-based architecture in semantic segmentation of multiclass images. The proposed model achieved 83.3% accuracy on the cityscapes test dataset, which is superior to most semantic segmentation state-of-the-art methods.**

Keywords—Semantic segmentation, U-net, Generative adversarial networks, Feature reconstruction loss

I. INTRODUCTION

Object detection, image classification, and image segmentation are very classic problems in computer vision and still challenging tasks in the current time. Identifying whether a particular object is present in an image or not and classifying the image according to it is object detection and image classification, respectively. But the semantic image segmentation phenomenon is very complex. We are interested in finding all the things present in a picture with exact regional boundaries and categorizing each pixel present in an image into a set of classes [1] like bus, car, sky, building, tree, road, etc., assigning a label to it. This is very challenging task in image processing because of multiple angles and viewpoints of different objects and illumination, texture of visual scenes, and the high variation in appearance. Semantic segmentation can be used in automated cars [2], virtual reality, human-computer interaction [3], sky monitoring cameras, delivery drones, etc.

Deep learning can solve these tasks using deep convolution architecture, and some of the problems are also resolved successfully. Convolution neural network architectures (CNN) [4, 5] pushed the accuracy and efficiency of image segmentation to a new height by surpassing other methods with a wide margin. CNNs can solve the problem of semantic segmentation. Still, it requires many modifications because originally, CNN is developed for the image classification task. Because of that, CNN suffers from various problems such as the low resolution of generated output, unable to predict object boundaries accurately. Using the capabilities of CNN and overcoming the problem, several deep architectures and research works have been proposed, such as Fully Convolution Network [6], Pyramid Scene Parsing Network [7], DeepLab [8], U-Net [9], etc. These techniques can solve image segmentation problems successfully at a certain level, but some issues like accurate boundary extraction and area prediction remain to solve.

A convolution neural network tries to learn a parametric translation function in a fully supervised environment. Significant manual efforts are required to compose accurate loss between ground truth pixels and predicted pixels. Ian J. Goodfellow proposed generative adversarial networks (GANs) [10] to address this problem. GAN tries to identify the relation between an input dataset and an output dataset. GANs can be used in several applications, such as object transfiguration, photo enhancement, style transfer, season, or daylight transfer. There are certain limitations of GANs, such as the training of GAN is sensitive and unstable, it cannot control the data being generated by the generator. To tackle these problems, Mehdi Mirza proposed Conditional Generative Adversarial Nets (CGAN) [11], in which condition has been imposed over discriminator to produce the required output. We use CGANs as basic architecture with feature reconstruction loss [12] for high quality and accurate regional segmentation in our proposed architecture.

II. RELATED WORK

This section describes the related literature in semantic segmentation and discussed four methods related to CNN-based and two ways associated with GAN based segmentation model.

A. Fully Convolutional Network (FCN)

FCN [6] is the first successful architecture in terms of accuracy gain in image segmentation, which uses a CNN to reconstruct image pixels into a set of categories. The last dense layer of convolutional neural networks is replaced by

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:46:11 UTC from IEEE Xplore. Restrictions apply.

transposed convolution layer in the proposed FCN architecture. Hence the predicted pixels can have one-to-one conformity in terms of spatial positions with the input image. FCN first uses the consecutive convolution layers to capture image features. Then it transforms the dimensions of the intermediate layer feature block to the dimensions of the input image by using the transposed convolution layer. The model's output will have the exact dimensions as the input image. The final output channel contains the class prediction of the pixel of the corresponding spatial dimension.

The dense layer in CNN has many trainable parameters, and it requires a lot of time to train; this layer is eliminated in FCN. Hence FCN takes less time for training in comparison to CNN. It does not have any dense layer and contains only convolutional layers, because of which it can accept variable size input images.

B. Pyramid Scene Parsing Network (PSPNet)

PSPNet [7] proposed the scene parsing concept, which refers to parse an image semantically into different regions that belong to a set of classes such as sky, road, and car. PSPNets employ pyramid pooling to use the capabilities of the global context of the image. PSPNet uses FCN in its core, where it has series of convolution layers, and then pyramid pooling adds a new layer between the convolution layer and the fully connected layers. It helps to map variable size input to constant output size. PSPNet provides a very efficient architecture for interpreting sophisticated scenes and pixel-level prediction ability in the image. It includes predicting the shape, size, and object position, similar to the human visual approach. PSPnet enormously enhances the ability of image description and provides an obvious direction for the pixel-level prediction.

C. DeepLab

When applying DCNN on images for segmentation, there are certain limitations, and DeepLab [8] tries to overcome these limitations. The first solution is dilated convolution, which solves the problem of reduced feature resolution. Due to multiple pooling and downsampling layers in DCNN reduce spatial resolution significantly from the resultant feature map. Conditional Random field has been employed to increase localization accuracy by considering the neighbor's context while predicting the label. After the first DeepLab model, various revised models have been proposed, including DeepLabV2 [13], DeepLabV3 [14], and DeepLabV3+ [15]. The main idea is to combine the capabilities of DCNN and conditional random fields.

D. U-net Architecture

U-net [9] was proposed for the segmentation of Biomedical Images. The architecture contains two paths. It is divided into two parts, same as FCN; the first part is the contraction path employed to extract the context and details of the image. It is also addressed as encoder architecture and just a stack of convolution layers without dense layers. The decoder part is the symmetric expanding path which is a stack of deconvolutions layers.

The problem in encoder-decoder architecture compresses the input linearly; hence, it creates a bottleneck in which all feature information cannot be passed to a decoder, leading to low resolution and inaccurate output. Here the U-net differs from other architecture. U-net allows the network to propagate context information from the lower layer to the higher layers. The higher resolution feature maps from the encoder's successive layers are concatenated with upsampled features of the decoder network. It helps to learn better representations with subsequent deconvolution operations. Upsampling or deconvolution is a sparse operation, so to increase the resolution and represent better localization, it needs prior knowledge from successive layers. This yields U-shaped architecture, and hence u-net architecture has achieved good accuracy in semantic segmentation [16] [17].

E. Semantic segmentation using Adversarial Networks

In all the previous models, the loss function is majorly based on pixel-to-pixel loss. As discussed previously in [8], a Conditional Markov random field (CRFs) is an effective way to emphasize spatial integrity in output labeled image. However, CRF Model has a limitation of higher-order consistency. To overcome this limitation, in 2016, facebook researchers proposed to use an Adversarial Network [18]. Adversarial loss of GAN architecture helps to enforce higher-order consistency forms, which is impossible with pixel-to-pixel loss.

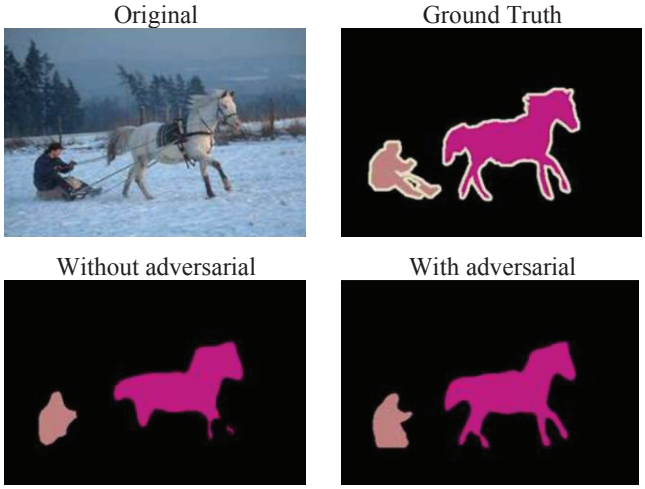


Fig. 1 with and without adversarial training results

The results are shown in Fig.1 prove that adversarial training helps to get spatial consistency and brings smoothness and sharpness.

F. SegGAN

SegGAN [19] has been proposed to utilize the capabilities of adversarial loss. It uses the GAN as a kind of loss to segmentation a given image and achieve effective results. In SegGAN architecture, there are mainly three components:

- Pretrained DeepLab model – it is used for semantic segmentation.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:46:11 UTC from IEEE Xplore. Restrictions apply.

- Generator G – it takes predicated masks from the DeepLab model and tries to generate an original image synthetically.
- Discriminator D – it distinguishes synthetically generated images from a generator and real images from a dataset.

Here, generator G follows a simple encoder-decoder architecture with four convolution and four deconvolution layers. The discriminator uses four convolution layers with ReLU as the activation method. In this paper, another significant thing is the loss function; they integrated three losses overall:

- Segmentation Loss – it is a loss between the predicated segmentation mask map from Segmentation Network and the actual mask.
- Content Loss – this is pixel-wise MSE loss between ground truth image and reconstructed image by Generator G.
- Adversarial Loss – generator G and discriminator D try to improve themselves by fooling each other.

The result obtained by SegGAN is superior in comparison to DeepLab and GAN-based architecture, and the research has successfully demonstrated GAN and adversarial loss capabilities.

III. Proposed Methodology

A. Review of Generative Adversarial Networks (GAN)

GANs [10] are powerful generative models. GAN is working on Game Theory, where it consists of a generator and a discriminator. The generator generates a new image, and the discriminator discriminates whether that newly generated image is actual (as real as training data) or fake. Suppose the discriminator assesses the generated image as a fake image. In that case, it is a loss for the generator network, and if it fails to discriminate as a fake image and classify as an actual image, then it is a loss for the discriminator network. So, it's like a game between these two networks, both are trying to fool each other, and they evaluate themselves by facing loss, using backpropagation, and adjusting weights. The optimization function of GANs is constructed as follows:

$$\min_G \max_D \mathbb{E}_{x\sim\mathbb{P}_r}[\log D(x)] + \mathbb{E}_{z\sim\mathbb{P}_g}\left[\log\left(1 - D\left(G(z)\right)\right)\right] \quad (1)$$

where $x\sim\mathbb{P}_r$ are images from the input dataset, and $z\sim\mathbb{P}_g$ are random noise units, $G(z)$ are the generated images and $D(x)$ is the probability of x being actual.

GAN cannot constrain the data being produced by the generator, and to address this issue, Mirza et al. [11] extended GANs to a CGAN (conditional generative adversarial networks). To achieve specific results from GAN networks, certain auxiliary conditions or information y emphasized on the generator and the discriminator, e.g., class labels. The objective function of CGANs is constructed as follows:

$$\min_G \max_D \mathbb{E}_{x\sim\mathbb{P}_r}[\log D(x,y)] + \mathbb{E}_{z\sim\mathbb{P}_g}\left[\log\left(1 - D\left(x, G(z)\right)\right)\right] \quad (2)$$

B. Feature Reconstruction Loss

In CNN, the feature extraction is done hierarchically. So, the response of the lower layer (layers that are nearest to the input layer) of CNN is much similar to the input image because feature extraction at lower layers is at a lower tendency, so it just reproduces much similar output with input. Higher layers give high-level content (e.g., objects in the scene, the relative positions of objects, boundaries of regions), so we prefer to use a higher-level feature map for feature reconstruction loss [20]. The primary purpose of feature reconstruction loss is to leverage the knowledge of a pre-trained network that can identify more minor details and boundaries of regions of image very effectively. U-net and other segmentation algorithms are struggling in the identification and classification of small object and their areas.

We use the content loss to bring more information, accurate boundaries, and smooth resultant segmented image to overcome this limitation.

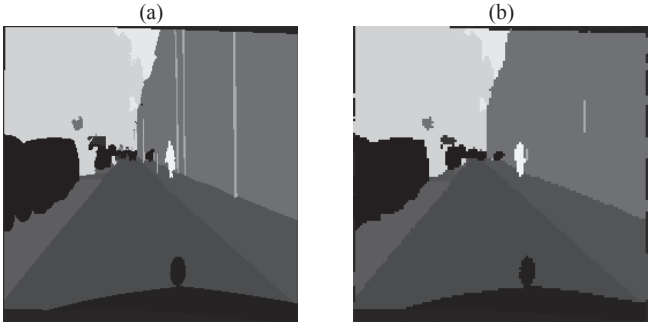


Fig. 2 (a) ground truth (b) output of U-net.

Fig. 2(a) is the ground truth segmented image from the dataset, whereas Fig. 2(b) is an output of U-net. Some details like street lights and stands are not visible, and boundaries are also not smooth in the predicated image compared to the ground truth image. If the original image (not segmented) is passed to pre-trained CNN, it can identify minor details where U-net is failed to classify it. Fig. 4 shows the comparison of the higher-level feature map of the actual image and predicted image.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:46:11 UTC from IEEE Xplore. Restrictions apply.

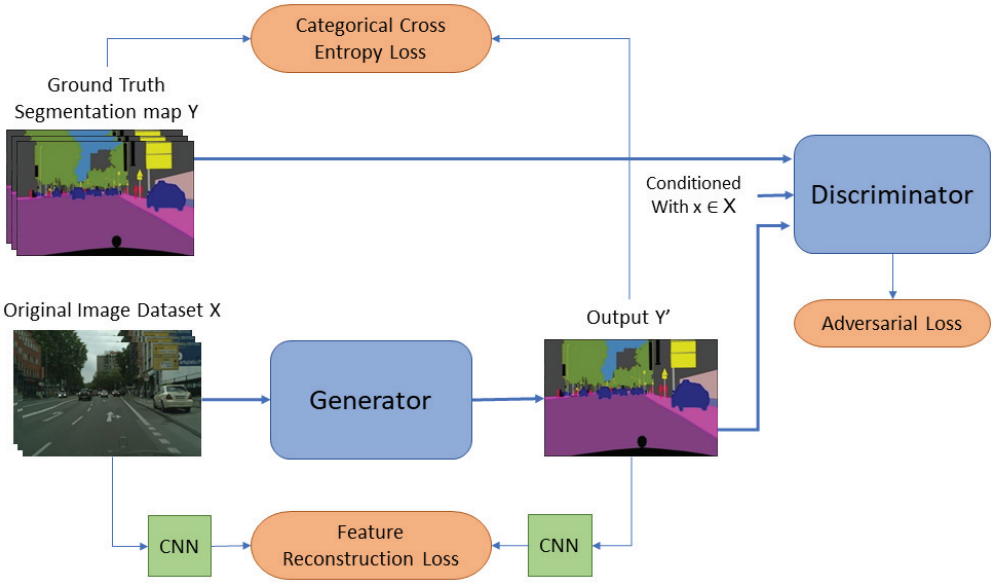


Fig. 3 The proposed architecture

The difference between the higher-level feature map of the original image and the predicted image can be employed as a feature reconstruction loss to gain a sharp and detailed result.

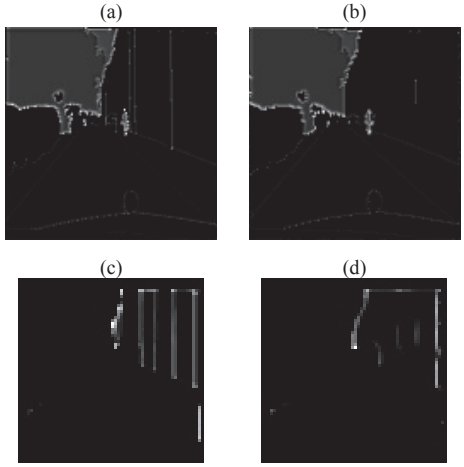


Fig. 4 feature map of (a) ground truth (b) U-net prediction (c) Ground truth (d) U-net prediction.

So, let $\vec{o}$ and $\vec{g}$ be the original and generated image, respectively, whereas O^l and G^l are their respective feature representation in layer l. C, H, and W represent the number of filters in feature map, height, and width of the image, respectively. We then define the squared-error loss (L2 Norms) between the two feature map representations. The feature reconstruction loss for a layer l can be formulated as:

$$\mathcal{L}_{feat}(\vec{o}, \vec{g}, l) = \frac{1}{C*H*W} \sum_{i=0}^{C} (G_i^l - O_i^l)^2 \qquad (3)$$

The above Equation is for one-layer content representation only; we can take multiple layers, with each layer has a different contribution factor.

C. *Network Architecture*

The Proposed method (Fig. 3) is a different technique in which the conditional GAN is combined with the pre-trained VGG19 [21] network for calculating the feature reconstruction loss. This hybrid architecture consists of a generator model, discriminator, and VGG nets.

U-net has shown excellent performance in image segmentation, and hence we are incorporated U-net with slight modification as generator architecture (Fig. 5). U-net is built on the encoder-decoder model, including skip connection, where skip connection helps u-net in shuffling low-level details directly between the encoder input to decoder output. The encoder aims to map the input data space into output data space, whereas the decoder forces encoder to do a meaningful reconstruction of output from a given input space. The encoder part of a generator (fig. 5) consists of 5 downsampling blocks, whereas the decoder part consists of 4 composite blocks of an upsampling block, concatenation layer, and downsampling block. ReLU activation function is used for both the convolution and deconvolution processes. Every downsampling block contains two convolution layers, followed by a batch normalization layer. Upsampling block contains transposed convolution layer followed by batch normalization layer. The output layer of the generator has softmax as an activation function because it predicts each class's probability for each pixel. In contrast, the discriminator accepts two images, generated image and the conditioned image (original image), with three channels. Generated image has N number of channels, and discriminator requires only three channels. Hence, we added argmax and stack layer between generator and discriminator.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:46:11 UTC from IEEE Xplore. Restrictions apply.

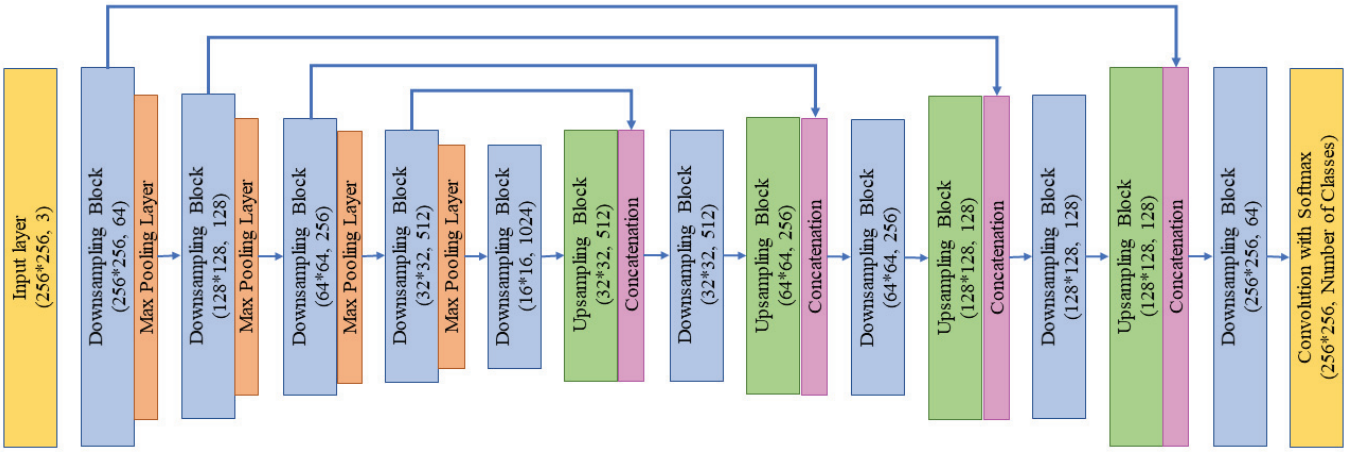


Fig. 5 Generator network for the proposed method

The PatchGAN is used as a discriminator, which is proposed in [22], aims to categorize every patch into real or fake of an image. Discriminator runs over the entire image patch by patch and does convolution over each patch. The mean of all the values will generate the final output of the discriminator. PatchGAN projects the image as same as CRF (Markov random field), and hence it acts as a form of texture.

We used VGG19 pre-trained architecture to calculate feature reconstruction loss. It is a very deep CNN trained on a large dataset and observed that it could extract more minor details accurately than other pre-trained networks.

D. Optimization

The objective function $\mathcal{L}$ of our method consists of three vital components: Adversarial Loss $\mathcal{L}_{GAN}$, Categorical cross-entropy $\mathcal{L}_{CCE}$, and feature reconstruction loss $\mathcal{L}_{C}$. In detail, the final loss function for our segmentation architecture, $\mathcal{L}_{SGAN}$, will be:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{CCE} + \lambda_2 \mathcal{L}_{FEAT} + \lambda_3 \mathcal{L}_{GAN} \quad (4)$$

Where λ_1, λ_2, and λ_3 are weighting hyperparameters of each loss.

Here we adopted categorical cross-entropy to calculate the distance between ground truth segmentation image and generated segmentation image because each pixel belongs to a particular class. Therefore, the discriminator task remains as it is, but the generator has to be near the pre-segmented image and fool the discriminator. So categorical cross-entropy is:

$$\mathcal{L}_{CCE}\big(y, G(x)\big) = \sum_i^N \sum_c^C G(x_{ic}) \log(y_{ic}) \quad (5)$$

Here x is from the original image, and y is the ground truth segmented image. $G(x_{ic})$ represents the probability of class c for generated pixel i and y_{ic} represents the probability of class c for ground truth pixel i.

Generated segmentation images will not be sharp, and the edges of objects will not be smooth and accurate without feature reconstruction loss. It guides the generator to follow the content and object edges of an image x. The block5_conv2 layer of VGG19 is used to calculates feature reconstruction loss because this layer contains the higher-level feature of an image which generally represents the shape and edges of objects. So, feature reconstruction loss will be:

$$\mathcal{L}_{FEAT}(x, G(x)) = \frac{1}{C*H*W} \sum [\| x^l - G(x)^l \|_2] \quad (6)$$

Here l is the layer of VGG19.

The adversarial loss is used to improve and enhance the constructed result by generator, which can be defined as:

$$\mathcal{L}_{GAN}(G, D) = \min_G \max_D \mathbb{E}_{x,y} [\log D(x, y)] + \mathbb{E}_{x,y} \Big[\log\Big(1 - D\big(x, G(x)\big)\Big)\Big] \quad (7)$$

IV. Experiments

A. Description of Dataset

We used the Cityscapes [23] dataset for this research. It is a huge dataset that focuses on semantic segmentation, through which we can understand urban street scenes. The dataset consists of around 20000 coarse annotated ones, and 5000 fine annotated images. Dataset contains daytime images captured from 50 different cities. There are mainly 34 classes like objects, humans, flat surfaces, constructions, sky, vehicles, nature, void, etc. The class contribution in images is hugely unbalanced in terms of region area, and it also includes a diverse scene layout and a large number of objects. It will help us to check the robustness of the model. It provides three types of annotation, e.g., instance-wise, semantic, and dense pixel annotations. Since this dataset contains high-resolution images, we converted it into 256*256 resolution so that class information assigned to each pixel is not loose.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:46:11 UTC from IEEE Xplore. Restrictions apply.

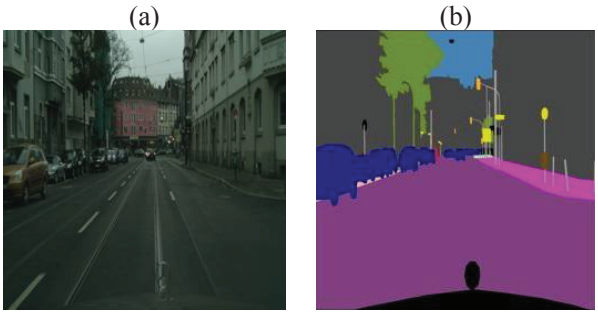


Fig. 6 Cityscapes Dataset (a) original (b) segmentation mask

We used a fine annotated dataset with 2975 training, 500 validation, and 1525 test images for our experiment

B. *Experimental Setup*

The first convolution layer contains 64 filters in both the generator and discriminator. We used upsampling factor and downsampling factor as 2 for every convolution and deconvolution block present in the generator and the discriminator with a kernel size of 3 × 3. Both the networks use an Adam optimizer with a 0.0002 learning rate. Moreover, this type of architecture uses one as the batch size; we use the same and train for 200 epochs. λ_1, λ_2, and λ_3 are adversarial loss, feature reconstruction loss, and categorical cross-entropy loss weights, respectively where $\lambda_1 = 1, \lambda_2 = 100$ and $\lambda_3 = 100$. Our networks were implemented with a Keras v2.0 with Tensorflow v1.15 framework.

V. Results and Analysis

The Mean Intersection over Union (mIoU) score [6] is used to evaluates the architecture performance, which is the standard accuracy matric for semantic segmentation tasks.

Using U-net architecture in CGAN with feature reconstruction loss, the proposed Segmentation GAN model has achieved a higher mIoU score than the original U-net. The obtained mIoU score is similar in such classes with larger region areas, e.g., road, sky, and buildings, compared to other methods. But the proposed architecture managed to improve mIoU score of minor objects such as humans, traffic lights, signs, etc., because of feature reconstruction loss adoption. Hence overall mIoU score has improved. The network managed to achieve an 83.30%, 86.90% and 97.5% mIoU score on the testing set, the validation set and the training set, respectively. It shows that more present classes in the dataset are classified accurately. The comparison table of segmentation methods performance is provided in Table I.

TABLE I. Segmentation performance on the cityscapes dataset

Method	IoU Score%
FCN-8s [6]	65.30
PSPNet [7]	81.20
DeepLabv2-CRF [13]	70.40
DeepLabv3 [14]	81.30
DeepLabv3+ [15]	82.10
U-net [9]	69.10
Proposed method	**83.30**

IoU Score of all the models mentioned in Table 1. are available in cityscapes dataset site [23] except U-net. Therefore, we trained U-net separately in the same hardware and software environmental setup. Previous GAN-based models used PASCAL VOC 2012 dataset [24] for the segmentation task; hence we trained our model on the PASCAL VOC dataset for the experiment. Our method achieved a better result than previous GAN-based segmentation models because this dataset includes a smaller number of categories, and objects presented in the dataset are also visually very large compared to the cityscapes dataset. The comparison table for GAN-based segmentation model is provided in Table II.

TABLE II. Segmentation performance on the PASCAL VOC dataset

GAN based method	IoU Score%
INRIA et al. 2016 [18]	54.30
SegGAN [19]	69.90
Proposed method	**82.10**

Here are some results generated by the proposed architecture in cityscapes dataset.

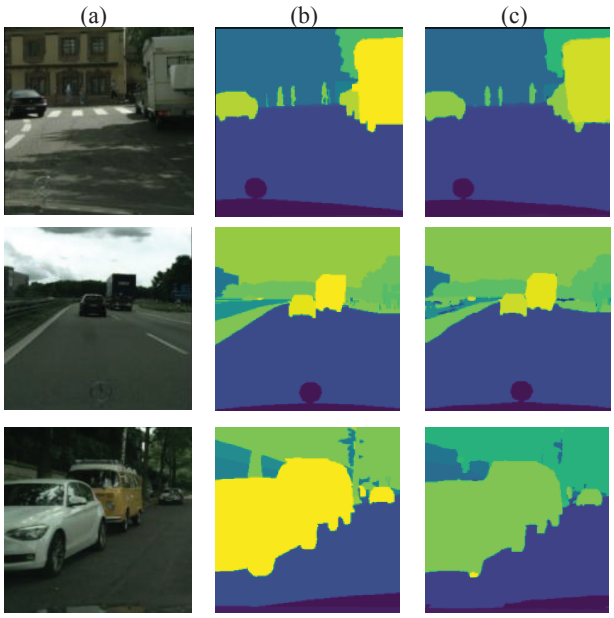


Fig. 7 Output sample generated by proposed model (a) original (b) ground truth (c) output

VI. Conclusion

In this paper, we proposed a GAN-based semantic segmentation method with the aim to identify more minor region details of an image and improve the segmentation model (generator) itself by competing with the discriminator. This aim has been fulfilled by employing feature reconstruction loss and adversarial loss. The several experiment results on the cityscapes dataset prove that the proposed method has significantly enhanced the quality and performance of the existing conventional and GAN-based semantic segmentation models.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:46:11 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez and J. G. Rodríguez, "A Review on Deep Learning Techniques Applied to Semantic Segmentation," CoRR, vol. abs/1704.06857, 2017.

[2] A. Ess, T. Mueller, H. Grabner and L. Van Gool, "Segmentation-Based Urban Traffic Scene Understanding.," in BMVC, 2009.

[3] M. Oberweger, P. Wohlhart and V. Lepetit, "Hands Deep in Deep Learning for Hand Pose Estimation," CoRR, vol. abs/1502.06807, 2015.

[4] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee and I. So Kweon, "Learning a deep convolutional network for light-field image super-resolution," in Proceedings of the IEEE international conference on computer vision workshops, 2015.

[5] D. Ciresan, A. Giusti, L. Gambardella and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," Advances in neural information processing systems, vol. 25, p. 2843–2851, 2012.

[6] J. Long, E. Shelhamer and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," CoRR, vol. abs/1411.4038, 2014.

[7] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, "Pyramid scene parsing network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," arXiv preprint arXiv:1412.7062, 2014.

[9] O. Ronneberger, P. Fischer and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention, 2015.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial networks," arXiv preprint arXiv:1406.2661, 2014.

[11] M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv preprint arXiv:1411.1784, 2014.

[12] J. Johnson, A. Alahi and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in European conference on computer vision, 2016.

[13] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, pp. 834-848, 2018.

[14] L.-C. Chen, G. Papandreou, F. Schroff and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," CoRR, vol. abs/1706.05587, 2017.

[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," CoRR, vol. abs/1802.02611, 2018.

[16] S. Suri, A. Gupta and K. Sharma, "Comparative Analysis of Ranking Algorithms Used On Web," Annals of Emerging Technologies in Computing (AETiC), vol. 4, 2020.

[17] A. K. Tripathi, K. Sharma, M. Bala, A. Kumar, V. G. Menon and A. K. Bashir, "A Parallel Military-Dog-Based Algorithm for Clustering Big Data in Cognitive Industrial Internet of Things," IEEE Transactions on Industrial Informatics, vol. 17, pp. 2134-2142, 2021.

[18] P. Luc, C. Couprie, S. Chintala and J. Verbeek, "Semantic Segmentation using Adversarial Networks," in NIPS Workshop on Adversarial Training, Barcelona, 2016.

[19] X. Zhang, X. Zhu, 3. X. Zhang, N. Zhang, P. Li and L. Wang, "SegGAN: Semantic Segmentation with Generative Adversarial Network," in 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM), 2018.

[20] L. A. Gatys, A. S. Ecker and M. Bethge, "A Neural Algorithm of Artistic Style," CoRR, vol. abs/1508.06576, 2015.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[22] P. Isola, J.-Y. Zhu, T. Zhou and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," CoRR, vol. abs/1611.07004, 2016.

[23] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," CoRR, vol. abs/1604.01685, 2016.

[24] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn and A. Zisserman, "The pascal visual object classes (voc) challenge," International journal of computer vision, vol. 88, p. 303–338, 2010.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:46:11 UTC from IEEE Xplore. Restrictions apply.

2021 Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 28-29, 2021

Sentiment Analysis of News Headlines Using Simple Transformers

Anurag Singh
Department of Applied Mathematics
Delhi Technological University
Delhi, INDIA
anuragsingh.ucer@gmail.com

Goonjan Jain
Department of Applied Mathematics
Delhi Technological University
Delhi, INDIA
goonjan_jain@hotmail.com

2021 Asian Conference on Innovation in Technology (ASIANCON) | 978-1-7281-8402-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ASIANCON51346.2021.9544806

***Abstract*—With the rate at which the data is being generated, it is vital to use it and get some insights from it. When reporting on events, news expresses the opinions of news entities like people, locations, and things. In this paper, we obtain the sentiment of the news headlines using a new technique called transformers, in particular simple transformers that have been a significant advancement in the natural language processing field. Sentiment Analysis has been performed using the four transformers model. These models are pre-trained on extensive data, and we have fine-tuned them by training them on our own news headlines dataset. For our sentiment analysis task, classification models (specific simple transformer model) are used to classify news headlines as negative, neutral, positive. The idea behind taking four different models that are Bidirectional Encoder Representations from Transformers (BERT) base-cased [1], Robustly optimized BERT approach (RoBERTa) base [1], Distilled BERT (DistilBERT) base-cased [1], XLNet base-cased [1], is the different dataset on which they are pre-trained, parameters used by them, and different method used by them, which improve their performance significantly in comparison to different machine learning classifiers and prior deep learning models. The model that performed the best is bert-base-cased with Matthews Correlation Coefficient (MCC) score of 90.1%, an F1 score of 93.6%, and an Accuracy score of 93.6%.**

***Index Terms*—sentiment analysis, news headlines, simple transformers, BERT, ROBERTa, DistilBERT, XLNet.**

I. Introduction

Opinions or Sentiments are primary to every human exercise and play a key role in influencing our behavior. Sentiment analysis is a method to extricate sentiments or opinions from the text data regarding positive, negative, or neutral opinions expressed in natural language. Sentiment analysis is applied in various fields like businesses, organizations, politics, sports, etc., and with the increasing amount of data, it has become a necessity nowadays. The sentiment analysis is conducted at four levels: word level, sentence level, document level, and aspect level [2]. According to [3], the Obama administration used sentiment analysis to measure public opinion before a presidential election. With the increase in news sources and access to the internet, the volume of news is increasing. As per [4], most people evaluate news reports by scanning the headlines rather than reading the whole news. Readers are more likely to use online news outlets due to two main factors: interactivity and immediacy [5].

Some of the approaches of extracting sentiment or opinion from the text are unsupervised learning [6], supervised learning [6], and semi-supervised learning [6]. When we train a machine using the data neither labeled nor classified, the first approach allows the algorithm to learn from the data without supervision. In the second approach, we train a machine using labelled data, and after that, the machine is given a new set of unknown data. The third approach is a combination of the first two approaches in which a machine is trained upon a combination of labelled and unlabeled data, therefore cutting down the labelled data expense.

In this paper, Sentiment analysis is done based on the news headlines, instead using the entire news article. The method proposed here starts with the data collection and preprocessing. Labels for the news headlines were assigned as 0(Negative),1(Neutral), 2(Positive) using the natural language toolkit's built in Vader sentiment analyzer that ranks a text as negative, neutral, positive based on the sentiments. We used Transformer models, specifically simple transformers, which are exceptional in advancing Natural Language Processing.

Advances in both model design and model pre-training have fueled recent advances in natural language processing. Transformer architectures have made it easier to create higher-capacity models, and pretraining has enabled this capacity to be used for a broad range of tasks [7]. Simple transformer is a library built on Hugging Face's Transformers library [8]. Every model is packed with functions and features that are most appropriate for the job at hand. It lets you train, test, and predict in a short span of time. They are designed with a specific natural language task in mind [8]. Simple transformers models can be used in the following steps: (1) initializing a task-specific model [8], (2) training the model using the training data [8], (3) evaluate the model using the testing data [8], (4) making predictions on unlabelled data [8]. The main differences between models will probably be differences in the format of input/output data as well as any task-specific features/configuration choices [8].

II. Literature Survey

Various researches have contributed to news sentiment analysis using various approaches. A brief review of work done

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:44:23 UTC from IEEE Xplore. Restrictions apply.

formally on sentiment analysis is given in this section.

Simon Fong, et al. [9] proposed a research and compared different classification and text mining algorithms for analysing sentiment in newspapers. Maximum Entropy, Naive Bayes, Decision Tree, winnow, and C 4.5, were the classifiers included. Instead of editorials, reviews, or blogs news articles were chosen as they contain less emotionally loaded texts than the former. Their results reveal that the Naive Bayes classifier outperforms all others by a wide margin.

Reis, et al. [10] investigated possible strategies used by web news corporations to design their news headlines. They used 69,907 news headlines from four major news organisations. They designed the methodology to use headlines as a key artifact because they are critical in gaining interest in online news. Their methodology employed sentiment analysis to understand inherent aspects of headlines in order to predict news popularity dependent on URL shortening service. They discovered that the headline's sentiment is closely linked to the news's popularity.

Godbole, Srinivasaiah, and Sekine in [11] presented an approach for awarding scores to every different entity in the text corpus that indicates positive or negative opinion. The approach included a component of sentiment identification, that associated expressed sentiments to each relevant entity, as well as a sentiment aggregation and scoring component that scored each and every entity relative to others of the same class [11]. For sentiment analysis from news and blogs, the news dimensions considered were crime, general, health, business, media, sports, and politics [11]. The following were two patterns that were tracked over time using sentiment score: (1) Polarity: positive or negative sentiment linked with the entity [11], and (2) the amount of sentiment an entity obtained [11].

According to [7] transformer architecture is especially well suited to pretraining on massive text corpora, resulting in significant improvements in accuracy on downstream tasks such as language understanding [12], coreference resolution [13], text classification [14], common-sense inference [15], summarization [16], and machine translation [17] among others.

The effectiveness of transformers for varied NLP tasks motivated us to implement transformers based sentiment analysis.

III. Proposed Work

This paper will analyze and introduce transformers using a simple transformers library in the easiest way possible. We will use the classification model from the simple transformers library to conduct our sentiment analysis of news headlines. The simple transformer models that are used here are Bidirectional Encoder Representations from Transformers (BERT) base-cased [1], Robustly optimized BERT approach (RoBERTa) base [1], Distilled BERT (DistilBERT) base-cased [1], XLNet base-cased [1]. Each of these models is pre-trained on large text data. The techniques and the volume of data used by these models are what sets them apart. The techniques used by different models are: (1) BERT uses masked language model, and next sentence prediction [1], (2) RoBERTa replaces the next sentence prediction task in bert with dynamic masking [1], (3) DistilBERT is a distilled version of bert using half of the number of parameters of the bert [1], (4) XLNet uses permutation language modelling [1].

The methodology for this experiment has been presented in Fig. 1.

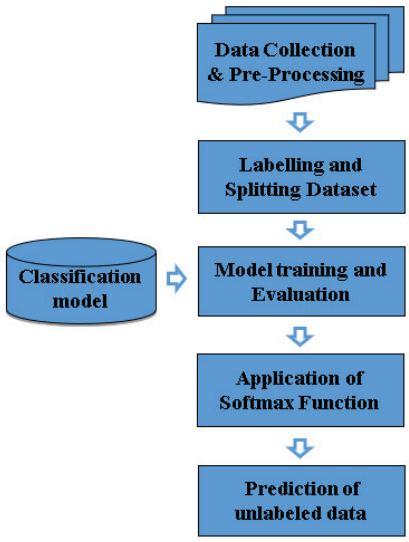


Fig. 1: Workflow for the News Headlines Dataset

The methodology consists of 6 steps, starting with data collection and pre-processing. The News headlines dataset has been used here for the experiment. The steps have been explained below in detail.

A. Data collection and Pre-processing

The news headlines dataset used here were collected from different news websites worldwide by web-scraping through a Python script and Python's newspaper package. News headlines dataset and code is uploaded at [1]. This dataset comprises a total of 33,801 news headlines. The news headlines belong to the business, sports, health, politics, tech, entertainment, education, science, lifestyle, environment, economy, crime, blogs, automobiles, travel, economics.

News headlines in the dataset do not require much pre-processing as most of the work is handled by classification models of simple transformer. Other than this, some of the pre-processing tasks done here are removing punctuation marks, numbers, special symbols, NA-values.

[1]https://github.com/singh1998anurag/Sentiment-analysis.git

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:44:23 UTC from IEEE Xplore. Restrictions apply.

B. Labelling and Splitting Dataset

The process of supervised sentiment analysis requires data to be labelled. The natural language toolkits built-in Vader sentiment analyser used here provides a compound value that ranges from -1 (Extremely Negative) to 1 (Extremely positive); with some experimentation and testing, we consider texts with a compound value less than 0 as negative, greater than 0 as positive, and equal to 0 as neutral.

After labelling, data was split in a ratio of 80% (training data) and 20% (testing data) using the python library sklearn model selection train test split with the shuffle parameter as true and random state parameter 2.

C. Model Training and Evaluation

Model training with a good range of hyperparameter values are critical to achieving a good outcome. Machine learning models can be highly sensitive to the hyperparameters used to train them [8]. Although large models, such as transformers, will work well over a broader range of hyperparameters. All the four classification models discussed above are initialized, and hyperparameters values are given to them after this training data is used for training each model with early stopping set to true used as one of the hyperparameters to prevent model overfitting [8].

Once the training part is over, the model evaluation performed on the test data. In the results and discussion section, the evaluation results for each model are listed.

D. Application of Softmax Function

After getting results on evaluation, the softmax function used to get the probabilities of each class of the news. The softmax function is another kind of activation function used in neural computation [18]. It converts a list of N real values to a list of N real values that add up to 1 [19]. The softmax converts input values that are zero, positive, negative, or more than one into values between 0 and 1, allowing them to be represented as probabilities [19]. Since several multi-layer neural networks generate real-valued scores that are difficult to scale and interact with, the softmax layer can be used in this situation because it converts the scores to a normalised probability distribution that can be presented to the user or used as input to other systems [19]. The formula used to compute softmax function is:

$$f(x_i) = \frac{exp(x_i)}{\sum_{j=1}^{N} exp(x_j)} \quad (1)$$

The mathematical definition of the terms and symbols in [18, eq. 1]: (1) x_i's are components of softmax function's input list/vector, and they can take positive, negative, zero, real, positive value. (2) $exp(x_i)$ is an exponential function that is applied to each component of the input vector. It yields a positive value greater than zero, which is very low when the input is less than zero and very high when the input is large [19]. (3) $\sum_{j=1}^{N} exp(x_j)$ this term ensures all the function's output values add up to 1 by normalization [19]. (4) N is the number of classes like in this paper there are 3 classes negative, neutral, positive.

The model output matrix obtained in the evaluation step fed to the softmax function, where the function returns probability values for each class, with the main target class having the highest probability. A sample model output matrix in Fig. 2:

```
[[-3.12695312  3.45898438 -0.77490234]
 [ 3.50585938 -1.58789062 -1.60058594]
 [ 6.69921875 -2.97460938 -2.28515625]
 ...
 [-2.77734375 -3.35546875  5.37109375]
 [-2.92382812  5.4609375  -2.64257812]
 [-3.09960938  5.5234375  -2.52929688]]
```

Fig. 2: Model Output matrix example

E. Prediction of unlabelled Data

In the real world, we do not know what an accurate label is, so we use the predict function, which is nearly the same as the evaluate function, to make predictions on unlabelled data. It accepts a text list and returns a list of predictions, and a model output matrix.

IV. Results and Discussion

The comparative analysis generated using the developed news headlines dataset and the implementation of four different classification models were plotted using a line graph, a ROC curve, a confusion matrix, and the analysis is done.

A python library wandb used to visualise each classification model at different stages wherever required. As per [20], "Wandb, a framework-agnostic experiment tracking tool for machine learning". It was created to make it easier for people to build machine learning models and save them.

Different machine learning classifiers can be used to perform sentiment analysis an application of text classification. With the amount of data generated, we are bound to see developments, which leads to the machine learning subfield of deep learning. Developments in deep learning resulted in neural network architectures such as convolutional neural networks (CNN) [21], and recurrent neural networks (RNN and LSTM) [21], showing a significant increase in efficiency while tackling natural language tasks. Transfer learning is defined as a methodology in which a deep learning model is trained on a massive dataset to handle related tasks on another dataset [21]. These deep learning models are referred to as pre-trained models. Since sequence modelling tasks like text classification, machine translation, etc., cannot be handled by traditional machine learning models and neural networks, this led to RNNs and CNNs.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:44:23 UTC from IEEE Xplore. Restrictions apply.

RNNs can take one token at a time as input since they cannot be parallelized, thus consuming a significant amount of time while training on massive datasets [21]. This issue prompted the introduction of transfer learning in natural language processing, which resulted in a wide variety of task-specific transformer models. One of the benefit of the transformers is that they accept a complete sequence as input rather than processing token by token [21], and therefore the training can be accelerated by using GPU's.

A. Measurements

- Matthews Correlation Coefficient (MCC): It is a measure of quality of multiclass classification. It takes into consideration all the blocks of the Confusion Matrix in its formula as seen in [22, eq. 2]. Its value lies between -1 to +1 with former indicating poor model and later indicating perfect model. This is shown in Fig. 3.

$$MCC = \frac{t_p * t_n - f_p * f_n}{\sqrt{(t_p + f_p)(t_p + f_n)(t_n + f_p)(t_n + f_n)}} \quad (2)$$

  t_p, t_n, f_p, and f_n are true positives, true negative, false positive, false negative.

- F1-Score: It is defined as the weighted average of the precision and recall with best value as 1 and worst value as 0. Precision and recall have equal contribution to the F1-score as seen in [22, eq. 3].

$$F1 - Score = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \quad (3)$$

- Accuracy: It is calculated by dividing the total number of correct predictions by the total number of predictions.
- Eval Loss: It is the evaluation loss occurred during evaluation of the model on the dataset. This is shown in Fig. 4.
- train loss: It is the training loss occurred during training of the model on the dataset. This is shown in Fig. 5.
- global steps: It is a global step in processing that gets incremented everytime Wandb.log is called. Wandb.log is used to log data from runs [20].

TABLE I: OUTPUT TABLE

Classification Models	Model Metrics			
	MCC Score	*F1-Score*	*Accuracy*	*Eval Loss*
BERT base cased	90.1%	93.6%	93.6%	38.8%
ROBERTa base	89.4%	93.2%	93.2%	40.0%
DistilBERT base cased	89.0%	92.9%	92.9%	39.5%
XLNet base cased	87.5%	91.9%	91.9%	43.9%

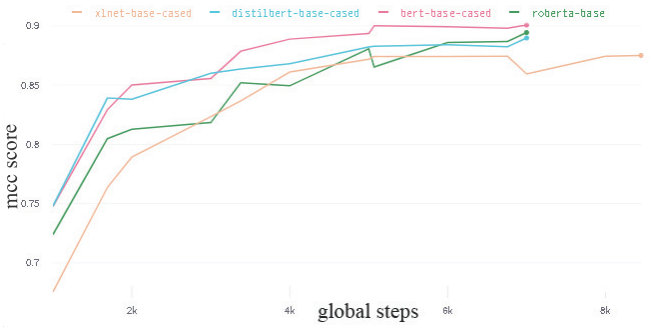


Fig. 3: mcc vs global steps curve

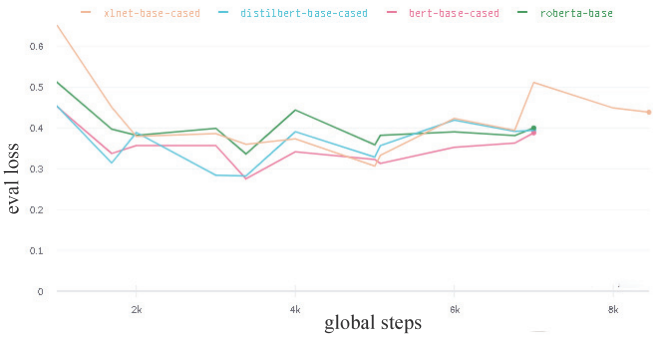


Fig. 4: eval loss vs global steps curve

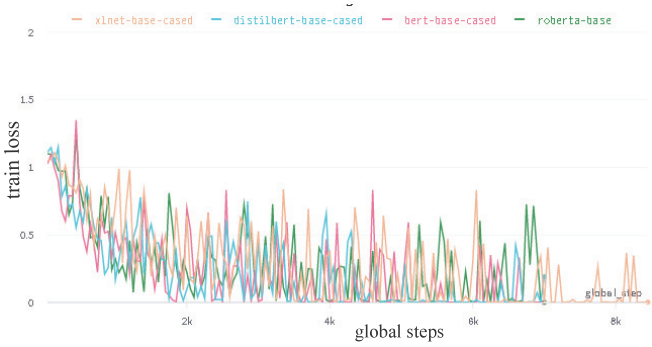


Fig. 5: train loss vs global steps curve

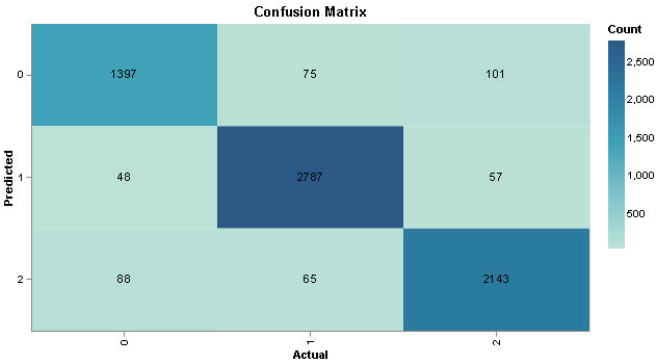


Fig. 6: Confusion matrix for BERT model

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:44:23 UTC from IEEE Xplore. Restrictions apply.

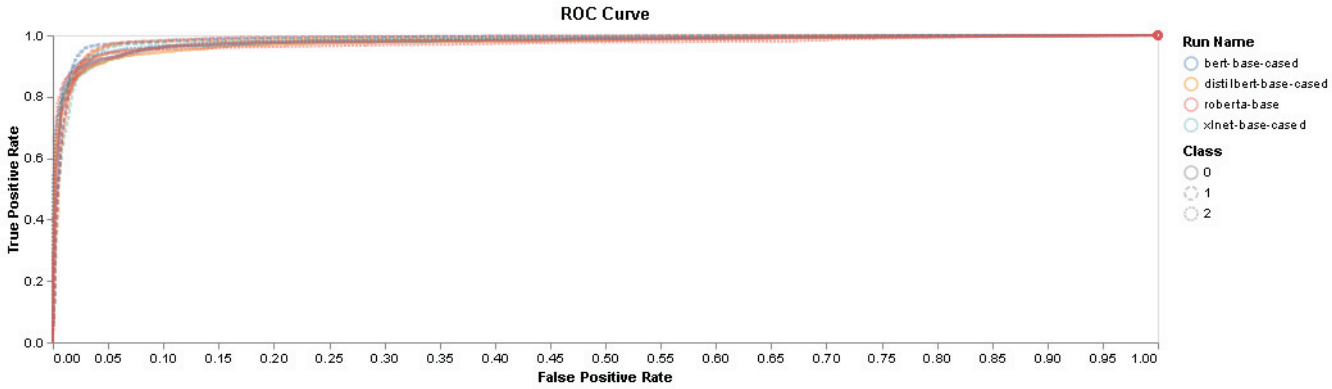


Fig. 7: ROC curve between true positive vs false positive rate

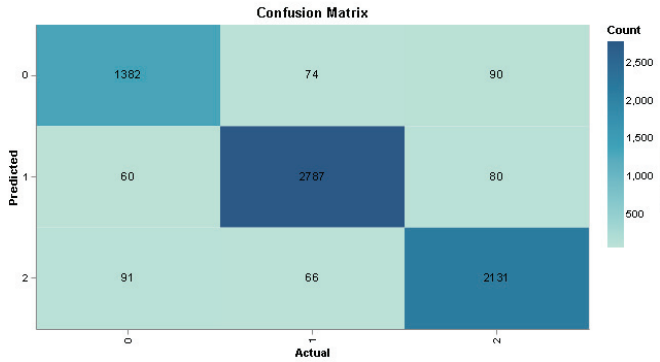


Fig. 8: Confusion matrix for ROBERTa model

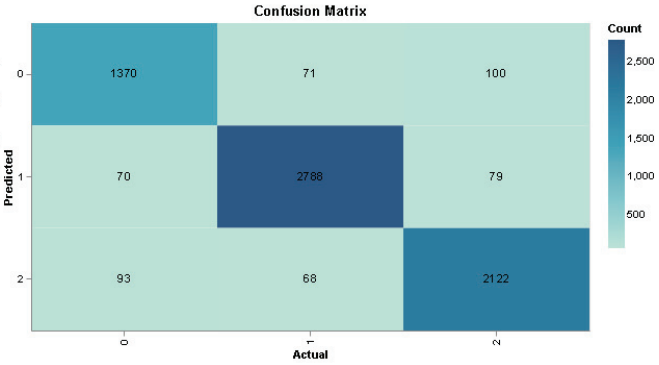


Fig. 9: Confusion matrix for DistilBERT model

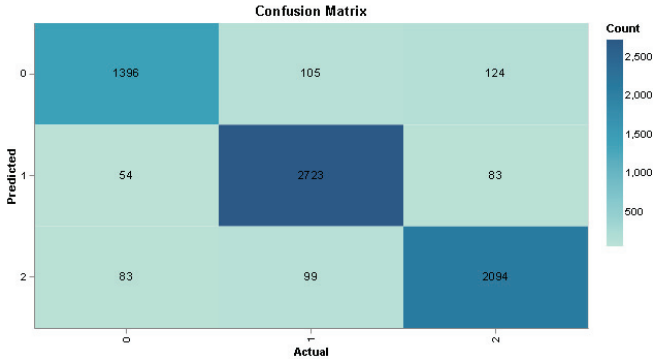


Fig. 10: Confusion matrix for XLNet model

V. Conclusion

Sentiment analysis is a way of analysing opinions and behaviour from text data; it can provide informative feedback and earnings to a business. According to Table I and the visualisations in the results and discussion section, all of the transformer models perform well on the news headline dataset, with the bert-base-cased model outperforming the others. In comparison to papers such as [23], [24], we see that our fine-tuned models outperform both traditional machine learning and deep learning models such as CNN, LSTM, and others.

Therefore, the transformers are the latest big thing in tasks like text classification, machine translation, language modelling, etc. For future work, the large number of hyperparameters available with simple transformers can be manipulated to allow our models to perform even better as good as human accuracy.

Acknowledgment

The author would like to thank Dr Goonjan Jain for her advice and assistance in writing this paper and all of the researchers whose research papers have been included in the reference section, which aided in the writing and presentation of this paper at the conference.

References

[1] "Transformers,". Accessed on: Feb. 1, 2021. [Online]. Available: https://huggingface.co/transformers/.

[2] C. Min, Z. Fan-wei, W. Ming-hui, and Y. Jing, "Survey of opinion mining," *Journal of Zhejiang University (Engineering Science)*, vol. 48, no. 8, pp. 1461–1472, 2014.

[3] J. Schectman, "Obama's campaign used salesforce.com to gauge feelings of core voters," *The Wall Street Journal*, para. 3, Dec. 7, 2012. [Online], Available: https://www.wsj.com/articles/BL-CIOB-1295. [Accessed Mar. 30, 2021].

[4] D. Dor, "On newspaper headlines as relevance optimizers," *Journal of Pragmatics*, vol. 35, no. 5, pp. 695–721, 2003.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:44:23 UTC from IEEE Xplore. Restrictions apply.

[5] M. Karlsson, "The immediacy of online news, the visibility of journalistic processes and a restructuring of journalistic authority," *Journalism*, vol. 12, no. 3, Apr., pp. 279—295, 2011.

[6] I. Salian, "SuperVize Me: What's the difference between supervised, unsupervised, semi-supervised and reinforcement learning?," *blogs.nvidia.com*, para. 4, 5, 15, Aug. 2, 2018. [Online]. Available: https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/. [Accessed Dec. 25, 2020].

[7] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," In Proc. Conference on Empirical Methods in Natural Language Processing: System Demonstrations '10, 2020, pp. 38—45.

[8] T. Rajapakse, "Simple transformers," 2019. [Online]. Available: https://www.simpletransformers.ai/. [Accessed Feb. 12, 2021].

[9] S. Fong, Y. Zhuang, J. Li, and R. Khoury, "Sentiment analysis of online news using MALLET," In Proc. 2013 International Symposium on Computational and Business Intelligence '08, 2013, pp. 301–304.

[10] J. Reis, F. Benevenuto, P. Olmo, R. Prates, H. Kwak, and J. An, "Breaking the news: First impressions matter on online news," In Proc. 9th International Conference on Web and Social Media '05, 2015, pp. 357-–366.

[11] N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs," In Proc. ICWSM 2007 - International Conference on Weblogs and Social Media '01, 2007.

[12] Y. Liu *et al.*, "RoBERTa: A robustly optimized bert pretraining approach," *ArXiv*, abs/1907.11692, 2019.

[13] M. Joshi *et al.*, "SpanBERT: Improving pre-training by representing and predicting spans," *ArXiv*, abs/1907.10529, 2020.

[14] Z. Yang *et al.*, "Xlnet: Generalized autoregressive pretraining for language understanding," *ArXiv*, abs/1906.08237, 2019.

[15] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, "COMET: Commonsense Transformers for Automatic Knowledge Graph Construction," In Proc. 57th Annual Meeting of the Association for Computational Linguistics '07, 2019, pp. 4762–4779.

[16] M. Lewis *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *ArXiv*, abs/1910.13461, 2019.

[17] G. Lample, and A. Conneau, "Cross-lingual Language Model Pretraining," *ArXiv*, abs/1901.07291, 2019.

[18] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation Functions: Comparison of trends in practice and research for deep learning," *ArXiv*, abs/1811.03378, 2018.

[19] T. Wood, "Softmax Function," *deepai.org*, para. 1, 3. [Online]. Available: https://deepai.org/machine-learning-glossary-and-terms/softmax-layer. [Accessed Mar. 1, 2021].

[20] L. Biewald, "Experiment tracking with weights and biases," wandb.com, 2020. [Online]. Available: https://www.wandb.com/. [Accessed Mar. 15, 2021].

[21] P. Joshi, "Transfer learning for NLP: Fine-Tuning BERT for text classification," analyticsvidhya.com, JULY 21, 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/07/transfer-learning-for-nlp-fine-tuning-bert-for-text-classification/. [Accessed Feb. 12, 2021].

[22] "Metrics and scoring: quantifying the quality of predictions," [Online]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html. [Accessed Feb. 12, 2021].

[23] C. J. Rameshbhai and J. Paulose, "Opinion mining on newspaper headlines using SVM and NLP," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 3, June, pp. 2152-2163, 2019.

[24] S. Jagtap, M. Sonwane, A. Shelke, Y. Bedmuthaont, and S. Gore, "Financial news analysis with NLP and machine learning," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 3, no. 4, Apr., pp. 623-628, 2021.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:44:23 UTC from IEEE Xplore. Restrictions apply.

Stacked Sparse Autoencoder and Machine Learning Based Anxiety Classification Using EEG Signals

Shikha*
Manan Agrawal*
Mohd Ayaan Anwar*
shikha_2k19phdco502@dtu.ac.in
mananagrawal_2k19co210@dtu.ac.in
mohdayaananwar_2k19co232@dtu.ac.in
Delhi Technological University (Formerly DCE)
New Delhi, India

Divyashikha Sethia
Department of Software Engineering
Delhi Technological University (Formerly DCE)
New Delhi, India
divyashikha@dtu.ac.in

ABSTRACT

Anxiety is an emotion characterized by trepidation, stress, or uneasiness that involves extreme worry or fear over future unwanted events or an actual situation. Careful analysis for anxiety is critical since approximately 2 to 4% of the general population have experienced adequate symptoms indicating an anxiety disorder. This paper aims to classify anxiety levels based on machine learning and deep learning algorithms with improved performance. This work uses the publically available DASPS Database (Database for Anxious States based on a Psychological Stimulation). The dataset consists of EEG recordings from 23 participants during anxiety elicitation through face-to-face psychological stimuli. This work uses RFECV with the classifiers to reduce redundancy between features and improve results. We achieve the highest classification accuracy of 83.93% and 70.25% using Stacked Sparse Autoencoder and Decision Tree for two-class anxiety classification.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**; **Feature selection**; **Supervised learning by classification**; *Neural networks*; Batch learning; **Regularization**; Ensemble methods.

KEYWORDS

signal processing, machine learning, feature selection, anxiety detection, stacked sparse autoencoders

ACM Reference Format:
Shikha, Manan Agrawal, Mohd Ayaan Anwar, and Divyashikha Sethia. 2021. Stacked Sparse Autoencoder and Machine Learning Based Anxiety Classification Using EEG Signals. In *The First International Conference on AI-ML-Systems (AIMLSystems '21), October 21–23, 2021, Bangalore, India.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3486001.3486227

*These authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
AIMLSystems '21, October 21–23, 2021, Bangalore, India
© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8594-7/21/10…$15.00
https://doi.org/10.1145/3486001.3486227

1 INTRODUCTION

Anxiety disorder and stress are highly prevalent and debilitating conditions that affect the daily life of an individual. These are the two states of emotional strain which many people have experienced over the last few years in day-to-day life. Anxiety disorder is the persistent and excessive worry that does not fade away even in the absence of a stressor, and it exists for an extended period. Various situations can trigger anxiety, such as having a job interview before making an important decision, chronic life-threatening disease, competitive exam, or painful medical procedure. According to the Anxiety and Depression Association of America (ADAA), anxiety disorder is the most common mental illness in the United States of America. 18.1% of the population is affected by anxiety disorder every year, and they are six times more likely to be hospitalized for a psychiatric disorder than normal people [2].

Physiologically, Positive stress is short-term stress that makes a person concentrate more on the problem and cope with it constructively. However, long-term and repetitive stress leads to anxiety, which causes damaging emotional and physical reactions in the body. The effects of anxiety disorder are visible in the form of physiological, neurobiological, and psychological parameters. Researchers show that anxiety disorder is strongly associated with cardiac disease and increases the risk of stroke and heart attack [10]. There are several types of Anxiety disorders such as General Anxiety Disorder (GAD), Social Anxiety Disorder (SAD), Obsessive-Compulsive Disorder (OCD), Post-traumatic Stress Disorder (PTSD), Phobias, Panic Disorder (PD) [21]. Table 1 shows the various symptoms of anxiety disorder.

Previously, clinicians had to assess anxiety symptoms by relying on Patients' self-report and measures. However, many researchers survey that anxiety and stress can be quantified objectively by various biomarkers such as heart rate, heart rate variability, brain activity [23], and cortisol [13]. There have been increasing reports showing that Electroencephalogram (EEG) based methods help measure a person's cognitive state for mental health diagnosis and explored various techniques to detect anxiety disorder. The advantage includes that EEG is a non-invasive method with high time resolution, great accessibility due to low cost, and does not evolve radiation exposure. Giannakakis et al. [11] present the effect of anxiety on EEG signal using DEAP Dataset. The EEG data is collected from 32 participants while watching 40 one-minute long videos. The author analyze the effects of anxiety based on EEG

Table 1: Types of Anxiety Disorder and its Physical Symptoms

Anxiety Disorder Types	Symptom
General Anxiety disorder (GAD)	Fatigue, Sweating, Muscle pain, Irritability, Nervousness, Difficulty in falling asleep and Feeling twitchy
Social Anxiety Disorder (SAD)	Blushing, Sweating, Upset stomach, Blank out, Shortness of breath, Dizziness, Fast heartbeat.
Obsessive-Compulsive Disorder (OCD)	There is no physical sign of OCD. A person with OCD can develop physical problems like a germophobic may wash his hands frequently, leading to raw, red, and painful skin.
Post-traumatic Stress Disorder (PTSD)	Sweating, Headaches, Tiredness, Sleep disturbances, Weakened Immune system, Chest pain, Stomach ache.
Panic Disorder (PD)	Accelerated heart rate, Shortness of breath, Dry mouth, Abdominal cramp, Feeling faint or Dizzy, Chest pain, Nausea, Chills, Trembling, or shaking.
Phobia	Confusion, feeling faint, Sensation of butterflies in the stomach, Choking sensation, Trembling, Tightness in Chest, Shortness of breath

signals and concludes that frontal asymmetry has reduced during anxiety state compared with a relaxed state. Frontal asymmetry is a typical indicator of asymmetric brain activity in the frontal cortex, which refers to asymmetrical activity between the left and right hemispheres. They further analyzes the effects of anxiety based on EEG signals and investigates EEG features and dataset reduction procedures that lead to efficient anxiety detection. Arsalan et al. [3] propose a multiclass classification of trait anxiety detection framework using resting-state electroencephalography (EEG) data. The study collects EEG readings from 65 participants for two minutes in an open eye state. The study uses the State-Trait Anxiety Inventory (STAI) questionnaire to measure the anxiety score and classified two-label EEG data into two and three-class anxiety levels. Further, the authors propose the impact of feature selection and channel selection for a trait anxiety detection framework. Chaitanya et al. [8] proposed a prototype for the diagnosis and treatment of anxiety-related disorders based on neurological consideration. The authors consider RGB light-emitting diodes (LED) and single-channel electroencephalography (EEG) Headband on Arduino driven platform to detect the presence of anxiety. The result shows strength in delivering a guaranteed diagnosis and remedying anxiety disorders. Several studies show that other than adults, students are too affected by stress, depression, and anxiety due to academic and social pressure. Meng et al. [17] present the Self-made and Standard datasets and used the Takagi-Sugeno-Kang (TSK) fuzzy system and deep features to diagnose anxiety levels among college students automatically.

Baghdadi et al. [6] proposed DASPS: 'A Database for Anxious States based on a Psychological Stimulation'. It contains recorded EEG signals of 23 participants during anxiety elicitation through six different situations for each subject. The author implements K-Nearest Neighbour (KNN), Support Vector Machine (SVM), and Stacked Sparse Autoencoder for 2 class and 4 class anxiety classification. Our work aims to enhance the two-class anxiety level classification performance using the classifier and deep learning algorithms. We extract 56 Time Domain, 126 Frequency Domain, and 336 Time-frequency Domain features from the pre-processed EEG signal. On contrasting the results, this work observes a maximum classification accuracy score of 83.98% and 70.25% using Stacked Sparse autoencoder and Decision Tree, respectively, which is significantly better than the earlier proposed results [6].

The organization of the remaining part of the paper is as follows: Section 2 represents the methodology of our work, followed by results and discussion for anxiety classification in section 3, and section 4 presents the conclusion.

2 METHODOLOGY

This paper aims to enhance the accuracy of two class anxiety level classification with reduced features using the classifier and deep learning algorithms. Figure 1 shows a block diagram of a general process of Anxiety classification, divided into five blocks: 1) Data acquisition from DASPS Database, 2) Data Preprocessing, 3) Feature extraction 4) Anxiety Classification.

2.1 Dataset Description

This paper uses the publicly available DASPS Dataset for anxiety level detection collected by Asma Baghdadi et al.[6]. The database contains raw EEG data in .edf format collected from the 23 participants, where 10 are men and 13 are women with the mean age of 30 years. The EEG signal is acquired using Emotiv EPOC, the wireless EEG headset with 14 channels and 2 mastoids with a 128Hz sampling rate. Along with the raw data, the database includes pre-processed data in the .mat format. During the recording of the EEG signal, the study presents six different situations to the subjects with closed eyes and minimized gestures. It divides each situation into two parts: In the first 15 seconds, the psychotherapist recites the situation, and in the next 15 seconds, the subject recalls the situation. The complete process for each subject proceeds as follows:

- The first phase includes 5 minutes of Hamilton Pre-Stimuli.
- In the second phase, the psychotherapist recites the situation for the first 15 seconds, and the subject recalls the situation in the next 15 seconds. The complete process repeats for each of the six situations for all the subjects.
- At the end of the six situations, Self Assessment Manikin (SAM) is used to rate how a subject felt during stimulation.

A database stores the SAM rating, consisting of two columns: Valence and Arousal [6].

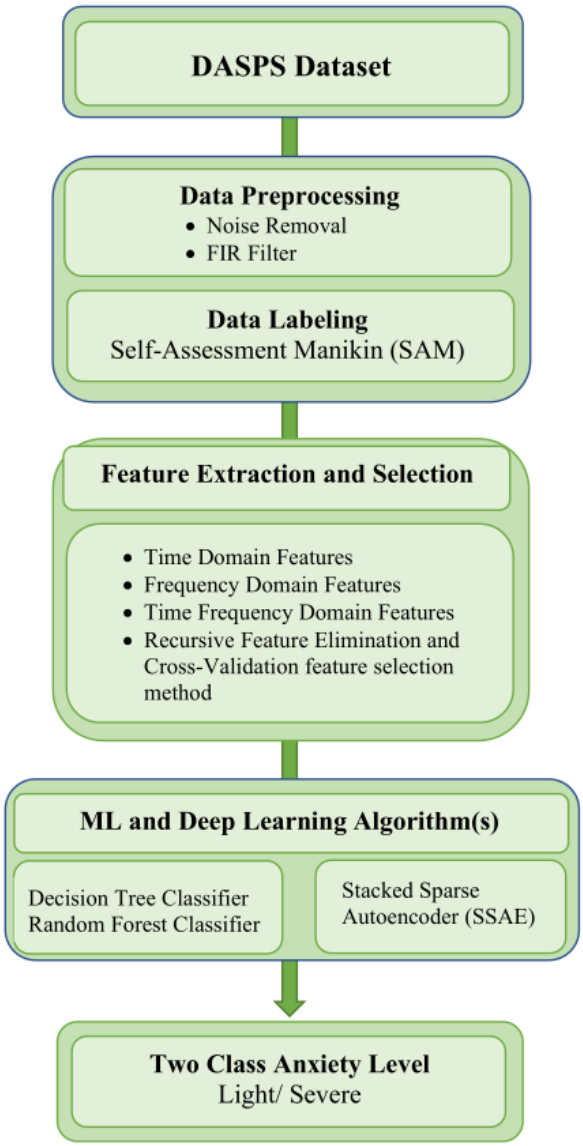


Figure 1: Anxiety Classification Framework

2.2 Data Preprocessing

EEG signals are electrical signals contaminated with various artifacts such as electrode misplacement, subject movement, environmental noise, or biological artifacts. Depending on the origin, the artifacts are classified into two categories:

(1) *Extrinsic artifacts:* Technical issues and some external factors in the data collecting environment are the causes of extrinsic artifacts such as electrode misplacement, cable movements and powerline interference (50-60 Hz).
(2) *Physiological artifacts:* these artifacts are the physiological signals originated from the body, such as Electromyogram(EMG) artifacts (muscle activities), Electroculogram (EOG) artifacts (eye movements or blinks), Electrocardiogram (ECG) artifacts (cardiac activities).

We apply Independent Component Analysis (ICA) to eliminate Physiological artifacts using the FastICA algorithm. A 5th Order Butterworth Bandpass filter with a frequency range of 4 - 44 Hz is applied to the raw EEG signal to improve the data quality and remove unwanted frequencies.

Data labeling and clipping: The duration of the experiment is about six minutes, divided into six different situations for each subject. Baghdadi et al. [6] consider the first 30 seconds of each trial and remove the rest 30 seconds of SAM during this step. This work follows the same procedure and considers the first 30 seconds of each situation and clips the data accordingly. After completing the experiment for each subject, the participant classifies and rates the excited (Arousal) and feeling (Valence) emotions throughout the stimulation. The arousal and valence score help classify the anxiety levels by using the algorithm mentioned in Baghdadi et al. [6].

2.3 Feature Extraction

The EEG features are of three categories: Time-domain features, Frequency-domain features, and Time-Frequency domain features.

2.3.1 *Time Domain Features.*

- *Hjorth Features:* Three different Hjorth parameters describe an EEG signal in the time domain, namely Activity, Mobility, and Complexity. Assuming y(t) represents the signal, then Equations (1), (2), and (3) denote the Hjorth parameters.

$$Activity = var(y(t)) \tag{1}$$

$$Mobility = \sqrt{\frac{var(\frac{dy(t)}{dt})}{var(y(t))}} \tag{2}$$

$$Complexity = \frac{Mobility(\frac{dy(t)}{dt})}{Mobility(y(t))} \tag{3}$$

The three parameters are computed for each of the 14 EEG channels to get 42 parameters per window (14 channels x 3 parameters).

- *Higuchi Fractal Dimensions:* Fractal dimensions measure the complexity of a time series [1]. Discretized signals are analyzed as time sequences X(1), X(2), ... , X(N). The resultant time series is given by Equation (4).

$$X_k^m : X(m), X(m+k), ..., X(m) + \frac{N-m}{k} \cdot k \tag{4}$$
$$(m = 1, 2, ..., k)$$

where, m denotes the starting time and k denotes the time interval. k sets of $L_m(k)$ are then calculated using Equation (5).

$$L_m(k) = \frac{\sum_{i=1}^{\frac{N-m}{k}} |X(m+ik) - X(m+(i-1)\cdot k)| \frac{N-1}{\frac{N-m}{k}\cdot k}}{k} \tag{5}$$

where $L(k)$ represents the mean value of $L_m(k)$. The relationship between $L(k)$ and k is given by Equation (6) [14].

$$\langle L(k) \rangle \propto k^{-D} \tag{6}$$

The fractal dimension for each of the 14 channels is computed to get the corresponding 14 Higuchi FD features.

2.3.2 Frequency Domain Features.

- *Relative Bandpower:* The Relative Bandpower is the ratio of the specific band power of the frequency band by the total power in all the bands of the EEG signal. This measure can provide new insight into the signal's power spectrum and characterize a particular band's contribution corresponding to all the bands. The work computes the absolute power of each of the four frequency bands- Theta, Alpha, Beta, and Gamma, and the total power by estimating the Power Spectral Density (PSD) using Welch's method. The Relative Power is computed from the Absolute Power using Equation (7).

$$Relative\ Power = \frac{Absolute\ Power\ of\ Band}{Total\ Power} \tag{7}$$

The four parameters are computed for each of the 14 EEG channels to get 56 parameters per window (14 channels x 4 parameters).

- *Alpha-Beta Ratio:* Alpha-Beta Ratio is computed by dividing the total power of the Alpha band by the total power of the beta band and is calculated as in Equation (8) [18].

$$Alpha - Beta\ Ratio = \frac{Absolute\ Power\ of\ Alpha\ Band}{Absolute\ Power\ of\ Beta\ Band} \tag{8}$$

The alpha-beta ratio is computed for each of the 14 EEG channels to get the corresponding 14 alpha-beta ratios.

- *Theta-Alpha Ratio:* Theta-Alpha Ratio is the ratio of the total power of the Theta band to the total power of the Alpha band, as calculated in Equation (9).

$$Theta - Alpha\ Ratio = \frac{Absolute\ Power\ of\ Theta\ Band}{Absolute\ Power\ of\ Alpha\ Band} \tag{9}$$

The theta-alpha ratio is computed for each of the 14 EEG channels to get the corresponding 14 theta-alpha ratios.

- *Spectral Entropy:* Spectral Entropy represents the complexity of the system. It is a form of Shannon Entropy that uses normalized PSD (Power Spectral Density) of the EEG signal [15]. The computation of the Spectral Entropy represents by Equation 11.

$$H(y(t)) = -\sum_{f=0}^{f_s/2} PSD(f)\log_2[PSD(f)] \tag{10}$$

where f_s is the sampling frequency of the signal. The spectral entropy is computed for each of the 14 EEG channels to get the corresponding 14 spectral entropy values.

- *Sample Entropy:* Sample entropy (SampEn) is an improvement over approximate Entropy (ApEn), reducing the bias caused by self-matching. It estimates the complexity of physiological time-series signals [20]. Equation 11 presents the computation of the Sample Entropy.

$$H(y, m, r) = -\log_e\left(\frac{C_{m+1(r)}}{C_m(r)}\right) \tag{11}$$

where, m is the embedding dimension (= order), r is the radius of the neighbourhood, $C_{m+1}(r)$ is the number of embedded vectors of length $m+1$ having a Chebyshev distance inferior to r and $C_m(r)$ is the number of embedded vectors of length m having a Chebyshev distance inferior to r. The sample entropy is computed for each of the 14 EEG channels to get the corresponding 14 sample entropy values.

2.3.3 Time-Frequency Domain Features.

- *Hilbert Huang Transform:* Hilbert–Huang transform is used to analyze non-linear and non-stationary signals. It involves two steps where the first step of HHT is to decompose the original signal into a finite number of intrinsic mode functions (IMFs); this process is called Empirical Mode Decomposition. The signal is decomposed into IMFs so that the first IMF captures the highest frequency component of each event in the signal. The second step includes the Hilbert transform, where, for each IMF, generates a pair whose phase is forward by $\frac{\pi}{2}$ [19]. IMFs are calculated till the fourth level for each channel to get IMF0, IMF1, IMF2, IMF3, and IMFR values. Then the Average Instantaneous Frequency, Variance, Maximum, and Shannon Entropy are computed to get a total of 280 parameters (14 channels x 5 IMFs x 4 values per IMF).

- /textitDiscrete Wavelet Transform (DWT): DWT is widely used in biomedical signals for time-frequency analysis, especially in an EEG signal, because of its non-stationary characteristics. In DWT, a series of high pass and low pass filters apply to the signal. The signal decomposes into various coefficients known as frequency bands. The threshold is applied to remove the signal artifacts, and then the artifact-free signal is added to reconstruct the EEG Signal. In our work, Root Mean Square (RMS), a statistical feature, is calculated from the decomposed coefficients of level 4 using the db5 wavelet. Equation 12 computes the RMS value of a signal.

$$RMS = \sqrt{\frac{1}{n}\sum_i x_i^2} \tag{12}$$

The RMS values of the four decomposition levels are computed for each of the 14 EEG channels to get 56 parameters per window (14 channels x 4 parameters).

2.4 Classification Algorithms

This paper applies Random Forest, Decision Tree, and Stacked Sparse Autoencoder for binary classification of anxiety levels.

- *Decision Tree:* The Decision Tree belongs to supervised learning algorithms and is used to solve classification problems and regression problems. A decision tree represents a tree-structured classifier, comprising two types of nodes: Decision nodes and Leaf nodes, used to make decisions. Decision nodes split the data, and leaf nodes are the decisions or the final outcomes. To obtain the best Decision Tree classification model, we have investigated the different values for criterion (gini and entropy), max_depth (maximum depth of the tree), and min_samples_leaf (minimum number of samples

Table 2: Machine Learning Results

Features	Decision Tree Accuracy (%)	Random Forest Accuracy (%)
Time Domain	**70.25**	61.54
Frequency Domain	**69.16**	64.29
Time-Frequency Domain	65.56	**66.70**
All Features	**67.00**	63.33

Table 3: SSAE Results

Features	No. of Features	Accuracy (%)
Time Domain	56	73.21
Optimal Time Domain (using RFECV)	31	71.42
Frequency Domain	112	80.35
Optimal Frequency Domain (using RFECV)	42	80.35
Time-Frequency Domain	336	75.00
Optimal Time-Frequency Domain (using RFECV)	336	75.00
All Features	504	83.98
Optimal All Features (using RFECV)	318	83.93

required to be at a leaf node). The max_depth values investigated ranged from 2 to 19 inclusive, while min_samples_leaf could be 1 or 2.

- *Random forest:* Random Forest is an ensemble tree-based machine learning method that reduces variance in the data. It considers the collection of decision trees in order to form a forest. This algorithm is also known as a group of decision tree algorithms used for classification and regression problems. Random forest follows specific rules for tree growing, tree combination, self-testing, and post-processing, and it always tries to find the best feature among all the features randomly. This algorithm is robust to overfitting and does not decrease the test performance as the number of trees increases. For the Random Forest classification model, we have investigated different values of the number of trees in the forest (from 100 to 200), maximum depth of a tree(2 to 20 inclusive), the maximum number of features(square root and log base 2 of the total number of features) for both the gini and entropy criterion.
- *Stacked Sparse Autoencoder:* An autoencoder is a neural network that is an unsupervised learning algorithm trained to copy its input to its output [12]. It consists of three layers categorized into three classes: input, hidden, and output. The network minimizes the reconstruction error to extract the best expression of the hidden layer by attempting to make the output layer equal to the input layer. Sparse autoencoders are an extension of the auto-encoder, in which sparse restrictions are introduced to the hidden nodes to control the number of activated neurons. As a result, sparse autoencoder can learn better features as the system complexity reduces parameters due to fewer activated neurons. A Stacked Sparse Autoencoder (SSAE) is created by stacking multiple sparse autoencoders such that the output of the first sparse autoencoder becomes the input of the second autoencoder, and so on [22].

3 RESULTS AND DISCUSSION

The novel contribution of this paper is the performance enhancement with the reduced number of features compared to Baghdadi et al. [6]. The anxiety levels of each subject are classified into four classes: Normal, Light, Moderate, and Severe, using the Self Assessment Manikin (SAM) scores. The four-class classification leads to unbalanced data, which affects the classification accuracies as found by Baghdadi et al. [6]. Consequently, this work bifurcates the data into two classes: Normal and Light anxiety states are grouped into 'Class-I' and Moderate and Severe into 'Class-II'.

Table 2 presents the anxiety classification performance for two-class levels using machine learning algorithms. This work applies GridsearchCV to obtain the optimal parameters for the Decision Tree and Random Forest classifiers. Then the results are verified using Stratified KFold Cross-Validation to ensure unbiased training models. We observe that the accuracy score increases by reducing the number of features using RFECV. Baghdadi et al.[6] have achieved the highest accuracy of 67% using KNN. In comparison, it is evident from the results that the Decision Tree classifier outperformed the best performance of the machine learning algorithm achieved by Baghdadi et al. [6]. In this work, the Decision Tree classifier performs remarkably well for Hjorth features (time-domain) and Power (frequency domain). For 2 class anxiety level classification, we achieve the maximum accuracy score of 70.25%, 69.16%, and 67.00% for the time domain, frequency domain, and all features combination, respectively.

Table 3 demonstrates the improved performance using Stacked Sparse Autoencoder (SSAE) with RFECV. A stacked sparse autoencoder (SSAE) is created by stacking multiple sparse autoencoders

Table 4: Result Comparison

Features	Our Results			**Baghdadi et al.** [6]	
	No. of Features	Best ML Classifier	SSAE	Best ML Classifier	SSAE
Time Domain	56	**70.25**	**73.21**	66.30	67.90
Frequency Domain	112	**69.16**	80.35	66.30	82.00
Time-Frequency Domain	336	66.67	**75.00**	67.00	67.10
All Features	318	**67.00**	**83.98**	-	83.50

such that the output of the first sparse autoencoder becomes the input of the second autoencoder, and so on [22]. In this work, we have set the train-test split ratio as 80:20. The two hidden sparse layers (Layer-1 and Layer-2) size is according to the number of features in the input layer. The model uses the 'ReLU' activation function for both hidden layers. The output layer, which consists of a single node, uses the 'Sigmoid function' as its activation function. The model compiles using the 'Adam' optimizer and binary cross-entropy as the loss function. We apply the Recursive feature elimination feature selection technique with cross-validation (RFECV), which removes highly correlated features and improves the performance with fewer features. The model achieves the highest accuracy of 83.93% for two-label classification problems with all features combinations.

In this paper, we compare the obtained results with the best results evaluated in the work of Baghdadi et al. [6]. The main reason to compare our work with Baghdadi et al. is that both worked on the same dataset. Table 4 demonstrates the comparison with the best results achieved by Baghdadi et al. [6]. We obtain the highest accuracy score of 70.25% using the Decision Tree on time-domain features. In contrast, Baghdadi et al. [6] reports the maximum accuracy score of 67% using the KNN machine learning algorithm. Other than the machine learning algorithm, we have obtained the highest accuracy score of 73.21%, 80.35%, 75.00%, and 83.98% using Stacked Sparse Autoencoder for time domain, frequency domain, time-frequency domain, and all features combination, respectively. As compared to the accuracy obtained by Baghdadi et al. [6], this paper improves the accuracy close to 5%, 8%, and 0.5% for time domain,time-frequency domain, and all features combination, respectively.

4 CONCLUSION

This paper presents an improved performance for two-class anxiety classification using machine learning and deep learning algorithms on DASPS Database. We extract several signal processing features from EEG signals and used the Recursive Feature Elimination (RFECV) method to select the most dominant features based on their importance and correlation. The selected subgroup of features is input into three distinct classifiers: Decision Tree, Random Forest, and Stacked Sparse Autoencoder. In conclusion, results indicate that after applying RFECV, we achieve the highest accuracy with fewer features. Decision Tree and Stacked Sparse Autoencoder provide the best results for two-class anxiety classification and improved accuracy close to 5%, 8%, and 0.5% for time domain, time-frequency domain, and all features combination. In future work, an extensive dataset is required to validate results and create a generalized anxiety disorder classification model.

REFERENCES

[1] Ali H Al-Nuaimi, Emmanuel Jammeh, Lingfen Sun, and Emmanuel Ifeachor. 2017. Higuchi fractal dimension of the electroencephalogram as a biomarker for early detection of Alzheimer's disease. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2320–2324.

[2] Anxiety and Depression Association of America. 2021. Anxiety and Depression Association of America: Facts & Statistics. Retrieved from https://adaa.org/understanding-anxiety/facts-statistics.

[3] Aamir Arsalan and Muhammad Majid. 2021. A study on multi-class anxiety detection using wearable EEG headband. *Journal of Ambient Intelligence and Humanized Computing* (2021), 1–11.

[4] Omneya Attallah. 2020. An effective mental stress state detection and evaluation system using minimum number of frontal brain electrodes. *Diagnostics* 10, 5 (2020), 292.

[5] Asma Baghdadi, Yassine Aribi, Rahma Fourati, Najla Halouani, Patrick Siarry, and Adel Alimi. 2020. Psychological stimulation for anxious states detection based on EEG-related features. *Journal of Ambient Intelligence and Humanized Computing* (2020), 1–15.

[6] Asma Baghdadi, Yassine Aribi, Rahma Fourati, Najla Halouani, Patrick Siarry, and Adel M Alimi. 2019. DASPS: A Database for Anxious States based on a Psychological Stimulation. *arXiv preprint arXiv:1901.02942* (2019).

[7] Alexander Bystritsky, Sahib S Khalsa, Michael E Cameron, and Jason Schiffman. 2013. Current diagnosis and treatment of anxiety disorders. *Pharmacy and Therapeutics* 38, 1 (2013), 30.

[8] M Nair Chaitanya, Sharmila Jayakkumar, Edwin Chong, and Chen-Hua Yeow. 2017. A wearable, EEG-based massage headband for anxiety alleviation. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 3557–3560.

[9] Vikrant Doma and Matin Pirouz. 2020. A comparative analysis of machine learning methods for emotion recognition using EEG and peripheral physiological signals. *Journal of Big Data* 7, 1 (2020), 1–21.

[10] Connor A Emdin, Ayodele Odutayo, Christopher X Wong, Jenny Tran, Allan J Hsiao, and Benjamin HM Hunn. 2016. Meta-analysis of anxiety as a risk factor for cardiovascular disease. *The American journal of cardiology* 118, 4 (2016), 511–519.

[11] Giorgos Giannakakis, Dimitris Grigoriadis, and Manolis Tsiknakis. 2015. Detection of stress/anxiety state from EEG features during video watching. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 6034–6037.

[12] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

[13] Karin Hek, Nese Direk, Rachel S Newson, Albert Hofman, Witte JG Hoogendijk, Cornelis L Mulder, and Henning Tiemeier. 2013. Anxiety disorders and salivary cortisol levels in older adults: a population-based study. *Psychoneuroendocrinology* 38, 2 (2013), 300–305.

[14] Tomoyuki Higuchi. 1988. Approach to an irregular time series on the basis of the fractal theory. *Physica D: Nonlinear Phenomena* 31, 2 (1988), 277–283.

[15] Tsuyoshi Inouye, Kazuhiro Shinosaki, H Sakamoto, Seigo Toi, Satoshi Ukai, Akinori Iyama, Y Katsuda, and M Hirano. 1991. Quantification of EEG irregularity by use of the entropy of the power spectrum. *Electroencephalography and clinical neurophysiology* 79, 3 (1991), 204–210.

[16] Daksh Maheshwari, SK Ghosh, RK Tripathy, Manish Sharma, and U Rajendra Acharya. 2021. Automated accurate emotion recognition system using rhythm-specific deep convolutional neural network technique with multi-channel EEG signals. *Computers in Biology and Medicine* 134 (2021), 104428.

[17] Xiangmin Meng and Jie Zhang. 2020. Anxiety Recognition of College Students Using a Takagi-Sugeno-Kang Fuzzy System Modeling Method and Deep Features. *IEEE Access* 8 (2020), 159897–159905.

[18] Andrew J Niemiec and Brain J Lithgow. 2006. Alpha-band characteristics in EEG spectrum indicate reliability of frontal brain asymmetry measures in diagnosis of depression.. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. IEEE, 7517–7520.

[19] Sonal K. Phan and Cathy Chen. 2017. 9 - Big Data and Monitoring the Grid. In *The Power Grid*, Brian W. D'Andrade (Ed.). Academic Press, 253–285. https://doi.org/10.1016/B978-0-12-805321-8.00009-4

[20] Joshua S Richman and J Randall Moorman. 2000. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology* 278, 6 (2000), H2039–H2049.

[21] JK Trivedi and Pawan Kumar Gupta. 2010. An overview of Indian research in anxiety disorders. *Indian Journal of Psychiatry* 52, Suppl1 (2010), S210.

[22] Wei Zhao, Qing-Hao Meng, Ming Zeng, and Pei-Feng Qi. 2017. Stacked sparse auto-encoders (SSAE) based electronic nose for Chinese liquors classification. *Sensors* 17, 12 (2017), 2855.

[23] Yali Zheng, Tracy CH Wong, Billy HK Leung, and Carmen CY Poon. 2016. Unobtrusive and multimodal wearable sensing to quantify anxiety. *IEEE Sensors Journal* 16, 10 (2016), 3689–3696.

2021 8th International Conference on Signal Processing and Integrated Networks (SPIN) | 978-1-6654-3564-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SPIN52536.2021.9566012

State Estimation Along WLS-Phasor Measurements in Power System

Ankit Kumar Dhaikar
Power Systems
Delhi Technological University
ankitkumardhaikar_2k19psy02@dtu.ac.in

S.T. Nagarajan
Electrical Engineering
Delhi Technological University
stnagarajan@dce.ac.in

***Abstract*— A procedure of approximating extent of state along with voltage angle built on measurements that have been agreed for every bus of system only at the buses is the Power System State Estimation. The power system state cannot be estimated only by using PMU measurements as PMUs are expensive. Along with the old measurement methods, to estimate the power system's state, phasor measurements are being used as additional measurement. A MATLAB program coded, PMU measurements described along with a simulation has been carried out on bus system's IEEE-30 and IEEE-14 for the technique's verification. A unique estimator model is used in the technique. The technique sets the estimate in current measurements, through post procession in the PMU voltage as well as for the state from the Weight Least Square (WLS) method. Using the measurements which are traditional, the estimation is done by the model in polar coordinates using WLS state estimation. The system's final state is estimated by expressing the PMU measurements along with the states in rectangular coordinates.**

Keywords*— Phasor Measurement Unit- PMU, Weight Least Square- WLS, State Estimator

I. Introduction

A. Power System State Estimation

In system of power, for every bus, at a certain time the voltage's magnitude and approximation of the angles there is the method of power state's estimation. Every bus-voltage system synchronized phasor measurements is used to achieve this.

PMUs being very costly and are also likely of having errors in the measurement. A more likely approach for the state's approximation would be to find an optimal estimate and for the filtration by using a set of superfluous measurements [1]. The power that these measurements have is traditional and for the voltage measurements. The current's magnitude or the phasor voltage measurements are synchronized. In the state of system, the bus-voltage-phasors is included in the definition of the state in the steady state.

The state's estimator is nothing but an online-security-analysis's core. The working of state estimator for various applications such as contingency-analysis, load-flow etc., is basically as a filter for giving the current states a dependable system [2]. The raw measurements include current, voltage, injections of the power along with the measurements of the flow of power.

Using topology-processor for the generation of a bus/branch-model in a system of power as shown in fig 1. the switch positions as well as the circuit breaker in powerhouse have been processed. For example, repeatedly used in WLS method, different techniques for state's estimation is used by the state's estimators of every bus in power system.

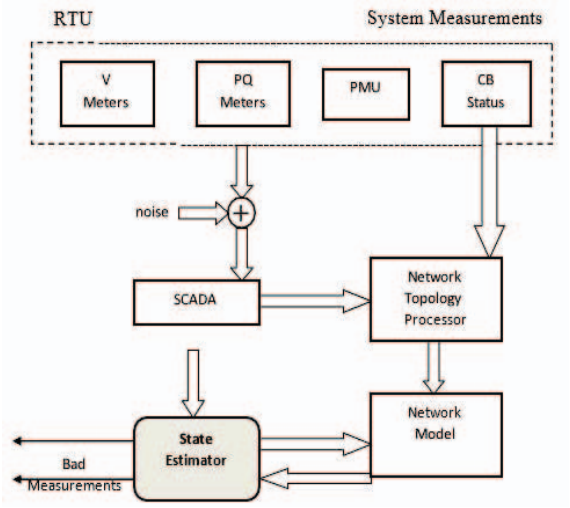


Fig 1 State Estimation Block Diagram

B. Synchronized Phasor Measurement Units (PMUs)

The current phasor's positive sequence in line is used by PMU along with the GPS signal's positive sequence in the phasors of the voltage for the synchronization at the network buses [3]. The real-time snapshot of power system state is given by a group of measurements and the correctness has been better than fraction of a second for the synchronization. One of the major problems of state approximation could be solved with the involvement of the phasor-measurements widely.

The state system's measurement procedure is result of the help from these measurements. Non-linear states functions used for the approximation. Being very costly, PMU has been taken as additional measurement along with the conventional ones. The working of typical PMU is shown as a block diagram in Fig 2.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:40:49 UTC from IEEE Xplore. Restrictions apply.

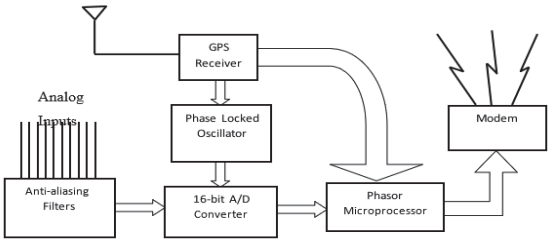


Fig.2 Functional block diagram of a typical PMU

C. State estimation using Phasor measurements

For the completion of the goal, the quantity required cannot be given by PMU measurements, even though it is a dependable technique for the old-style-state-estimators. Other measurements are more benefited from phasor measurements enhancing the correctness of the state-estimate [3]-[4]. Different methods have been accepted for including in traditional measurements the phasor measurements. An easy and uncomplicated way is to add the voltages along with the currents in the phasor measurements as additional to the traditional set of measurement. There is a requirement of eminent refurbishment in the non-linear state-estimator and requires changes in the authentic EMS-software. In other case, we can leave the software as it is and just embody the phasor measurements along with the traditional style estimator in the post-processed linear estimator.

II. WLS STATE ESTIMATION USING PHASOR MEASUREMENTS

A. Introduction

There is proposal of different methods for the amalgamation in traditional WLS algorithm of the phasor measurements. For controlling the center computer, the working state estimators in the contemporary control centers work as a feature for large software package belonging to EMS role. The phasor V along with I measurements is boarded in set of input in state estimation software where transformation takes place in respectively P and Q measurements. Additional Phasor measurements have been one of the best and effective technique that is post processed [5].

B. WLS State Estimation with Phasor Measurements

Assuming that, at bus injections and at the voltage magnitude buses, the reactive power and the active power flowing into the network elements consists of non-synchronized data containing 1z set of measurements. Supposedly, temporary techniques have contributed to the eradication of poor data from the measurement. At all of network buses, the set of voltages is the state vector X whose measurements with non-linear role have positive sequence.

$$z_1 = h_1(x) + e_1$$

In which the 'h_1' has been the nonlinear function of state vector x that has been presented in the polar coordinates as well as the e_1 in error of the measurement vector that has a covariance matrix R_1.

The Jacobian matrix, H_1 given by [6],

$$H_1(x) = \frac{\partial h_1(x)}{\partial x} \quad (1)$$

The Gain matrix, $G_1(x^k)$ given by

$$G_1(x^k) = \left[H_1^T(x_k) R_1^{-1} H_1(x_k)\right]^{-1} \quad (2)$$

The error covariance matrix of approximation x is given by

$$Con([x]) = [H_1^T R_1^{-1} H_1] \quad (3)$$

And the state vector is obtained from,

$$[x^{k+1}] = [x^k] + [G_1(x^k)]^{-1} G_1(x^k)[H_1^T R_1^{-1}][z_1 - h_1(x^k)] \quad (4)$$

This method carries on till one of two circumstances has been fulfilled, also known as reiterative method. The circumstance one which need to be fulfilled is surpassing the permissible reiterations. The second circumstance would be the falling of changes in a bearable range of the state variables.

$$max|\triangle x^k| \leq \varepsilon$$

The method of post-processing

Using rotation matrix, changing state vector attained from equation 4 which is in the polar co-ordinates and changing into the co-ordinates that are rectangular given by,

$$K = \begin{bmatrix} \cos\delta_1 & 0 & 0 & -|E_1|\sin\delta_1 & 0 & 0 \\ 0 & \cos\delta_2 & 0 & 0 & -|E_2|\sin\delta_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \sin\delta_1 & 0 & 0 & |E_1|\sin\delta_1 & 0 & 0 \\ 0 & \sin\delta_2 & 0 & 0 & |E_2|\cos\delta_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (5)$$

Where, $|E_1|, |E_2| \ldots |E_n|$ are magnitudes and $|\delta_1|, |\delta_2|, \ldots \ldots |\delta_n|$ are respective angles.

And the converted state vector becomes,

$$\begin{bmatrix} V_R \\ V_I \end{bmatrix}_{SE} = [K][x] \quad (6)$$

Let a group of the voltage having positive sequence as well as the current phasors be, z_2 . Error measurement covariance matrix of phasor measurements has been supposed of being, R_2. Both measurement phasors as well as the error covariance matrix are in polar coordinates, and hence, they must be transformed to rectangular coordinates.

$$\begin{bmatrix} \begin{bmatrix} V_R \\ V_I \end{bmatrix}_{PM} \\ \begin{bmatrix} I_R \\ I_I \end{bmatrix}_{PM} \end{bmatrix} = [K] \begin{bmatrix} \triangle V_{polar} \\ \triangle I_{polar} \end{bmatrix} \quad (7)$$

$$R_2' = K.R_2.K^T \quad (8)$$

The novel measurement set M is consists of the following [7]:

1. The direct state output from WLS, (Equation 6)
2. PMU voltage measurements; (Equation 7)
3. PMU current measurements. (Equation 7)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:40:49 UTC from IEEE Xplore. Restrictions apply.

The new measurement model can be written as:

$$M = \begin{bmatrix} \begin{bmatrix} V_R \\ V_I \end{bmatrix}_{SE} \\ \begin{bmatrix} V_R \\ V_I \end{bmatrix}_{PM} \\ \begin{bmatrix} I_R \\ I_I \end{bmatrix}_{PM} \end{bmatrix} = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \\ H_{31} & H_{32} \\ H_{41} & H_{42} \\ H_{51} & H_{52} \\ H_{61} & H_{62} \end{bmatrix} \begin{bmatrix} V_R \\ V_I \end{bmatrix} + \begin{bmatrix} e_{VR}^{SE} \\ e_{VI}^{SE} \\ e_{VR}^{PM} \\ e_{VI}^{PM} \\ e_{IR}^{PM} \\ e_{II}^{PM} \end{bmatrix} \quad (9)$$

Equation 9 can be written as:

$$[M] = [H'].[V] + [e]$$

where the hypothetical along with real features of current and voltage measurements are reviewed in subscript R and I. The value approximated of the traditional state estimator and the phasor measurements are shown by SE and PM.

Using the succeeding Equation (9)'s WLS solution, state estimation could be resolved with no reiteration straight away:

$$V = [G']^{-1}.[H'][R']^{-1}.M \quad (10)$$

where $[G']$ is the new gain matrix given by,

$$[G'] = [H']^T[R']^{-1}[H']$$

$[R']$ is the new covariance matrix.

C. Elements of new Jacobian matrix [H']

These sub matrices of coefficient matrix that follow from structure of measurement model:

$$H_{11} = I \qquad H_{12} = 0$$

$$H_{21} = 0 \qquad H_{11} = I$$

in which, I is an $N \times N$ Identity matrix and 0 is the $N \times N$ Null matrix as well as that N is number of the buses in system. H_{31} is a $P \times N$ matrix, P denotes quantity of the PMUs [8]. Every row i resembles the PMU i as well as it is having all zeros except one at ith column equivalent with the V_{Ri}^{pmu} index in state vector.

$$H_{31i} = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0]$$
$$H_{32} = 0_{(P \times N)}$$
$$H_{41} = 0_{(P \times N)}$$

H_{31} is a $P \times N$ matrix, P denotes quantity of the PMUs. Every row, i resembles the PMU i as well as it is having all zeros but except one at jth column equivalent to V_{Ri}^{pmu} index in state vector.

$$H_{42i} = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0]$$

Now, we will grow the components of the H' equivalent to PMU current measurements. It is shown that π -model of the communication division in the Fig 3. The current I_{pq} has been presented:

$$I_{pq} = V_p Y_{po} + (V_p - V_q) Y_{pq}$$

Transforming everything in the rectangular element yields:

$$\begin{bmatrix} I_{pq\,R} \\ I_{pq\,I} \end{bmatrix} = \begin{bmatrix} (G + G_{po}) & -(B + B_{po}) & -G & B \\ (B + B_{po}) & (G + G_{po}) & -B & -G \end{bmatrix} \begin{bmatrix} V_{pR} \\ V_{pI} \\ V_{qR} \\ V_{qI} \end{bmatrix}$$

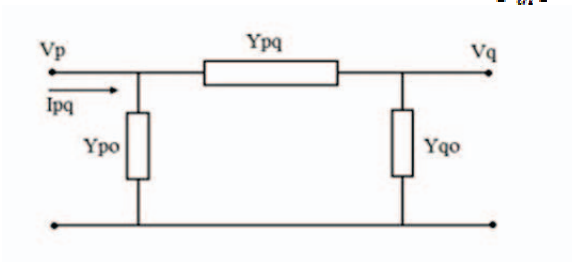


Fig 3. Π-model of the transmission branch

Covariance matrix ' R

In the relations of variations, the size along with the typical deviations on angle, the distinctions in the meter measurements have been given [9]. There is need of covariance matrix components in this model that are alike to the phasor in the elements in the rectangular shape. So, following changes are important:

$$\sigma_{|V_R|}^2 \approx \sigma_{|V|}^2.cos^2\delta + \sigma_{|\delta|}^2.|V|^2 sin^2\delta$$
$$\sigma_{|V_I|}^2 \approx \sigma_{|V|}^2.sin^2\delta + \sigma_{|\delta|}^2.|V|^2 cos^2\delta$$
$$\sigma_{|I_R|}^2 \approx \sigma_{|I|}^2.cos^2\phi + \sigma_{|\phi|}^2.|V|^2 sin^2\phi$$
$$\sigma_{|I_I|}^2 \approx \sigma_{|I|}^2.sin^2\phi + \sigma_{|\phi|}^2.|V|^2 cos^2\phi$$

In which, δ and ϕ are phasor angle having compound as well as currents. $\sigma_{|V|}$ and $\sigma_{|\delta|}$ from typical state estimator could be assessed from the gain matrix directly to meeting opinion, i.e., R_1' can be obtained from the Equation 11 as given below:

$$[R_1'] = [K'][H_1^T R_1^{-1} H_1].[K']^T \quad (11)$$

where,

$$K' = \begin{bmatrix} \cos\delta_1 & 0 & 0 & -|V_1|\sin\delta_1 & 0 & 0 \\ 0 & \cos\delta_2 & 0 & 0 & -|V_2|\sin\delta_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \sin\delta_1 & 0 & 0 & |V_1|\sin\delta_1 & 0 & 0 \\ 0 & \sin\delta_2 & 0 & 0 & |V_2|\cos\delta_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$R'_2 = \begin{bmatrix} R'_1 & 0 \\ 0 & R'_2 \end{bmatrix} \quad (12)$$

III. SIMULATION AND RESULTS

A. SIMULINK DIAGRAM OF COMPLETE SYSTEM

On the IEEE-14 and IEEE-30, a simulation has been approved on above estimation. The base's load flow taken as the initial point and considering the years of measurements

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:40:49 UTC from IEEE Xplore. Restrictions apply.

with proper errors, it is desired to be the solution of the load flow.

With the extent of the voltage, active along with the injection of the reactive power and the flow of reactive power is contained in measurements for traditional state estimation. Current Phasors along with voltages are the measurements of the Phasor. Over the system, there is a uniform spread of measurements.

The state vector's original value is dependent on the solution of the load flow for which quantities measured with standard deviation and generated randomly for addition of errors to quantities are being used. The line-currents along with the bus voltage in every line arising on the bus is desired to be inspected by each PMU.

Conforming with the tradition that the bus swing's angle is 0°, Phasor angles have been adjusted.

A. Bus System IEEE 14

The results of the simulation have been presented on the following figures.

- Fig 4 and 5 shows voltage angle estimation errors and voltage magnitude estimation error respectively, with no use of the phasor measurements (PMU) on IEEE-14 bus system.
- In Fig 4 and 5, it can be seen from that Voltage Angle Estimation error and voltage magnitude (%) error is keep on increasing which is making the system practically unstable.

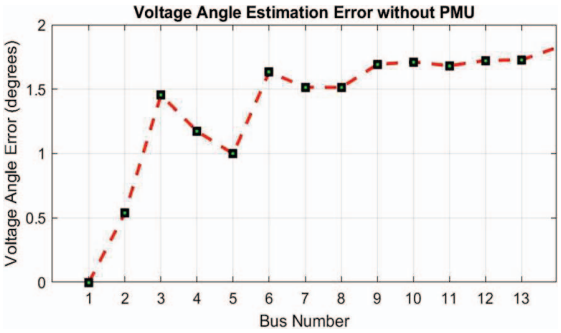


Fig 4. Voltage Angle Estimation Error without PMU for IEEE 14 Bus System

- Fig 6 and fig 7 shows voltage angle estimation errors and voltage magnitude estimation error respectively using phasor measurements (PMU) on IEEE-14 bus system.
- Fig 6 and fig 7, it can be seen from that Voltage Angle Estimation error and voltage magnitude (%) error is tending towards zero.

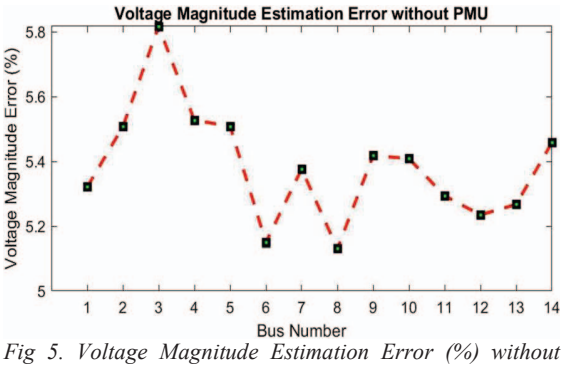


Fig 5. Voltage Magnitude Estimation Error (%) without PMU for IEEE 14 Bus System

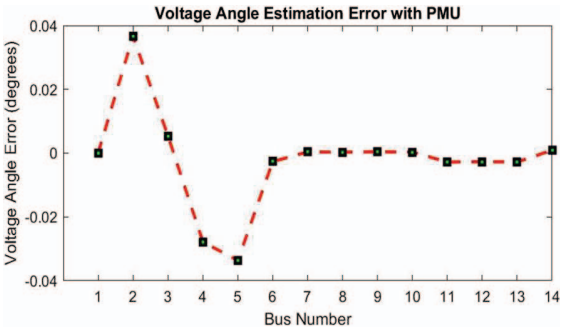


Fig 6. Voltage Angle Estimation Error with PMU for IEEE 14 Bus System

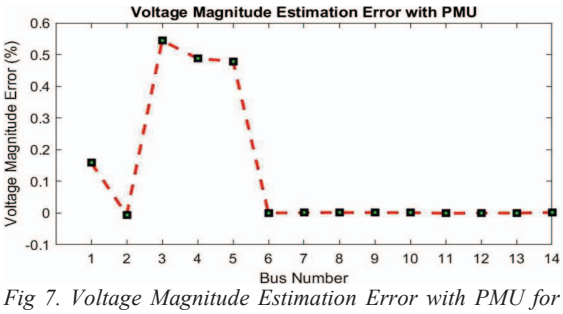


Fig 7. Voltage Magnitude Estimation Error with PMU for IEEE 14 Bus System

B. Bus System IEEE 30

- Fig 8 and 9 shows voltage angle estimation errors and voltage magnitude estimation error respectively, with no use of the phasor measurements (PMU) on IEEE-30 bus system.
- Fig 8 and fig 9, it can be seen from that Voltage Angle Estimation error and voltage magnitude (%) error is keep on increasing which is making the system practically unstable.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:40:49 UTC from IEEE Xplore. Restrictions apply.

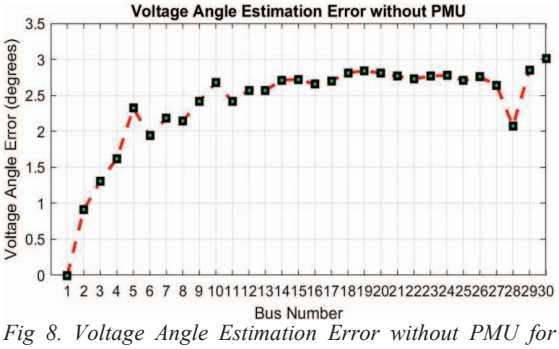


Fig 8. Voltage Angle Estimation Error without PMU for IEEE 30 Bus System

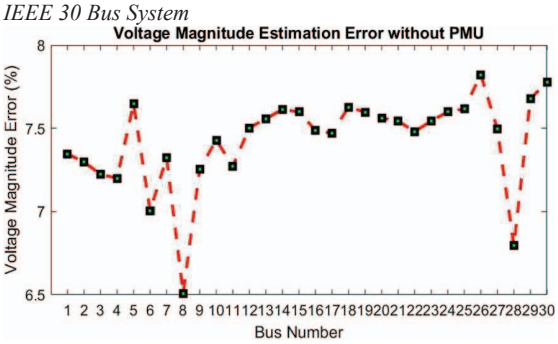


Fig 9. Voltage Magnitude Estimation Error (%) without PMU for IEEE 30 Bus System

- Fig 10 and 11 shows voltage angle estimation errors and voltage magnitude estimation error respectively, with no use of the phasor measurements (PMU) on IEEE-30 bus system.
- Fig 10 and fig 11, it can be seen from that Voltage Angle Estimation error and voltage magnitude (%) error is tending towards zero.

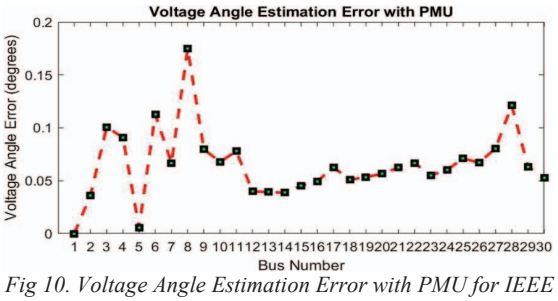


Fig 10. Voltage Angle Estimation Error with PMU for IEEE 30 Bus System

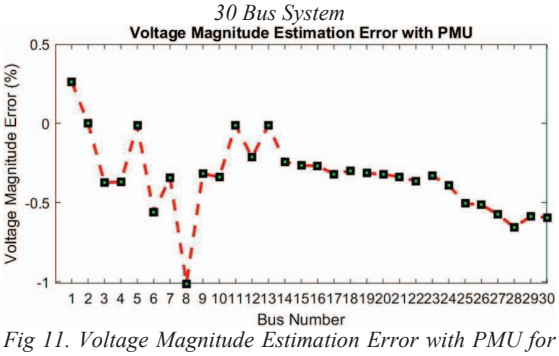


Fig 11. Voltage Magnitude Estimation Error with PMU for IEEE 30 Bus System

IV. CONCLUSION

Using the measurements of PMU there is explanation about coherent technique for state power system with WLS method. With the help of supporting theories and simulations on IEEE-30 and IEEE-14 bus systems, the results have been set up.

Comparatively, the error in PMU estimation is less or say more accurate than without it.

The working of the state estimation algorithm have efficiently increased as shown in the conclusion.

Calculation time is reduced in a step which is post-processed using phasor measurements.

And the power system and estimation accuracy are inversely related.

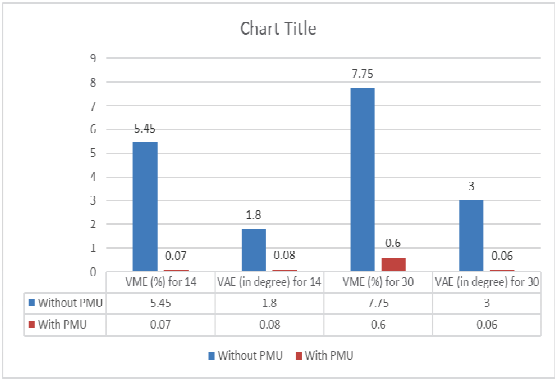


Fig 12. Graphical Representation of Errors Comparison Before and After Designing PMU

REFERENCES

[1] Jonghoek Kim, Hyun-Tae Kim; Sungyun Choi, "Performance Criterion of Phasor Measurement Units for Distribution System State Estimation", IEEE Access Volume: 7, August 2019.

[2] Yu Zhang, Ramtin Madani, Javad Lavaei, "Power system state estimation with line measurements", 2016 IEEE 55th Conference on Decision and Control (CDC).

[3] Monti, C. Muscas and F. Ponci, "Phasor Measurement Units and Wide Area Monitoring Systems," Elsevier, 2016.

[4] N. Zhou, D. Meng, Z. Huang, and G. Welch, "Dynamic state estimation of a synchronous machine using pmu data: A comparative study," IEEE Transactions on Smart Grid, vol. 6, no. 1, pp. 450-460, 2015.

[5] A. G. Phadke, "Synchronized phasor measurements in Power Systems", in Proc. IEEE Computer Applications in Power, vol. 6, no. 2, pp. 10-15, Apr. 1993.

[6] Y. C. Chen, J. Wang, A. D. Dom´ınguez-Garc´ıa, and P. W. Sauer, "Measurement- based estimation of the power flow Jacobian matrix," IEEE Transactions on Smart Grid, vol. 7, no. 5, pp. 2507-2515, 2016.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:40:49 UTC from IEEE Xplore. Restrictions apply.

[7] Hongga Zhao, "*A New State Estimation Model utilising PMU Measurements*", in International Conference on Power Technology, 2006, pp. 1-5.

[8] Fang Chen, Xueshan Han,Zhiyuan Pan,Li Han, "*State Estimation Model and Algorithm Including PMU*", Electric Utility Deregulation and Restructuring and Power Technologies, Third International Conference, April 2008, pp - 1097 – 1102.

[9] Yunzhi Cheng, Xiao Hu and Bei Gou, "*A New State Estimation Using Synchronized Phasor Measurements*", Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on 18-21 May 2008, pp. 2817 – 2820.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:40:49 UTC from IEEE Xplore. Restrictions apply.

2021 8th International Conference on Signal Processing and Integrated Networks (SPIN) | 978-1-6654-3564-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SPIN52536.2021.9566129

Studying Electromagnetic field pattern for Breast Cancer Detection by Hexagonal Patch Antenna

Shruti Awasthi
Electronics and Communication Engineering Department
Delhi Technological University
New Delhi, India
shruti798@gmail.com

Dr. Priyanka Jain
Electronics and Communication Engineering Department
Delhi Technological University
New Delhi, India
priyankajain@dtu.ac.in

***Abstract*— Cancer describes the cellular changes which lead to an uncontrolled division or growth of cells. Some types of cancer show visible growth of cells referred to as tumors. Every year hundreds of people are diagnosed with one or the other form of cancer with breast cancer as the most common type in females. This demands early detection, for which microwave imaging is considered to be the most promising method. A lot of contribution has been made in this field with materials having different permittivity and conductivity. In this paper, a 3-D structure of breast and microstrip patch antenna of hexagonal shape, operated at 2.45 GHz is designed using finite element method (FEM) in HFSS 15.0 with FR4 epoxy as a substrate to measure electromagnetic field patterns. The tumor present in breast is detected by observing Electric field patterns and specific absorption rate (SAR).**

Keywords— SAR, hexagonal patch antenna, 3-D breast structure, electric field.

I. INTRODUCTION

According to cancer statistics report 2020, no. of cancer patients in India was reported as 1,392,179 with breast cancer leading on the charts. 1 in 29 females suffer from breast cancer during their lifetime. Early diagnosis of cancer is thus, of utmost importance. There are various techniques to detect breast cancer namely, ultrasound, X-ray mammography, magnetic resonance imaging (MRI). The above mentioned techniques, suffers from various shortcomings. The X-ray mammography technique generates harmful radiation which can be risky for the patient [1]. Ultrasound is a painless technique but it provides lower resolution in detection of the tumor. In MRI, a contrast liquid is used for the identification of tumor but this technique is costly and is rarely adopted for the early detection purpose. Considering the disadvantages of all the above mentioned techniques, various developments has been made by the researches involving microwaves for the early detection of breast tumor [2-3]. The usage of microwaves provides the significant contribution in the detection of breast cancer because of its various advantages- safe, comfortable and harmless radiations. Therefore, Microwave imaging (MWI) proves to be the most promising method for the early detection of cancer. Microwave imaging method determines the differences in the dielectric constant between tumor and healthy tissues [4]. It is based on the transmission of electromagnetic waves from implanted antenna on tissue and receiving the waves reflected from the tissues by the receiving antenna [5] to determine the presence of tumor. It is an inexpensive method which is able to detect small size tumors at an early stage [6-8]. In [9], a microwave imaging based prototype for early detection of breast cancer over a broad frequency range was developed. The usage of non-ionizing radiation in this prototype [9] gave an advantage over conventional breast cancer detection methods like mammography. The microwave imaging method can be categorized into two forms- monostatic radar based microwave imaging (RBMI) and multistatic radar based microwave imaging. A monostatic RBMI was utilized for breast cancer detection [10] using cubical shaped dielectric resonator antenna. Based on this microwave imaging method, 16 ultra-wideband (UWB) antenna array was fabricated [11] which was capable of detecting 5mm and 10mm tumor inside the breast tissue. An UWB tapered slot antenna was designed in [12] for cancer detection which was compact in size with high directivity. To study the cancerous tissue, choice of antenna plays a crucial role as it helps in determining the electric and magnetic field patterns of the malignant tissue. Although various types of antennas are available such as horn antenna, dipole antenna etc. but microstrip antenna is preferred in this paper because of its low cost, small size, easy to design and availability.

In this paper, an inset fed microstrip patch antenna is designed for detecting cancerous tissue. A simple 3-D structure is used as a model to define the malignant tissue (tumor). The idea of using this 3-D model has been inspired from [13].The design is simulated on HFSS 15.0 using FEM. The antenna is designed for ISM (industrial, scientific and medical) band (2.45 GHz), considering as safe frequency band for antenna implantation on human tissues. The designed antenna is placed on breast skin to observe electromagnetic radiations for both the cases of with and without tumor in breast tissue. The SAR and electric field values help in identifying the tumor inside the tissue. The paper is sectioned into three parts with section II explains the antenna design, section III illustrates the EM (electromagnetic) values and SAR values in section IV which is then concluded with section V.

II. ANTENNA DESIGN

When electromagnetic signals illuminate the specimen, the signals gets scattered by the specimen [14] and based on this back scattering technique, presence of malignant tissue (tumor) is identified. Various factors are responsible for the strength of the back scattered waves, such as, the specimen's characteristics, environmental conditions etc.
In this paper, inset fed hexagonal patch antenna is designed for the detection of tumor in the breast tissues. Fig. 1 shows the implemented structure of antenna.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:36:55 UTC from IEEE Xplore. Restrictions apply.

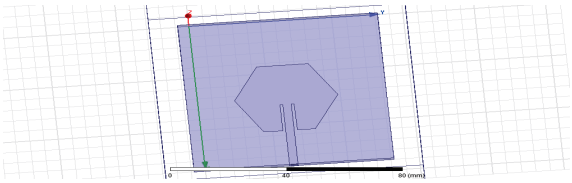


Fig.1: Hexagonal Patch Antenna

The dimension of substrate is taken as 70mm X 70mm with FR4-epoxy as substrate material (permittivity=4.4F/m) and thickness as 3.6mm. The hexagonal patch is placed at x=38mm, y=49mm, z=3.6mm location. The dimension of inset and feedline is 12.5mm X 5mm and is 34.2mm X 3mm respectively [15].3-D breast structure is considered for breast cancer detection with three layers, breast skin, fat and tumor (outwards to inwards) as shown in fig. 2.

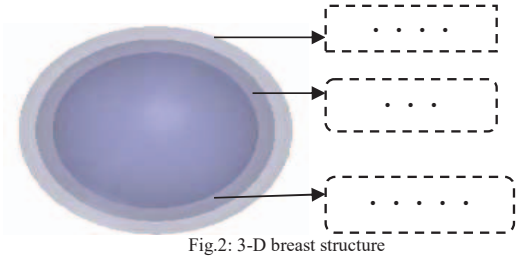


Fig.2: 3-D breast structure

The outer layer (skin) of this 3-D proposed breast structure has relative permittivity, bulk conductivity and mass density of 38, 1.49S/m and 1109 respectively. The middle layer contains fat tissue having relative permittivity as 5.14 with bulk conductivity and mass density of 0.141S/m and 911 respectively. The innermost layer is considered as tumor of radius 12mm with relative permittivity of 67, mass density and bulk conductivity as 84 and 49S/m [16] respectively. The permeability of all the layers in the 3-D breast structure is considered as 1.The designed antenna is placed on the breast skin as shown in fig.3. The whole design (3-D breast structure and antenna) is contained in radiation box to observe electromagnetic radiation pattern of the 3-D breast structure.

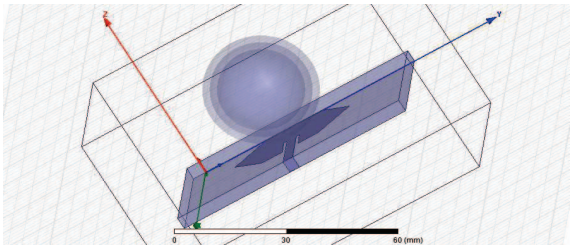


Fig.3: 3-D breast structure and proposed antenna design.

III. Simulation Results

The antenna is designed at 2.45GHz by providing frequency sweep from 1-3 GHz with step size of 0.01. On analysis, return loss is found to be -22db at 2.45 GHz as shown in fig.4

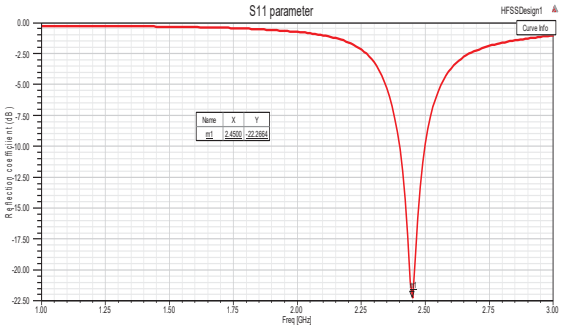


Fig. 4: Reflection coefficient of proposed antenna

The changes in electromagnetic radiation pattern by placing antenna over the breast tissue without tumor and with tumor cases are hence, observed. Firstly, the antenna is placed on breast structure (without tumor) having breast skin of radius 16 mm and fat tissue of radius 14mm and then antenna is placed on the breast tissue containing tumor. The reflection coefficient values for breast structure with and without tumor are shown in fig.5.

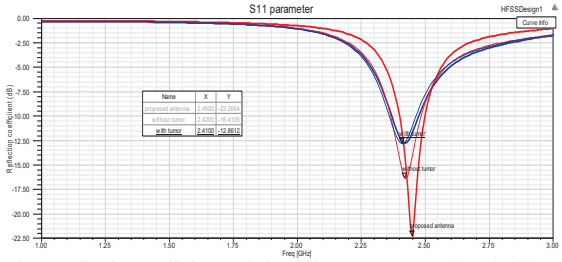


Fig.5: Reflection coefficient variation of breast structure with and without tumor.

In case of normal breast tissue, reflection coefficient value is observed as -16.4 dB whereas for tumorous tissue, it is found to be – 12.66 dB, thus, giving the difference of 23.17%. These differences can be used to identify the presence of malignant cells in the breast tissues. Moreover, a slight shift in the center frequency is observed after placing the designed antenna on 3-D breast structure. The malignant tissue can be also be distinguished from healthy tissue by observing the electric field patterns. The Far field electric field values in 3-D breast structure for both healthy tissues (without tumor) and malignant tissues (with tumor) are shown in fig. 6.

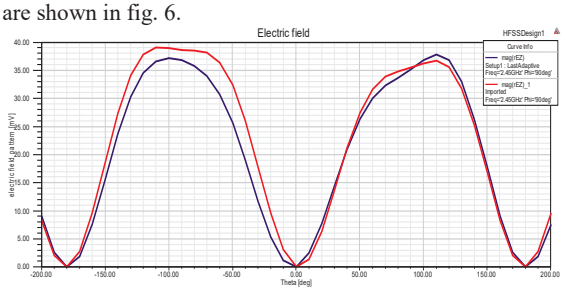


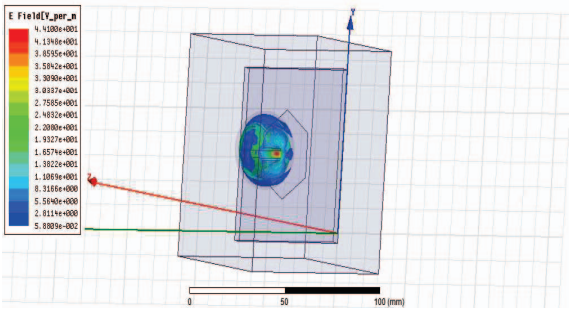
Fig.6: Graphical representation of electric field distribution in the far field region with tumor (blue line) and without tumor (red line).

The blue line is representing far field electric field distribution in malignant breast tissue (tumor) and red line is

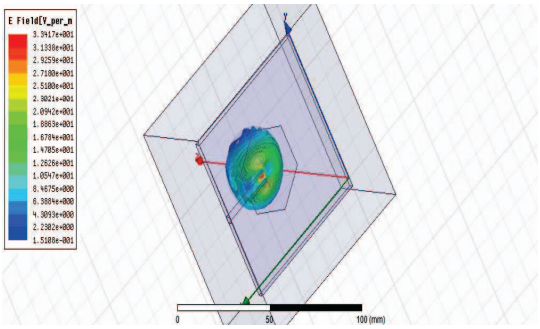
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:36:55 UTC from IEEE Xplore. Restrictions apply.

showing electric field in the far field region in the absence of tumor. The alterations in electric and magnetic field are significant parameters to distinguish cancerous tissues from healthy tissues.

The 3-D electric field radiation pattern between tumor and without tumor cases can be observed in fig.7 with Electric field strength as 44.1 V/m in presence of tumor and 33.41 V/m in the absence of tumor. The values of Electric, magnetic field and current density in case of with and without tumor's presence in the Breast tissues are tabulated in table 1.



(a)



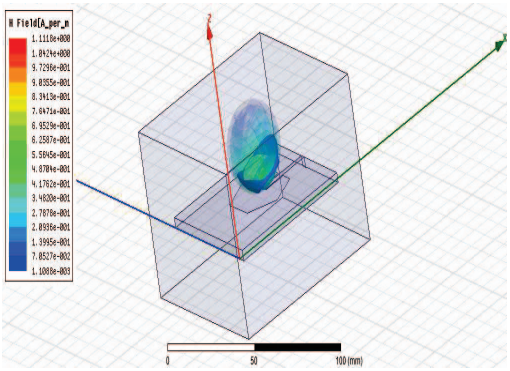
(b)

Fig.7 : 3-D plot of electric field pattern (a) with tumor (b) without tumor.

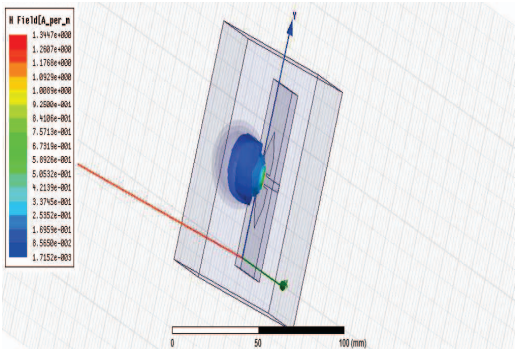
	Electric field (V/m)	Magnetic field (A/m)	Current Density (A/m^2)
Without tumor	33.41	1.11	51.45
With tumor (12mm)	44.1	1.94	65.7

Table 1: Comparison of Electric field, Magnetic field and current densities values in breast tissues containing tumor and without tumor.

The differences in magnetic field and current densities can also be considered as a distinguishing parameter between healthy (without tumor) and malignant tissues (with tumor) as shown in fig.8 and fig.9.

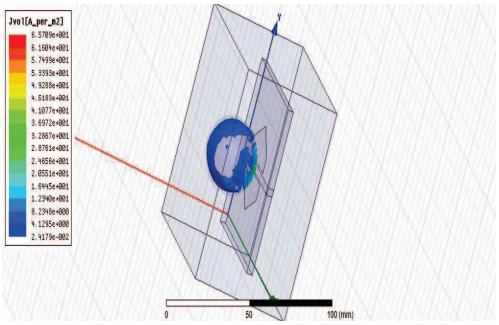


(a)

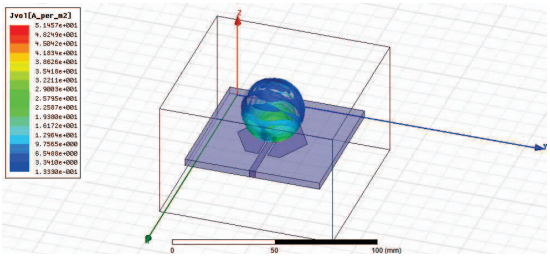


(b)

Fig.8 : Magnetic field distribution (a) without tumor and (b) with tumor.



(a)



(b)

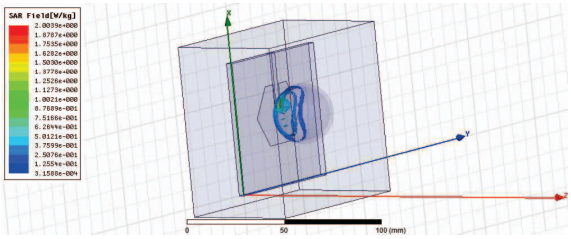
Fig.9: current density distribution (a) with tumor (b) without tumor.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:36:55 UTC from IEEE Xplore. Restrictions apply.

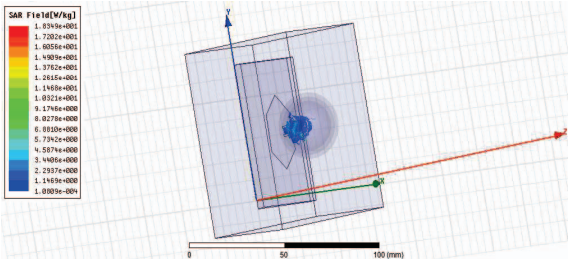
From above figures, it can be observed that there is a decrease in electric field, magnetic field and current density values in breast tissues containing tumor in comparison to without tumor condition, therefore, specific absorption rate is represented for the clear identification of tumor in the breast tissue.

IV. SPECIFIC ABSORPTION RATE

Specific absorption rate is the amount of energy absorbed by tissues, expressed in W/kg. The proposed antenna calculate the value of SAR for the purpose of detecting the tumor. The obtained values of SAR in healthy and malignant tissue are shown in fig.10.



(a)



(b)

Fig. 10: SAR values for tissues (a) without tumor (b) with tumor.

SAR value of breast tissue without tumor is found to be 2.0039W/kg whereas for breast tissue with tumor is 18.34 W/kg. This shows SAR value of tissue with tumor is highly greater than tissue without tumor. Therefore, SAR can be considered as an important parameter in detecting tumor inside the breast tissue.

V. CONCLUSION

A cost effective, compact size hexagonal antenna is designed for breast cancer detection. The antenna is designed for 2.45GHz using FR4 epoxy as substrate which provides a return loss of -22dB. A 3-D breast structure is used as a representative model for human breast structure to study the effectiveness of this proposed antenna in detecting breast cancer. Electric field, magnetic field and current density patterns are studied for both healthy (without tumor) and malignant (with tumor) tissues. Also, SAR is calculated to detect the presence of tumor inside the breast tissue. The future scope of this work is to create antenna arrays around the breast structure to enhance the functioning of microwave imaging system and image reconstruction to detect the exact location of the tumor inside the breast tissues. Also, the experimental verification on human breast tissue sample will be executed for the validation of the proposed design.

REFERENCES

[1] A. L. Siu, "Screening for breast cancer: U.S. Preventive services task force recommendation statement," Annals of Internal Medicine, vol. 164, no. 4, pp. 279–296, 2016

[2] Aldhaeebi, M.A.; Almoneef, T.S.; Attia, H.; Ramahi, O.M. Near-Field Microwave Loop Array Sensor for Breast Tumor Detection. IEEE Sens. J. 2019, 19, 11867-11872. [CrossRef]

[3] Aldhaeebi, M.A.; Almoneef, T.S.; Attia, H.; Ramahi, O.M. Electrically Small Magnetic Probe with PCA for Near-Field Microwave Breast Tumors Detection. Prog. Electromagn. Res. M 2019, 48, 177–186. [CrossRef]

[4] Wang, L., & Huang, B., "Design of Ultra-Wideband MIMO antenna for breast tumor detection", International Journal of Antennas and Propagation , 1-7,2012

[5] Mahalakshmi, N., & Jeyakumar, V., "Design and development of single layer microstrip patch antenna for breast cancer detection", Bonfring International Journal of Research in Communication Engineering, 14-18, July 2012.

[6] Lazaro, A, Villarino, R and Girbau, D, "Design of tapered slot Vivaldi antenna for UWB breast cancer detection", Microwave and Optical Technology Letters 53, 639–64, 2011,CrossRefGoogle Scholar.

[7] Adnan, S, "A compact UWB antenna design for breast cancer detection", PIER Letters 6, 129–132, 2010.Google Scholar

[8] Pozar, DM , "A micro strip antenna aperture coupled to a microstrip lin". Electronics Letters 21, 49–5, 1985.CrossRefGoogle Scholar.

[9] Janjic A, Cayoren M, Akduman I, Yilmaz T, Onemli E, Bugdayci O, Aribal ME, "SAFE: A Novel Microwave Imaging System Design for Breast Cancer Screening and Early Detection-Clinical Evaluation", Diagnostics (Basel). 2021 Mar 16;11(3):533. doi: 10.3390/diagnostics11030533. PMID: 33809770; PMCID: PMC8002280.

[10] Kaur, G, Kaur, A. ," Monostatic radar-based microwave imaging of breast tumor detection using a compact cubical dielectric resonator antenna", *Microw Opt Technol Lett*. 2021; 63: 196– 204, 2021. https://doi.org/10.1002/mop.32557

[11] Ahadi, M., Nourinia, J. & Ghobadi, C., "Square Monopole Antenna Application in Localization of Tumors in Three Dimensions by Confocal Microwave Imaging for Breast Cancer Detection: Experimental Measurement", *Wireless Pers Commun* 116, 2391–2409 2021. https://doi.org/10.1007/s11277-020-07801-5

[12] Al-Zuhairi, D.T., Gahl, J.M., Al-Azzawi, A., & Islam, N.,. "Simulation design and testing of a dielectric embedded tapered slot UWB antenna for Breast Cancer detection", *Progress in Electromagnetics Research C, 79*, 1-15, 2017

[13] Rabia Çalışkan, S. Sinan Gültekin, Dilek Uzer, Özgür Dündar, ",A Microstrip Patch Antenna Design for Breast Cancer Detection",Procedia - Social and Behavioral Sciences,Volume 195 ,Pages 2905-2911, 2015. https://doi.org/10.1016/j.sbspro.2015.06.418

[14] Park, S., & Jeyaraman, S. (2003). Enhancing the quality of life through wearable technology. IEEE Engineering in Medicine and Biology Magazine, 22, 41–48

[15] "Hexagonal Microstrip Patch Antenna at 2.45 Ghz Using HFSS Software",(sept.20,2019). Accessed:April1,2021. [Online].Available: https://www.youtube.com/watch?v=R8zkAwX8nQk.

[16] A. Kahwaji, H. Arshad, S. Sahran, A. G. Garba and R. I. Hussain, "Hexagonal microstrip antenna simulation for breast cancer detection," *2016 International Conference on Industrial Informatics and Computer Systems (CIICS)*, Sharjah, United Arab Emirates, 2016, pp. 1-4. doi: 10.1109/ICCSII.2016.7462400.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:36:55 UTC from IEEE Xplore. Restrictions apply.

Rendiconti del Circolo Matematico di Palermo Series 2
https://doi.org/10.1007/s12215-021-00677-9

The family of Bernstein–Kantorovich operators with shifted knots

Ram Pratap[1] · **Naokant Deo**[1]

Received: 25 July 2020 / Accepted: 28 October 2020
© The Author(s), under exclusive licence to Springer-Verlag Italia S.r.l., part of Springer Nature 2021

Abstract
In the present paper, we have proposed the Stancu type generalization of the family of Bernstein Kantorovich operators involving parameter $\alpha \in [0, 1]$ with Shifted Knots. These operators provide the pliability to approximate on the interval [0,1] and over its sub intervals. For the proposed operators, we investigate some basic results of approximation and their rate of convergence in terms of first and second order modulus of continuity, Lipschitz class and Lipschitz-type function. We also estimate the global rate of convergence of the operators with the help of Ditzian-Totik modulus of smoothness. Moreover, the pth order generalization of the operators is established. Some numerical simulation and graphical comparisons are given for better depiction of theoretical results.

Keywords Bernstein–Stancu operators · Korovkin theorem · Modulus of continuity · Lipschitz-type function · Ditzian-Totik modulus of smoothness

Mathematics Subject Classification 41A25 · 41A36 · 41A10

1 Introduction

Bernstein operators in 1912 were constructed by Bernstein [6] to give a more subtle proof of Weierstrass Approximation Theorem [29]. For a continuous function f(x) on [0,1], the Bernstein operators are as follows:

$$B_n(f;x) = \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} f\left(\frac{k}{n}\right). \tag{1.1}$$

✉ Ram Pratap
rampratapiitr@gmail.com

Naokant Deo
naokantdeo@dce.ac.in; dr_naokant_deo@yahoo.com

[1] Department of Applied Mathematics, Formerly Delhi College of Engineering, Delhi Technological University, Bawana Road, Delhi 110042, India

Chen et al. [7] constructed a new generalization of classical Bernstein operators with the help of a non negative real parameter $\alpha \in [0,1]$. Using these operators, another proof to Weierstrass Approximation Theorem [29] was provided, and the operators are as follows:

$$B_{n,\alpha}(f;x) = \sum_{k=0}^{n} p_{n,k}^{(\alpha)} f\left(\frac{k}{n}\right), \tag{1.2}$$

where $p_{1,0}^{(\alpha)}(x) = 1 - x$, $p_{1,1}^{(\alpha)}(x) = x$ and for $n \geq 2$

$$p_{n,k}^{(\alpha)}(x) = \left[\binom{n-2}{k}(1-\alpha)x + \binom{n-2}{k-2}(1-\alpha)(1-x) + \binom{n}{k}\alpha x\right] x^{k-1}(1-x)^{n-k-1}$$

Since Bernstein operators are not suitable in approximating Lebesgue integrable functions, therfore for such functions Kantorovich [20] modified the Bernstein operators for $f \in C[0,1]$ as follows:

$$K_n(f;x) = (n+1)\sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} \int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} f(t)dt. \tag{1.3}$$

Many authors have discussed Kantorovich modification of various linear positive operators [1, 2, 5, 8, 11, 12, 15, 22, 25] and studied several approximation results.

Recently, Mohiuddine et al. [22] considered the Kantorovich form of operators (1.2) given by:

$$K_{n,\alpha}(f;x) = (n+1)\sum_{k=0}^{n} p_{n,k}^{(\alpha)} \int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} f(t)dt. \tag{1.4}$$

Recently, Gadjiev and Ghorbanalizadeh [14] proposed another generalization of Bernstein-Stancu type polynomials with shifted knots. These shifted knots provide pliability to the operators and also allow us to approximate on the interval [0,1] and over its sub intervals. These operators are defined as follows:

$$B_n^{(r,s)}(f;x) = \left(\frac{n+s_2}{n}\right)^n \sum_{k=0}^{n} b_n^{(r_2,s_2)}(x) f\left(\frac{k+r_1}{n+s_1}\right), \tag{1.5}$$

where

$$b_n^{(r_2,s_2)}(x) = \binom{n}{k}\left(x - \frac{r_2}{n+s_2}\right)^k \left(\frac{n+r_2}{n+s_2} - x\right)^{n-k}$$

and $\frac{r_2}{n+s_2} \leq x \leq \frac{n+r_2}{n+s_2}$, $r_j, s_j (j = 1,2)$ are non-negative real numbers provided $0 \leq r_1 \leq r_2 \leq s_1 \leq s_2$. For $r_2 = s_2 = 0$, we obtain Bernstein-Stancu operators and for $r_1 = r_2 = s_1 = s_2 = 0$, we get classical Bernstein operators. Içöz and Kantorovich [16] considered the Kantorovich-type modification of the operators (1.5) and discussed its rate of convergence using usual modulus of continuity. Recently, Rahman et al. [27] defined the λ

-Bernstein-Kantorovich operators with shifted knots and discussed some basic results along with its rate of convergence, graphical comparisons and error estimation tables. For more articles, we refer to readers [4, 16–18, 23, 28].

Motivated from the above stated work, we consider the family of Bernstein-Stancu-Kantorovich operators with shifted knots. For $f \in C[0,1]$, the operators are defined as follows:

$$K_{n,\alpha}^{(r,s)}(f;x) = \left(\frac{n+s}{n}\right)^n (n+s+1) \sum_{k=0}^{n} p_{n,k,\alpha}^{(r,s)}(x) \int_{\frac{k+r}{n+s+1}}^{\frac{k+r+1}{n+s+1}} f(t)dt, \tag{1.6}$$

where

$$p_{n,k,\alpha}^{(r,s)}(x) = \left[\binom{n-2}{k-1}(1-\alpha)\left(x-\frac{r}{n+s}\right)+\binom{n-2}{k-2}(1-\alpha)\left(\frac{n+r}{n+s}-x\right)\right.$$
$$\left.+\binom{n}{k}\alpha\left(x-\frac{r}{n+s}\right)\left(\frac{n+r}{n+s}-x\right)\right]\left(x-\frac{r}{n+s}\right)^{k-1}\left(\frac{n+r}{n+s}-x\right)^{n-k-1},$$

$x \in \left[\frac{r}{n+s}, \frac{n+r}{n+s}\right]$ and r, s are non-negative real numbers such that $0 \le r \le s$. In special cases:

Case 1. For $r = s = 0$, the operators (1.6) reduce to family of Bernstein Kantorovich operators (1.4).

Case 2. For $r = s = 0$ and $\alpha = 1$ the operators (1.6) reduce to classical Bernstein-Kantorovich operators (1.3).

The aim of this paper is to define the family of Bernstein–Stancu–Kantorovich operators with shifted knots and estimate its moments and central moments in order to study local and global rate of convergence and discuss the pth order generalization of operators (1.6). In support of theoretical results, we have also shown numerical and graphical comparisons for the proposed operators.

2 Auxiliary results

Lemma 2.1 *The moments of the operators* (1.6) *are given as:*

(i) $K_{n,\alpha}^{(r,s)}(1;x) = 1$;

(ii) $K_{n,\alpha}^{(r,s)}(t;x) = \frac{1}{n(n+s+1)}[(n+s)x+1)(n+(1-\alpha)s]$;

(iii)
$$K_{n,\alpha}^{(r,s)}(t^2;x) = \frac{1}{3n^3(n+s+1)^2}\left[(3n^5+3((3-\alpha)s+(3\alpha-4))n^4+3((3-2\alpha)s^2\right.$$
$$+(10\alpha-12)s+4(1-\alpha))n^3+3((1-\alpha)s^3+(12-11\alpha)s^2+12(1-\alpha)s)n^2$$
$$+12(1-\alpha)(s-3)s^2n-12(1-\alpha)s^3)x^2+((12-6\alpha)n^4+(21-15\alpha$$
$$-6(1-\alpha)r)s+6(3-2\alpha)r-3(1-\alpha))n^3+((9-6r)(1-\alpha)s^2-3((1-\alpha)$$
$$-2(7-6\alpha)r)s-24(1-\alpha)r)n^2+24(1-\alpha)(s-2)srn-24(1-\alpha)rs^2)x$$
$$\left.+(4-3\alpha-6r)n^3+(3(1-\alpha)(3-r)rs+3r((1-\alpha)-(2-\alpha)r))n^2\right.$$
$$\left.-12(1-\alpha)s(s-2)rn+12(1-\alpha)r^2s\right].$$

Lemma 2.2 *The central moments* $\mu_{n,\alpha,j}^{(r,s)}(x) = K_{n,\alpha}^{(r,s)}((t-x)^j;x)$ *for* $j = 1, 2$ *of the operators* (1.6) *are given as:*

(i) $\mu_{n,\alpha,1}^{(r,s)}(x) = \frac{1}{n(n+s+1)}[(((1-\alpha)s-1)n + (1-\alpha)s^2)x + (n+(1-\alpha)s)]$

$\mu_{n,\alpha,2}^{(r,s)}(x) = \frac{1}{3n^3(n+s+1)^2}[(-3((1-\alpha)s + (4-3\alpha))n^4 - 6((1-\alpha)s^2$
$+ (7-6\alpha)s)n^3 - 3((1-\alpha)s^3 + (14-13\alpha)s^2 + 4(1-\alpha)s)n^2$
$- 12(1-\alpha)(s-3)s^2n + 12(1-\alpha)s^3)x^2 + (6(1-\alpha)n^4$
$+ 3((1-\alpha)(3-2r)s + 2(3-2\alpha)r - (3-\alpha))n^3$
$+ (3(1-\alpha)(1-2r)s^2 + (-9(1-\alpha) + 6(7-6\alpha)r)s$
$- 24(1-\alpha)r)n^2 + 24(1-\alpha)(s-2)rn - 24(1-\alpha)rs^2)x$
$+ (4-3\alpha-6r)n^3 + 3((1-\alpha)(r-3)rs + (r-2)r^2$
$+ (1-\alpha)r)n^2 - 12(1-\alpha)(s-1)r^2n + 12(1-\alpha)r^2s]$

(ii) .

Lemma 2.3 *For* $\alpha \in [0,1]$, $f \in C[0,1]$ and $x \in \left[\frac{r}{n+s}, \frac{n+r}{n+s}\right]$,*we have*

$$\left|K_{n,\alpha}^{(r,s)}(f;x)\right| \leq \|f\|.$$

Proof From (1.6) and Lemma 2.1, we have

$$\left|K_{n,\alpha}^{(r,s)}(f;x)\right| \leq \left(\frac{n+s}{n}\right)^n (n+s+1) \sum_{k=0}^{n} p_{n,k,\alpha}^{(r,s)}(x) \int_{\frac{k+r}{n+s+1}}^{\frac{k+r+1}{n+s+1}} |f|dt$$

$$\leq \|f\| \left(\frac{n+s}{n}\right)^n (n+s+1) \sum_{k=0}^{n} p_{n,k,\alpha}^{(r,s)}(x) \int_{\frac{k+r}{n+s+1}}^{\frac{k+r+1}{n+s+1}} 1dt$$

$$\leq \|f\|.$$

□

3 Main results

Let $C[0,1]$ be the space of all continuous functions on $[0,1]$ with sup-norm $\|f\| := \sup_{x\in[0,1]} |f(x)|$. Let $f \in C[0,1]$ and $\delta > 0$. Then the modulus of continuity $\omega(f,\delta)$ is given as:

$$\omega(f,\delta) = \sup_{\substack{|v-w| \leq \delta \\ v,w \in [0,1]}} |f(v) - f(w)|.$$

It is well-known $\lim_{\delta\to 0} \omega(f;\delta) = 0$. For $f \in C[0,1]$ and $x, t \in [0,1]$, we have

$$|f(t) - f(x)| \leq \omega(f;\delta)\left(1 + \frac{|t-x|}{\delta}\right), \tag{3.1}$$

and Peetre K-functional [10] is given by

$$K_2(f;\delta) = \inf_{g\in C^2}\{|f-g| + \delta\|g''\|\},$$

where $\delta > 0$ and $C^2 = \{g \in C[0,1] : g', g'' \in C[0,1]\}$. In [10], there exists an positive constant $\lambda > 0$, such that

$$K_2(f;\delta) \le \lambda\omega_2(f;\sqrt{\delta}). \tag{3.2}$$

The second order modulus of continuity $\omega_2(.;\delta)$ for $f \in C[0,1]$ is as follows:

$$\omega_2(f;\delta) = \sup_{0<h\le\delta}\ \sup_{x,x+h,x+2h\in[0,1]} |f(x+2h) - 2f(x+h) + f(x)|$$

Theorem 3.1 *For $f \in C[0,1]$ and $\alpha \in [0,1]$,we have*

$$\lim_{n\to\infty} \max_{\frac{r}{n+s}\le x\frac{n+r}{n+s}} \left|K_{n,\alpha}^{(r,s)}(f;x) - f(x)\right| = 0.$$

Proof From Lemma 2.1 and for $k = 0, 1, 2$, we have

$$\lim_{n\to\infty} \max_{\frac{r}{n+s}\le x\le\frac{n+r}{n+s}} \left|K_{n,\alpha}^{(r,s)}(t^k;x) - x^k\right| = 0 \tag{3.3}$$

Now consider a sequence of operators:

$$K_{n,\alpha}^{(r,s)}(f(t);x) = \begin{cases} K_{n,\alpha}^{(r,s)}(f(t);x) \text{ if } \frac{r}{n+s} \le x \le \frac{n+r}{n+s}, \\ f(x) \qquad\qquad \text{if } x \in \left[0, \frac{r}{n+s}\right] \bigcup \left[\frac{n+r}{n+s}, 1\right]. \end{cases}$$

Then,

$$\left|K_{n,\alpha}^{(r,s)}(f(t);x) - f(x)\right| = \max_{\frac{r}{n+s}\le x\le\frac{n+r}{n+s}} \left|K_{n,\alpha}^{(r,s)}(f(t);x) - f(x)\right|. \tag{3.4}$$

Using (3.3) and (3.4), we get

$$\lim_{n\to\infty} \left|K_{n,\alpha}^{(r,s)}(t^k;x) - x^k\right| = 0, \ \ k = 0, 1, 2.$$

Using well-known Korovkin theorem ([3, 21]) and in view of (3.4), we obtain the required result. □

Theorem 3.2 *For $f \in C[0,1]$, $\alpha \in [0,1]$ and $x \in \left[\frac{r}{n+s}, \frac{n+r}{n+s}\right]$,* we have

$$\left|K_{n,\alpha}^{(r,s)}(f;x) - f(x)\right| \le C\omega_2\left(f; \frac{\mu_{n,\alpha,2}^{(r,s)}(x)}{4}\right) + \omega\left(f; \left|\mu_{n,\alpha,1}^{(r,s)}(x)\right|\right),$$

where $\mu_{n,\alpha,1}^{(r,s)}(x)$ and $\mu_{n,\alpha,2}^{(r,s)}(x)$ are given in Lemma 2.2.

Proof Define an auxiliary operators as follows:

$$\widehat{K_{n,\alpha}^{(r,s)}}(f;x) = K_{n,\alpha}^{(r,s)}(f;x) - f\left(x + \mu_{n,\alpha,1}^{(r,s)}(x)\right) + f(x). \tag{3.5}$$

With the help of Lemma 2.1, we obtain

$$\widehat{K_{n,\alpha}^{(r,s)}}(1;x) = 1, \quad \widehat{K_{n,\alpha}^{(r,s)}}(t;x) = x.$$

For $g \in C^2$, by Taylor's theorem, we have

$$g(t) = g(x) + (t-x)g'(x) + \int_x^t (t-v)g''(v)dv.$$

Applying $\widehat{K_{n,\alpha}^{(r,s)}}(.;x)$ on both sides of the above equation, we obtain

$$\begin{aligned}
\widehat{K_{n,\alpha}^{(r,s)}}(g(t);x) - g(x) =& \widehat{K_{n,\alpha}^{(r,s)}}((t-x);x)g'(x) + \widehat{K_{n,\alpha}^{(r,s)}}\left(\int_x^t (t-v)g''(v)dv;x\right) \\
=& \widehat{K_{n,\alpha}^{(r,s)}}\left(\int_x^t (t-v)g''(v)dv;x\right) \\
& - \widehat{K_{n,\alpha}^{(r,s)}}\left(\int_x^{x+\mu_{n,\alpha,1}^{(r,s)}(x)} \left(x + \mu_{n,\alpha,1}^{(r,s)}(x) - v\right)g''(v)dv;x\right).
\end{aligned}$$

$$\begin{aligned}
\left|\widehat{K_{n,\alpha}^{(r,s)}}(g(t);x) - g(x)\right| \leq& \widehat{K_{n,\alpha}^{(r,s)}}\left(\left|\int_x^t (t-v)g''(v)dv\right|;x\right) \\
& + \widehat{K_{n,\alpha}^{(r,s)}}\left(\int_x^{x+\mu_{n,\alpha,1}^{(r,s)}(x)} \left|\left(x + \mu_{n,\alpha,1}^{(r,s)}(x) - v\right)\right||g''(v)|dv;x\right) \\
\leq& K_{n,\alpha}^{(r,s)}((t-x)^2;x)\|g''\| + \left(\mu_{n,\alpha,1}^{(r,s)}(x)\right)^2\|g''\| \\
\leq& \left(\mu_{n,\alpha,2}^{(r,s)}(x) + \left(\mu_{n,\alpha,1}^{(r,s)}(x)\right)^2\right)\|g''\|.
\end{aligned} \tag{3.6}$$

From (3.5), we have

$$\left|\widehat{K_{n,\alpha}^{(r,s)}}(f(t);x)\right| \leq 3\|f\|. \tag{3.7}$$

Since

$$\begin{aligned}\left|K_{n,\alpha}^{(r,s)}(f;x)-f(x)\right| &= \left|\widehat{K_{n,\alpha}^{(r,s)}}(f(t);x)-f(x)+f\left(x+\mu_{n,\alpha,1}^{(r,s)}(x)\right)-f(x)\right| \\ &\leq \left|\widehat{K_{n,\alpha}^{(r,s)}}(f-g;x)\right| + \left|\widehat{K_{n,\alpha}^{(r,s)}}(g;x)-g(x)\right| \\ &\quad + |f(x)-g(x)| + \left|f\left(x+\mu_{n,\alpha,1}^{(r,s)}(x)\right)-f(x)\right|.\end{aligned}$$

From (3.6) and (3.7), we have

$$\begin{aligned}\left|K_{n,\alpha}^{(r,s)}(f;x)-f(x)\right| \leq & 4\|f-g\| + \left(\mu_{n,\alpha,2}^{(r,s)}(x)+\left(\mu_{n,\alpha,1}^{(r,s)}(x)\right)^2\right)\|g''\| \\ & + \omega\left(f;\left|\mu_{n,\alpha,1}^{(r,s)}(x)\right|\right).\end{aligned}$$

Taking infimum over $f \in C^2$, we have

$$\left|K_{n,\alpha}^{(r,s)}(f;x)-f(x)\right| \leq 4K_2\left(f;\frac{\mu_{n,\alpha,2}^{(r,s)}(x)}{4}\right)+\omega\left(f;\left|\mu_{n,\alpha,1}^{(r,s)}(x)\right|\right).$$

In view of relation (3.2), we obtain the required result. □

For $M > 0$ and $0 < \xi \leq 1$. The Lipschitz class $Lip_M(\xi)$ is defined as:

$$Lip_M(\xi) = \left\{f : |f(t)-f(x)| \leq M|t-x|^{\xi} \text{ and } x \in [0,1]\right\} \tag{3.8}$$

Theorem 3.3 *For $f \in Lip_M(\xi)$, $\alpha \in [0,1]$ and for each $x \in \left[\frac{r}{n+s}, \frac{n+r}{n+s}\right]$, we have*

$$\left|K_{n,\alpha}^{(r,s)}(f;x)-f(x)\right| \leq M\left(\mu_{n,\alpha,2}^{(r,s)}(x)\right)^{\frac{\xi}{2}},$$

where $\mu_{n,\alpha,2}^{(r,s)}(x)$ is given in Lemma 2.2.

Proof By linearity and monotonicity of the operators $K_{n,\alpha}^{(r,s)}$, we have

$$\begin{aligned}\left|K_{n,\alpha}^{(r,s)}(f;x)-f(x)\right| = \left|K_{n,\alpha}^{(r,s)}(f(t)-f(x);x)\right| &\leq K_{n,\alpha}^{(r,s)}(|f(t)-f(x)|;x) \\ &\leq MK_{n,\alpha}^{(r,s)}\left(|t-x|^{\xi};x\right).\end{aligned}$$

Using Hölder's inequality with $p = \frac{\xi}{2}$, $q = \frac{2-\xi}{2}$, we have

$$\begin{aligned}\left|K_{n,\alpha}^{(r,s)}(f;x)-f(x)\right| &\leq M\left(K_{n,\alpha}^{(r,s)}((t-x)^2;x)\right)^{\frac{\xi}{2}}\left(K_{n,\alpha}^{(r,s)}(1;x)\right)^{\frac{2-\xi}{2}} \\ &\leq M\left(\mu_{n,\alpha,2}^{(r,s)}(x)\right)^{\frac{\xi}{2}}.\end{aligned}$$

Hence the proof. □

In next theorem, we discuss rate of convergence for the operators (1.6) using Lipschitz-type function based on two parameters $a \geq 0$ and $b \geq 0$. The Lipschitz-type function $Lip_M^{(a,b)}(\xi)$ is given as:

$$Lip_M^{(a,b)}(\xi) = \left\{ f \in C[0,1] : |f(t) - f(x)| \leq M \frac{|t-x|^{\xi}}{(ax^2+bx+t)^{\frac{\xi}{2}}}; x \in (0,1], t \in [0,1] \right\},$$

for $\xi \in (0,1]$ and M is positive constant (see [24]).

Theorem 3.4 *For $f \in Lip_M^{(a,b)}(\xi)$, we have*

$$\left| K_{n,\alpha}^{(r,s)}(f;x) - f(x) \right| \leq M \left(\frac{\mu_{n,\alpha,2}^{(r,s)}(x)}{ax^2+bx} \right)^{\frac{\xi}{2}},$$

where $\mu_{n,\alpha,2}^{(r,s)}(x)$ is given in Lemma 2.2.

Proof First, we show the statement is true for $\xi = 1$.

We can write

$$\begin{aligned}
&\left| K_{n,\alpha}^{(r,s)}(f;x) - f(x) \right| \\
&\leq K_{n,\alpha}^{(r,s)}(|f(t) - f(x)|;x) \\
&\leq \left(\frac{n+s}{n} \right)^n (n+s+1) \sum_{k=0}^{n} p_{n,k,\alpha}^{(r,s)}(x) \int_{\frac{k+r}{n+s+1}}^{\frac{k+r+1}{n+s+1}} |f(t) - f(x)| dt \\
&\leq M \left(\frac{n+s}{n} \right)^n (n+s+1) \sum_{k=0}^{n} p_{n,k,\alpha}^{(r,s)}(x) \int_{\frac{k+r}{n+s+1}}^{\frac{k+r+1}{n+s+1}} \frac{|t-x|}{(ax^2+bx+t)^{\frac{1}{2}}} dt.
\end{aligned}$$

Since $(ax^2+bx+t)^{-\frac{1}{2}} \leq (ax^2+bx)^{-\frac{1}{2}}$, we have

$$\left| K_{n,\alpha}^{(r,s)}(f;x) - f(x) \right| \leq \frac{M}{(ax^2+bx)^{\frac{1}{2}}} \left(\frac{n+s}{n} \right)^n (n+s+1) \sum_{k=0}^{n} p_{n,k,\alpha}^{(r,s)}(x) \int_{\frac{k+r}{n+s+1}}^{\frac{k+r+1}{n+s+1}} |t-x| dt.$$

Applying Cauchy–Schwarz inequality in above, we obtain

$$\left| K_{n,\alpha}^{(r,s)}(f;x) - f(x) \right| \leq M \sqrt{\frac{\mu_{n,\alpha,2}^{(r,s)}(x)}{ax^2+bx}}.$$

Hence the statement is true for $\xi = 1$. Next, we prove the same for $0 < \xi < 1$. Using *Hölder*'s inequality with $p = \frac{2}{\xi}$ and $q = \frac{2}{2-\xi}$, we obtain

$$\left|K_{n,\alpha}^{(r,s)}(f;x)-f(x)\right| \leq \left(\frac{n+s}{n}\right)^n (n+s+1)\sum_{k=0}^{n} p_{n,k,\alpha}^{(r,s)}(x) \int_{\frac{k+r}{n+s+1}}^{\frac{k+r+1}{n+s+1}} |f(t)-f(x)|dt$$

$$\leq \frac{M}{(ax^2+bx)^{\frac{\xi}{2}}}\left(\frac{n+s}{n}\right)^n (n+s+1)\sum_{k=0}^{n} p_{n,k,\alpha}^{(r,s)}(x) \int_{\frac{k+r}{n+s+1}}^{\frac{k+r+1}{n+s+1}} |t-x|^{\xi}dt$$

$$\leq \frac{M\left(K_{n,\alpha}^{(r,s)}((t-x)^2;x)\right)^{\frac{\xi}{2}}}{(ax^2+bx)^{\frac{\xi}{2}}} = M\left(\frac{\mu_{n,\alpha,2}^{(r,s)}(x)}{ax^2+bx}\right)^{\frac{\xi}{2}}.$$

which is clearly true for $0<\xi<1$. The proof is complete. □

The first and second order Ditzian–Totik modulus of continuity are defined as:

$$\omega_{\eta}(f;\delta) = \sup_{0<|h|\leq\delta}\ \sup_{x,x\pm h\eta(x)\in[0,1]} \{|f(x+h\eta(x))-f(x)|\}$$

and

$$\omega_2^{\phi}(f;x) = \sup_{0<|h|\leq\delta}\ \sup_{x,x\pm h\phi(x)\in[0,1]} \{|f(x+h\phi(x))-2f(x)+f(x-h\phi(x))|\},$$

where $\phi(x)=[(x-a)(b-x)]^2$ is known as admissible step-weight function if $x\in[a,b]$ (see [10]).

The Peetre K-functional is given by

$$K_2^{\phi}(f;\delta) = \inf_{g\in W_{\phi}^2}\left\{\|f-g\|_{C[0,1]}+\delta\left\|\phi^2 g''\right\|_{C[0,1]};g\in C^2[0,1]\right\},\ (\delta>0),$$

where

$$W_{\phi}^2 = \left\{g\in C[0,1]:\ g'\in AC[0,1],\phi^2 g''\in C[0,1]\right\}$$

and

$$C^2[0,1] = \left\{g\in C[0,1]:\ g',g''\in C[0,1]\right\}.$$

Here $g'\in AC[0,1]$ means that g' is absolutely continuous on [0, 1]. In [9], there exists an absolute constant $C>0$, such that

$$C^{-1}\omega_2^{\phi}\left(f;\sqrt{\delta}\right)\leq K_2^{\phi}(f;\delta)\leq C\omega_2^{\phi}\left(f;\sqrt{\delta}\right). \tag{3.9}$$

Theorem 3.5 *Let $\alpha\in[0,1]$ and $f\in C[0,1]$ such that ϕ^2 ($\phi\neq 0$) is concave, then*

$$\left|\widehat{K_{n,\alpha}^{(r,s)}}(f;x)-f(x)\right| \leq C\omega_2^{\phi}\left(f;\frac{\left(\mu_{n,\alpha,2}^{(r,s)}(x)+\left(\mu_{n,\alpha,1}^{(r,s)}(x)\right)^2\right)^{\frac{1}{2}}}{2\phi(x)}\right)+\omega_{\eta}\left(f;\frac{\mu_{n,\alpha,1}^{(r,s)}(x)}{\eta(x)}\right),$$

where $\mu_{n,\alpha,1}^{(r,s)}(x)$ *and* $\mu_{n,\alpha,2}^{(r,s)}(x)$ *are given in Lemma 2.2.*

Proof From (3.5), it is clear that $\widehat{K_{n,\alpha}^{(r,s)}}(1;x)=1$, $\widehat{K_{n,\alpha}^{(r,s)}}(t;x)=x$ and $\widehat{K_{n,\alpha}^{(r,s)}}(t-x;x)=0$. Let $y=\lambda x+(1-\lambda)t$, $\lambda\in[0,1]$. Since ϕ^2 $(\phi\neq 0)$ is concave on $[0,\ 1]$, we have $\phi^2(y)\geq\lambda\phi^2(x)+(1-\lambda)\phi^2(t)$ and hence

$$\frac{|t-y|}{\phi^2(y)}\leq\frac{\lambda|x-t|}{\lambda\phi^2(x)+(1-\lambda)\phi^2(t)}\leq\frac{|t-x|}{\phi^2(x)}. \tag{3.10}$$

We can write

$$\begin{aligned}\left|\widehat{K_{n,\alpha}^{(r,s)}}(f;x)-f(x)\right| \leq&\left|\widehat{K_{n,\alpha}^{(r,s)}}(f-g;x)\right|+\left|\widehat{K_{n,\alpha}^{(r,s)}}(g;x)-g(x)\right|+|f(x)-g(x)|\\ \leq&4\|f-g\|+\left|\widehat{K_{n,\alpha}^{(r,s)}}(g;x)-g(x)\right|.\end{aligned} \tag{3.11}$$

By Taylor's formula and operating $\widehat{K_{n,\alpha}^{(r,s)}}(.;x)$, we have

$$\begin{aligned}&\left|\widehat{K_{n,\alpha}^{(r,s)}}(g;x)-g(x)\right|\\ &\leq K_{n,\alpha}^{(r,s)}\left(\int_x^t|t-y|\left|g''(y)\right|dy;x\right)+\left|\int_x^{x+\mu_{n,1,\alpha}^{(r,s)}(x)}\left|x+\mu_{n,\alpha,1}^{(r,s)}(x)-y\right|\left|g''(y)\right|dy\right|\\ &\leq\left\|\phi^2g''\right\|_{C[0,1]}K_{n,\alpha}^{(r,s)}\left(\int_x^t\frac{|t-y|}{\phi^2(y)}dy;x\right)+\left\|\phi^2g''\right\|_{C[0,1]}\left|\int_x^{x+\mu_{n,\alpha,1}^{(r,s)}(x)}\frac{\left|x+\mu_{n,\alpha,1}^{(r,s)}(x)-y\right|}{\phi^2(y)}dy\right|\\ &\leq\phi^{-2}(x)\left\|\phi^2g''\right\|_{C[0,1]}\left(\mu_{n,\alpha,2}^{(r,s)}(x)+\left(\mu_{n,\alpha,1}^{(r,s)}(x)\right)^2\right).\end{aligned} \tag{3.12}$$

From (3.10), (3.11) and using the definition of K-functional, we obtain

$$\begin{aligned}\left|\widehat{K_{n,\alpha}^{(r,s)}}(f;x)-f(x)\right| \leq&\phi^{-2}(x)\left\|\phi^2g''\right\|_{C[0,1]}\left(\mu_{n,\alpha,2}^{(r,s)}(x)+\left(\mu_{n,\alpha,1}^{(r,s)}(x)\right)^2\right)+4\|f-g\|_{C[0,1]}\\ \leq&C\omega_2^{\phi}\left(f;\frac{\left(\mu_{n,\alpha,2}^{(r,s)}(x)+\left(\mu_{n,\alpha,1}^{(r,s)}(x)\right)^2\right)^{\frac{1}{2}}}{2\phi(x)}\right).\end{aligned} \tag{3.13}$$

By using first order Ditzian–Totik modulus of continuity, we have

$$\left|f\left(x+\mu_{n,\alpha,1}^{(r,s)}(x)\right)-f(x)\right| = \left|f\left(x+\eta(x)\frac{\mu_{n,\alpha,1}^{(r,s)}(x)}{\eta(x)}\right)-f(x)\right| \leq \omega_\eta\left(f;\frac{\mu_{n,\alpha,1}^{(r,s)}(x)}{\eta(x)}\right)$$

Therefore the following inequality satisfy:

$$\left|K_{n,\alpha}^{(r,s)}(f;x)-f(x)\right| \leq \left|\widehat{K_{n,\alpha}^{(r,s)}}(f;x)-f(x)\right| + \left|f\left(x+\mu_{n,\alpha,1}^{(r,s)}(x)\right)-f(x)\right| \leq C\omega_2^{\phi}\left(f;\frac{\left(\mu_{n,\alpha,2}^{(r,s)}(x)+\left(\mu_{n,\alpha,1}^{(r,s)}(x)\right)^2\right)^{\frac{1}{2}}}{2\phi(x)}\right)+\omega_\eta\left(f;\frac{\mu_{n,\alpha,1}^{(r,s)}(x)}{\eta(x)}\right).$$

This completes the proof. □

4 The generalization of *p*th order of the operators

In this section, we consider the pth order generalization of the operators (1.6) as follows:

$$K_{n,\alpha,p}^{(r,s)}(f;x)=\left(\frac{n+s}{n}\right)^n(n+s+1)\sum_{k=0}^{n}p_{n,k,\alpha}^{(r,s)}(x)\int_{\frac{k+r}{n+s+1}}^{\frac{k+r+1}{n+s+1}}\sum_{j=0}^{p}f^{(j)}(u)\frac{(x-u)^j}{j!}du, \tag{4.1}$$

where $f\in C^p[0,1]$, $p=0,1,2....$ For the above generalized operators (4.1), it can be easily seen that at $p=0$ reduces to the operators (1.6). For more details, the readers can refer to [13, 16, 26, 27].

Theorem 4.1 *Suppose* $\alpha\in[0,1]$, $f\in C^p[0,1]$ and $f^{(p)}\in Lip_M(\xi)$, *then for all* $x\in\left[\frac{r}{n+s},\frac{n+r}{n+s}\right]$, *we have*

$$\left|K_{n,\alpha,p}^{(r,s)}(f;x)-f(x)\right| \leq \frac{M}{(p-1)!}\frac{\xi}{\xi+p}B(\xi,p)*\left\|K_{n,\alpha}^{(r,s)}(f;x)-f(x)\right\|,$$

where B is Beta function.

Proof From (4.1), we have

$$f(x) - K_{n,\alpha,p}^{(r,s)}(f;x) = \left(\frac{n+s}{n}\right)^n (n+s+1) \sum_{k=0}^{n} p_{n,k,\alpha}^{(r,s)}(x) \times \int_{\frac{k+r}{n+s+1}}^{\frac{k+r+1}{n+s+1}} \left(f(x) - \sum_{j=0}^{p} f^{(j)}(u) \frac{(x-u)^j}{j!} \right) du \tag{4.2}$$

By Taylor's theorem (see [19]), we have

$$f(x) - \sum_{j=0}^{p} f^{(j)}(u) \frac{(x-u)^j}{k!} = \frac{(x-u)^p}{(p-1)!} \int_0^1 (1-y)^{p-1} \left(f^{(p)}(u + y(x-u)) - f^{(p)}(u) \right) dy \tag{4.3}$$

Since $f^{(p)} \in Lip_M(\xi)$, we have

$$\left| f^{(p)}(u + y(x-u)) - f^{(p)}(u) \right| \leq M y^{\xi} |x-u|^{\xi}. \tag{4.4}$$

From the definition of Beta function, we get

$$\int_0^1 y^{\xi} (1-y)^{p-1} dy = B(1+\xi, p) = \frac{\xi}{\xi + p} B(\xi, p). \tag{4.5}$$

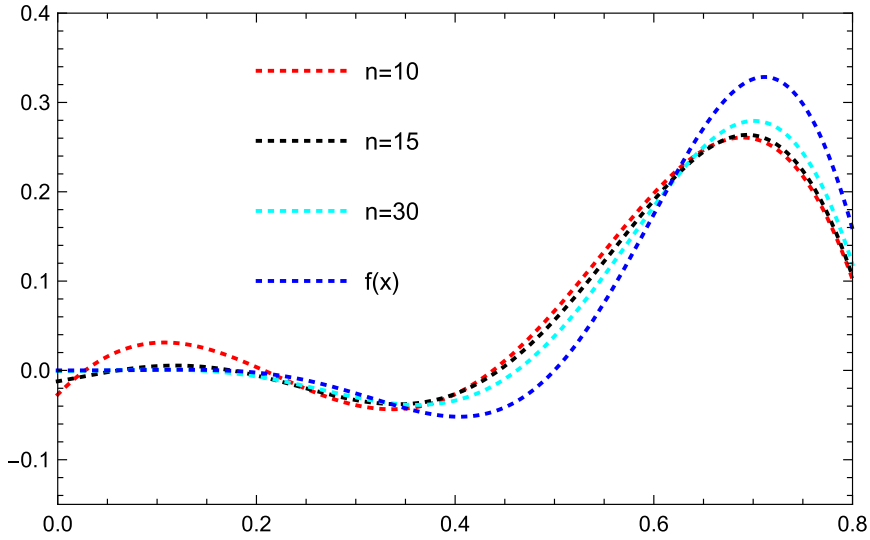
Using relation (4.4) and (4.5) in (4.3), we get

$$\begin{aligned} \left| f(x) - \sum_{j=0}^{p} f^{(j)}(u) \frac{(x-u)^j}{k!} \right| &\leq M \frac{|x-u|^p}{(p-1)!} \int_0^1 y^{\xi} |x-u|^{\xi} (1-y)^{p-1} dy \\ = M \frac{|x-u|^{p+\xi}}{(p-1)!} \int_0^1 y^{\xi} (1-y)^{p-1} dy &= M \frac{|x-u|^{p+\xi}}{(p-1)!} \frac{\xi}{\xi + p} B(\xi, p). \end{aligned} \tag{4.6}$$

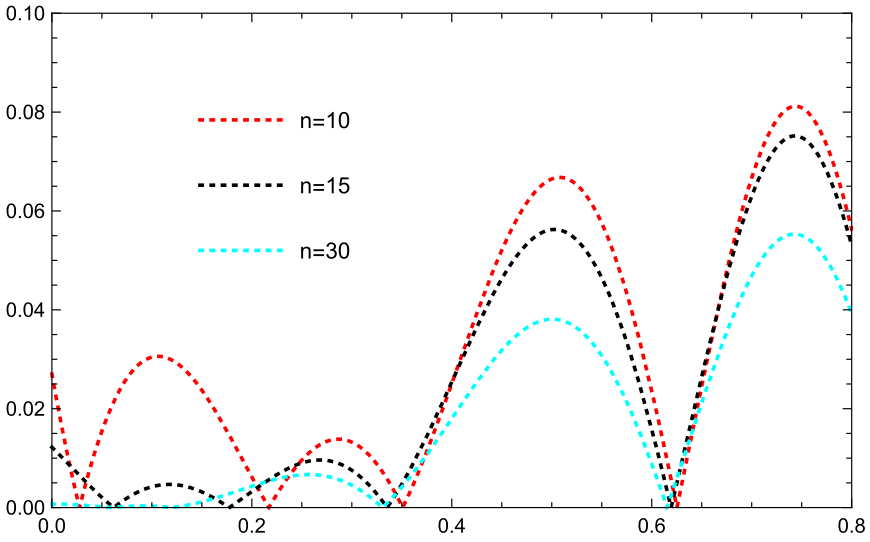
From (4.2) and (4.6), we obtain required result. □

5 Numerical results

Example 5.1 Let $f(x) = x^3 \cos(3\pi x)$, $\alpha = 0.2$, $r = 4$, $s = 5$ and $n \in \{10, 15, 30\}$. The convergence of the defined operators $K_{n,\alpha}^{(r,s)}$ towards the function $f(x)$ and the absolute error $\varepsilon_{n,\alpha}^{(r,s)}$ of the operators are shown in Fig. 1a and b respectively. The absolute error of the operators are also computed for some values in $\left[\frac{r}{n+s}, \frac{n+r}{n+s}\right]$ which is shown in Table 1.



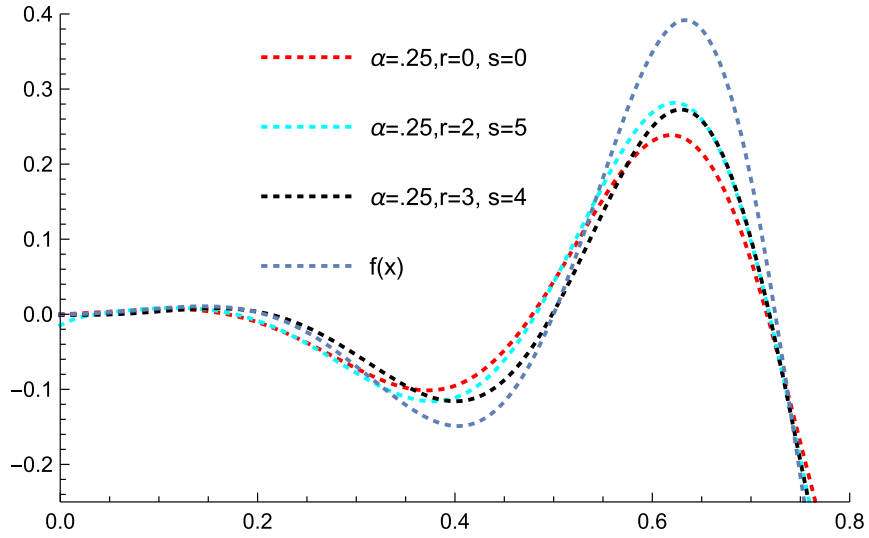
(a) Convergence of the operator $K_{n,\alpha}^{(r,s)}(f;x)$ towards function $f(x)$ for $(\alpha = 0.2,\ r = 4,\ s = 5)$.



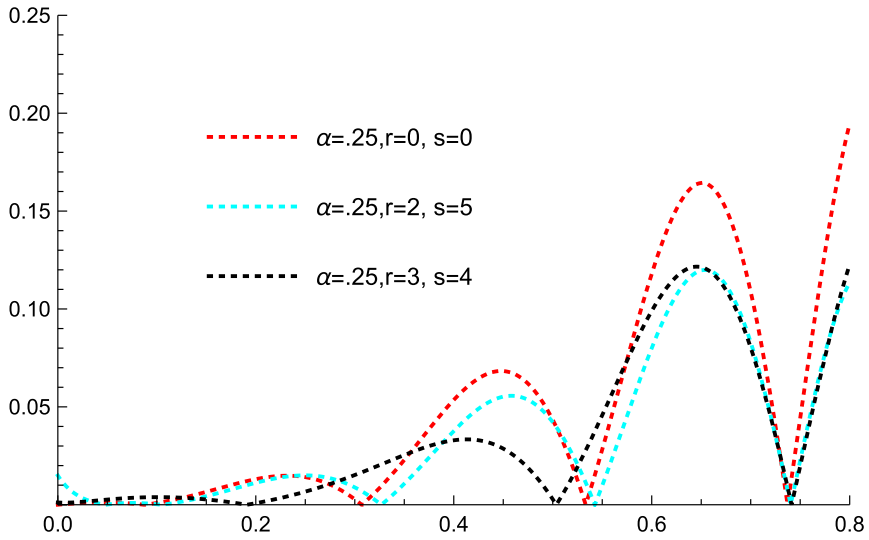
(b) The absolute error $\varepsilon_{n,\alpha}^{(r,s)}$ for $(\alpha = 0.2,\ r = 4,\ s = 5)$.

Fig. 1 **a**: Convergence of the operator $K_{n,\alpha}^{(r,s)}(f;x)$ towards function $f(x)$ for $(\alpha = 0.2,\ r = 4,\ s = 5)$; **b**: The absolute error $\varepsilon_{n,\alpha}^{(r,s)}$ for $(\alpha = 0.2, r = 4, s = 5)$

Example 5.2 Let $f(x) = x^2 \sin\left(2\pi\left(x + \frac{1}{2}\right)^2\right)$ and $n = 50$. The comparison of convergence of proposed operators $K_{n,\alpha}^{(r,s)}$ (cyan, $\alpha = 0.25$, $r = 2$, $s = 5$), (black, $\alpha = 0.25$, $r = 3$, $s = 4$) and the family of Bernstein–Kantorovich operators (1.4) (red, $\alpha = 0.25$, $r = 0$, $s = 0$) is presented in Fig. 2a. The absolute error $\varepsilon_{n,\alpha}^{(r,s)}$ of the operators is depicted in Fig. 2b and also computed for some values in $\left[\frac{r}{n+s}, \frac{n+r}{n+s}\right]$, which is shown in Table 2.



(a) x for $(n = 50)$



(b) The absolute error $\varepsilon_{n,\alpha}^{(r,s)}$ for $(n = 50)$.

Fig. 2 **a**: x for $(n = 50)$; **b**: The absolute error $\varepsilon_{n,\alpha}^{(r,s)}$ for $(n = 50)$

Table 1 Absolute error of the operators $K_{n,\alpha}^{(r,s)}$ with function $f(x) = x^3 \cos(3\pi x)$ for $\alpha = 0.2$, $r = 4$, $s = 5$

x	n = 10	n = 15	n = 30
0.1	.030426	.004204	.000300
0.2	.006182	.003235	.004314
0.3	.013229	.007541	.004520
0.4	.025076	.025405	.018002
0.5	.066466	.056224	.038098
0.6	.023637	.015840	.008822
0.7	.066703	.062968	.046879
0.8	.056277	.053033	.039276

Table 2 Absolute error of the operators with function $f(x) = x^2 \sin\left(2\pi\left(x+\frac{1}{2}\right)^2\right)$ for $n = 50$

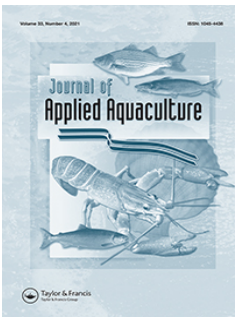
x	$\alpha = 0.25$ $r = 0$ $s = 0$	$\alpha = 0.25$ $r = 2$ $s = 5$	$\alpha = 0.25$ $r = 3$ $s = 4$
0.1	.000904	.000260	.003880
0.2	.013064	.011202	.000805
0.3	.002768	.008868	.015166
0.4	.053540	.037694	.032918
0.5	.043356	.042120	.002801
0.6	.118005	.080048	.098996

References

1. Agratini, O.: Kantorovich-type operators preserving affine functions. Hacettepe J. Math. Statis. **45**(6), 1657–1663 (2016)
2. Acar, T., Aral, A., Mohiuddine, S.A.: On Kantorovich modification of (p, q)-Bernstein operators. J. Inequal. Appl. **42**(3), 1459–1464 (2018)
3. Altomare, F., Campiti, M.: Korovkin-Type Approximation Theory ant Its Application. De Gruyter Studies in Mathematics Vol. 17, Walter de Gruyter and Co. Berlin, 1995
4. Agrawal, P.N., Gupta, V., Kumar, A.S.: On q-analouge of Bernstein–Schurer–Stancu operators. Appl. Math. Comput. **219**(14), 7754–7764 (2013)
5. Agrawal, P. N., Ispir, N., Kajla, A.: Approximation properties of Lupas–Kantorovich operators based on Polya distribution, Rendiconti del Circolo Matematico di Palermo Series 2, **65**(2) (2016), 185–208
6. Bernstein, S.N.: Demonstration of the Weierstrass theorem based on the calculation of probabilities. Common Soc. Math. Charkow Ser. 2t, **13** (1912), 1–2
7. Chen, X., Tan, J., Liu, Z., Xie, J.: Approximation of functions by a new family of generalized Bernstein operators. J. Math. Anal. Appl. **450**, 244–261 (2017)
8. Deo, N., Pratap, R.: α-Bernstein–Kantorovich operators. Afr. Mat. **31**, 609–618 (2019)
9. DeVore, R.A., Lorentz, G.G.: Constructive Approximation. Springer, Berlin, Germany (1993)
10. Ditzian, Z., Totik, V.: Moduli of Smoothness. Springer, New York (1987)
11. Dhamija, M., Deo, N.: Better approximation result of Bernstein-Kantorovich operators. Lobachevskii J. Math **38**(1), 94–100 (2017)
12. Dhamija, M., Pratap, R., Deo, N.: Approximation by Kantorovich form of modified Szász–Mirakyan operators. Appl. Math. Comput. **317**, 109–120 (2018)
13. Doğru, O.: Approximation order and asymptotic for generalized Meyer-König and Zeller operators. Math. Balkanica. **12**(34), 359–368 (1998)
14. Gadjiev, A.D.: Ghorbanalizadeh, approximation properties of a new type of Bernstein–Stancu polynomials of one and two variables. Appl. Math. Comput. **216**(3), 890–901 (2010)
15. Gupta, V., Tachev, G., Acu, A.M.: Modified Kantorovich operators with better approximation properties. Numer. Algorithms **81**, 125–149 (2019)
16. Içöz, G., Kantorovich, A.: Variant of new type Bernstein-Stancu polynomials. Appl. Math. Comput. **218**, 8552–8560 (2012)
17. Indrea, A.D.: A particular class of linear and positive Stancu type operators. Acta Univ. Apulensis Math. Inform. **31**, 249–246 (2012)
18. Jiang, B., Yu, D.: On Approximation by Bernstein–Stancu polynomials in movable compact disks. Results Math. **72**, 1535–1543 (2017)
19. Kirov, G.H.: Approximation with Quasi-Splines, (1st ed), CRC Press (1992). https://doi.org/10.1201/9781003062929
20. Kantorovich, L. V.: Sur certains développements suivant les polynomes de la forme de S. Bernstein, I, II, C.R Acad URSS., **20**, 563-568 (1930)
21. Korovokin, P.P.: On convergence of linear operators in the space of continuous functions (Russian). Dokl. Acad. Nauk. SSSR (NS) **90**, 961–964 (1953)
22. Mohiuddine, S.A., Acar, T., Alotaibi, A.: Construction of a new family of Bernstein–Kantorovich operators. Math. Meth. Appl. Sci. **40**(18), 7749–7759 (2017)

23. Mursaleen, M., Ansari, K.J., Khan, A.: Approximation by Kantorovich type q-Bernstein Stancu operators. Complex Anal. Oper. Theory **11**, 85–107 (2017)
24. Özarslan, M.A., Aktuglu, H.: Local approximation for certain King type operators. Filomat **27**, 173–181 (2013)
25. Özarslan, M.A., Duman, O., Srivastava, H.M.: Statistical approximation results for Kantorovich-type operators involving some special polynomials. Math. Comput. Model **48**(3–4), 388–401 (2008)
26. Özarsalan, O., Doğru, M.A., Taşdelen, F.: On positive operators involving a certain class of generating functions. Studia Sci. Math. Hungar. **41**(4), 415–429 (2004)
27. Rahman, S., Mursaleen, M., Acu, A.M.: Approximation properties of λ-Bernstein Kantorovich operators with shifted Knots. Math. Meth. Appl. Sci. **42**(11), 4042–4053 (2019)
28. Srivastava, H.M., Özger, F., Mohiuddine, S.A.: Construction of Stancu-Type Bernstein operators based on Bézier bases with shape parameter λ. Symmetry (2019). https://doi.org/10.3390/sym11030316
29. Weierstrass, K.: Über die analytische Darstellbarkeit sogenannter willkürlicher Functionen einer reellen Veränderlichen. Sitzungsberichteder der Koniglich preussischen Akademie der Wissenschcaften zu Berlin **633–639**, 789–805 (1885)

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



Journal of Applied Aquaculture

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/wjaa20

The study of effect of vitamin C and *Achyranthes aspera* seeds enriched diets on the growth, biochemical composition, digestive enzyme activities and expressions of genes involved in the biosynthesis of fatty acids in Snow trout *Schizothorax richardsonii* (Gray, 1832)

Guddu Kumar, JaiGopal Sharma, Ravi Kumar Goswami, Avanish Kumar Shrivastav, Neelesh Kumar, Suresh Chandra & Rina Chakrabarti

To cite this article: Guddu Kumar, JaiGopal Sharma, Ravi Kumar Goswami, Avanish Kumar Shrivastav, Neelesh Kumar, Suresh Chandra & Rina Chakrabarti (2021): The study of effect of vitamin C and *Achyranthes aspera* seeds enriched diets on the growth, biochemical composition, digestive enzyme activities and expressions of genes involved in the biosynthesis of fatty acids in Snow trout *Schizothorax richardsonii* (Gray, 1832), Journal of Applied Aquaculture, DOI: 10.1080/10454438.2021.1985679

To link to this article: https://doi.org/10.1080/10454438.2021.1985679

View supplementary material

Published online: 19 Oct 2021.

Submit your article to this journal

Article views: 59

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=wjaa20

JOURNAL OF APPLIED AQUACULTURE
https://doi.org/10.1080/10454438.2021.1985679

Taylor & Francis
Taylor & Francis Group

Check for updates

The study of effect of vitamin C and *Achyranthes aspera* seeds enriched diets on the growth, biochemical composition, digestive enzyme activities and expressions of genes involved in the biosynthesis of fatty acids in Snow trout *Schizothorax richardsonii* (Gray, 1832)

Guddu Kumar[a], JaiGopal Sharma[b], Ravi Kumar Goswami[a], Avanish Kumar Shrivastav[b], Neelesh Kumar[b], Suresh Chandra[c], and Rina Chakrabarti[a]

[a]Department of Zoology, University of Delhi, Delhi, India; [b]Department of Biotechnology, Delhi Technological University, Delhi, India; [c]Directorate of Coldwater Fisheries Research, Bhimtal, Uttrakhand, India

ABSTRACT

Snow trout, *Schizothorax richardsonii* (Gray, 1832), is an economically important fish in the Himalayan region. Snow trout fry (118.53 ± 5.42 mg) were cultured in flow-through systems for 60 days. Fry were fed with four different diets containing: 0.08% vitamin C (D1), 0.5% *Achyranthes aspera* seeds (D2), both vitamin C and seeds (D3), and a control (D4). Average weight (271.82 ± 4.90 g) and specific growth rate (1.39 ± 0.03) were significantly higher in fish fed D3 compared to the others. Digestive enzyme activities were significantly higher in D3 compared to the others. EPA content was significantly higher in D2 (110.33 ± 0.94 mg 100 g^{-1}) and D3 (109.87 ± 1.87 mg 100 g^{-1}); DHA (342.70 ± 0.77 mg 100 g^{-1}) and *n*-3 PUFA (496.17 ± 1.82 mg 100 g^{-1}) contents were significantly higher in D3 compared to the others. The expression of *fads2d6* was significantly higher in D3 (6.70-fold) and *elovl5* in D2 and D3 (1.51–1.45-fold) compared to the others. Enriched diets improved fish flesh composition.

KEYWORDS

Schizothorax richardsonii; Vitamin C; *Achyranthes aspera*; DHA and EPA; *fads2d6* and *elovl5*

Introduction

Snow trout, *Schizothorax richardsonii* (Gray, 1832), belongs to the family Cyprinidae and is a commercially important fish of the Himalayan region. It contributes the maximum yield in the capture fisheries of hills and is a potential species for aquaculture (Kamalam et al. 2019; Petr 2002). Although artificial breeding of snow trout is successful (Joshi 2004), the proper culture technique has not been developed (Sarma et al. 2013). Moreover, the growth rate of snow trout is very slow, and the fish are susceptible to diseases

CONTACT Rina Chakrabarti aquaresearchlab@yahoo.co.in Department of Zoology, University of Delhi, Delhi-110007, India.

Supplemental data for this article can be accessed on the publisher's website.

© 2021 Taylor & Francis

(Moses, Nilssen, and Chakrabarti 2017). Nutrition plays a significant role in the health status of the cultivable species (Lall and Olivier 1993). Supplementation of growth promoters and immunostimulants in the diet enhances the growth and improves the immunity of the fish (Kord et al. 2021). An earlier study shows that the enrichment of feeds with vitamin C and *Achyranthes aspera* seeds increases the growth rate of snow trout larvae. The enhanced levels of digestive enzymes, myeloperoxidase, nitric oxide synthase, etc., are found in fish fed enriched diets (Moses, Nilssen, and Chakrabarti 2017). The enzyme L-gulonolactone oxidase, which helps in the synthesis of ascorbic acid from L-gulonolacetone, is lacking in teleosts (Sato, Nishikimi, and Udenfriend 1976). Many studies show that supplementation of vitamin C in diets of fishes improves health conditions, e.g., in rohu *Labeo rohita* (Misra et al. 2007), Jian carp *Cyprinus carpio* (Liu et al. 2011), tilapia *Oreochromis karongae* (Nsonga et al. 2009), GIFT tilapia *O. niloticus* (Huang et al. 2016), and discus fish *Symphysodon haraldi* (Liu et al. 2018). Various factors like species, development stage, and environment etc., influence the requirement of vitamin C. The supplementation is needed more at the earlier stages compared to adults (Imanpoor, Imanpoor, and Roohi 2017; Kolkovski et al. 2000).

The herb *Achyranthes aspera* L. belongs to the family Amaranthaceae. It grows in tropical climatic conditions and has pharmaceutical applications. Many studies show that the seed of *A. aspera* enhances the immunity of different carps, and it protects the fishes against pathogens (Chakrabarti and Srivastava 2012; Chakrabarti et al. 2014; Kumar et al. 2019; Rao and Chakrabarti 2005; Rao et al. 2004; Sharma et al. 2019). The feeding of a seed-enriched diet protects the carps from harmful UV-B radiation (Singh, Sharma, and Chakrabarti 2013a, 2013b). The seed-supplemented diet also enhances the survival and growth rate of fishes. Chakrabarti et al. (2012) have observed the presence of saturated fatty acids (SFA), palmitic acid (23%), stearic acid (2%), and two long-chain polyunsaturated fatty acids (LC-PUFA)—oleic acid (33%) and linolenic acid (42%)—in the seeds of *A. aspera*.

Earlier studies show that freshwater teleosts are capable of converting linoleic acid (LA; 18:2 *n*-6) and α-linolenic acid (ALA; 18:3 *n*-3) to LC-PUFA—arachidonic acid (AA; 20: 4 *n*-6), eicosapentaenoic acid (EPA; 20:5 *n* − 3), and docosahexaenoic acid (DHA; 22:6 *n*-3) (Buzzi, Henderson, and Sargent 1997; Cook 1996; Glencross 2009). Therefore, the essential fatty acid requirements of freshwater teleosts can usually fulfill dietary LA and ALA (Tocher et al. 2001; Tocher, Dabrowski, and Hardy 2010). The conversion of LA and ALA to LC-PUFA requires a series of fatty acyl desaturase (*fad*) and elongation of very long-chain fatty acids (*elovl*) enzymes such as *elovl5* and *elovl2* (Torstensen and Tocher 2010). Fatty acid synthase (*fas*) is a key enzyme that regulates the de novo biosynthesis of long-chain fatty acids from acetyl-CoA and malonyl-CoA in the presence of NADPH (Dong et al. 2014). So far,

no study has been conducted to evaluate the effect of seeds of *A. aspera* on the fatty acid composition and the expressions of genes involved in the biosynthesis of fatty acids in snow trout.

The present study aims to evaluate the effect of vitamin C and seeds of *Achyranthes aspera* supplemented diets on the growth performance, digestive enzyme activities, biochemical composition, and expressions of specific genes involved in the biosynthesis of fatty acids—*fads2d6, fas, elovl5*, and *elovl2*—in the fry of snow trout (*Schizothorax richardsonii*).

Materials and methods

Culture of fry

Snow trout were bred in the wet laboratory of DCFR, Champawat, Uttrakhand, India, in September 2017, and larvae were grown up to fry stage (118.53 ± 5.43 mg). The experiment was started October 8 and was continued up to December 6, 2017. Fry were cultured under four different feeding regimes and maintained in four separate flow-through systems. Each flow-through system consisted of four culture units (10 L each); these four units were considered replicates. The flow rate of the water was 1.5 L min^{-1}. The stocking density was 10 fry per $unit^{-1}$ (10 × 4 = 40 fry per treatment).

Formulation of diets and feeding

Four diets were formulated using fish meal as protein source (Table 1). The three experimental diets were: diet 1 (D1) was supplemented with vitamin C (0.08%), diet 2 (D2) was supplemented with *A. aspera* (0.5%) seeds, and diet 3 (D3) was enriched with vitamin C (0.08%) and seeds (0.5%); the control diet 4 (D4) had no vitamin C or seeds. The concentrations of vitamin C and seeds in the diets were selected based on a previous experiment by Moses, Nilssen, and Chakrabarti (2017). The amino acids and fatty acids compositions of the

Table 1. Formulation of diets and their composition.

Ingredient (g kg^{-1})	Diet 1	Diet 2	Diet 3	Diet 4
Fish meal	450	450	450	450
Fish oil	10	10	10	10
Wheat flour	535.2	531	530.2	536
Vitamins & minerals	4	4	4	4
Vitamin C	0.8	–	0.8	–
A. aspera	–	5	5	–
Proximate composition of diets (% dry weight)				
Lipid	7.43 ± 0.03	6.86 ± 0.03	7.00 ± 0.05	7.09 ± 0.21
Protein	35.84 ± 1.12	35.92 ± 0.26	36.00 ± 0.23	36.15 ± 1.13
Moisture	7.37 ± 0.08	7.39 ± 0.12	7.00 ± 0.08	7.38 ± 0.03
Ash	9.36 ± 0.11	9.20 ± 0.03	8.84 ± 0.11	9.04 ± 0.05

Note. Diet 1, vitamin C 0.08%; Diet 2, *A. aspera* seeds 0.5%; Diet 3, vitamin C 0.08% and *A. aspera* seeds 0.5%; Diet 4, control.

formulated diets were evaluated (Tables S1 and S2 in the supplementary materials). Snow trout fry were fed at the rate of 5% of body weight every day. The whole amount of feed was divided into two parts and was given at 9:00 a.m. and 4:00 p.m. Satiation limit was assessed by feeding the fish with different amount of feeds; it was observed that feeding of fish at the rate of 2.5% of body weight was suitable. After 1 h of feeding, the uneaten feed was removed. The experiment was continued for 60 days.

Survival rate, growth, and feed conversion ratio

Survival rate (%) of snow trout was recorded. The weight of individual fish was recorded (Libror AEG-120, Shimadzu, Japan) after 60 days of culture. Specific growth rate (SGR) and feed conversion ratio (FCR) were calculated as follows:

$$Specific\ growth\ rate\ (SGR\ \%) = [ln\ (final\ weight) - ln\ (initial\ weight)] \times 100/study\ period$$

$$Feed\ conversion\ ratio\ (FCR) = feed\ intake/body\ weight\ gain$$

Digestive enzyme assays

The digestive tract was collected from individual snow trout. The digestive tracts of six fish were pooled to make one replicate, and four replicates were used per treatment. The sample was prepared in the following sequence: homogenization of pooled sample in ice-cold distilled water, following centrifugation (10,000 g for 30 min at 4°C) and collection of supernatant. The supernatant was used for the enzyme assay. All enzymes were assayed using a fluorometer (Biotek Synergy H1, USA). The EnzChek ® Ultra Amylase Assay Kit (E33651, Invitrogen, Oregon, USA) was used for the estimation of amylase activity. The excitation and emission were recorded at 485 nm and 520 nm respectively. The total protease activity was estimated based on a green fluorescence technique. The EnzChek® Protease Assay Kit (E6638, Molecular ProbesTM, Invitrogen) was used for the study. The change in fluorescence was estimated at excitation 485 nm and emission 530 nm. The trypsin activity was assayed using Na-benzoyl-L-arginin-methyl-coumarinylamide (Sigma, USA) as a substrate (Ueberschar 1988). The change in fluorescence was recorded at 380 nm (excitation) and 440 nm (emission). The chymotrypsin activity was assayed using succinyl-Leu-Val-Tyr-4-methyl-coumaryl -7-aminde (Sigma) as substrate (Cao et al. 2000). The excitation and emission were recorded at 380 nm and 450 nm respectively. Lipase was estimated using 4-methylumbelliferyl butyrate (4-MUB, Sigma) as substrate (Roberts 1985).

The excitation and emission were studied at 365 nm and 450 nm respectively. Total protein was estimated using bovine serum albumin as a standard (Bradford 1976).

Biochemical composition

The proximate composition—crude protein, crude lipid, moisture, and ash contents were assayed following standard methods (AOAC 2000). The crude protein was estimated following the Kjeldahl technique. In this assay, total nitrogen (%) was measured and was multiplied with 6.25 (N × 6.25) to obtain the protein (%) content. The crude lipid content was assayed gravimetrically (Folch, Lees, and Sloane-Stanley 1957). Moisture content was estimated by keeping the sample at 105°C for 3 h; for the estimation of ash, the sample was kept at 550°C for 8 h.

Fatty acids

The acid transesterification reaction was used for FAME (fatty acid methyl esters) estimation using a methylation reagent (1%). Crude lipid (1 mg) isolated from each sample was added with 2 mL of methylation reagent, 1 mL of toluene, and 1 mL of (C17:0) internal standard. After vortexing and proper mixing, the lipid sample was flushed with nitrogen. The sample was incubated at 50°C for 16 h (Christie 2003). The FAME (1 mg mL^{-1}) was isolated, purified, separated, and quantified using a Perkin Elmer Gas Chromatograph (Clarus 580 Autosampler, USA). Zebronics wax column 60 m (length) × 0.32 mm (internal diameter) × 0.25 µm (film thickness) was used for fatty acid analysis. FAMEs were compared with the known standards and the published data (Tocher and Harvie 1988). The fatty acid content was expressed as mg 100 g^{-1} of wet sample.

Amino acids

The amino acid composition of the sample was assayed using a Hitachi L-8900 Amino acid analyzer (Japan). All amino acids (except methionine, cysteine, and tryptophan) were hydrolyzed using HCl (6 N). In a digestion tube, a 30 mg sample was taken; 10 mL HCl was added to the sample, then flushed with nitrogen and sealed. The tube was kept for 22 h at 110°C for hydrolysis (Chakrabarti et al. 2018). Then the sample was kept in a nitrogen evaporator (PCi Analytics Pvt. Ltd., Maharastra, India) to evaporate acid. The dried samples were dissolved in HCl (0.02 N), and the final concentration of protein was 0.5 mg mL^{-1}. The sample (20 µL) was injected in the autosampler. The performic acid was first used to oxidize methionine and cysteine, and then hydrobromic acid (48%) was added in

the sample. The methanesulfonic acid (4 N) and 3-(2-aminoethyl) indole were used for the measurement of tryptophan. The amino acid concentrations were measured using standard solutions (Wako Pure Chemical Industries, Limited). The glutamine and tryptophan were estimated using freshly prepared standard (Sigma-Aldrich, USA). The amino acid content was expressed as g 100 g^{-1} of sample (wet weight).

Gene expression

Four genes—*fads2d6* (delta-6 desaturase), *fas* (fatty acid synthase), *elovl5* (elongase of very long chain fatty acids 5), and *elovl2* (elongase of very long chain fatty acids 2) were selected for the study; β-actin was used as reference gene for normalization. Primers were designed as per the sequences of genes available on the NCBI (Table 2). RNA was isolated using Trizol reagent following the instructions provided with the chemical. The purity of isolated RNA was estimated. The RNA was treated with DNase 1 to avoid DNA contamination (DNase kit, Sigma Aldrich). Purified RNA was converted to cDNA with an ABi kit (USA). The PCR amplification with β-actin confirmed cDNA, and bands were separated on agarose gel. SYBR green was used for the study of relative mRNA expressions (RT-PCR, Quant Studio 6 Flex, ABi, USA). The amplification efficiency of primers was estimated with their melt curves. Total volume of reaction mixture per well was 10 µL that was composed of 5 µL of SYBR green, 0.5 µL of forward primer, 0.5 µL of reverse primer, 3 µL of nucleic acid-free water, and cDNA 1 µL. The amplification cycle in the RT-PCR was as follows: 10 min for preliminary denaturation, number of cycles was 40 (like 15 s at 95°C and 1 min at 60°C). The plate was sealed with a 96-well optical adhesive cover. The optical plate was placed in RT-PCR and a program was run using QuantStudio 6 Flex software v1.1 on the computer. On the basis of CT value, the fold change of a particular mRNA was estimated (Livak and Schmittgen 2001).

Table 2. Primers used for gene expression analysis.

Gene	Accession No.	Direction	Primer Sequence
β-actin	KU715835	Forward	5′-GCCGTGACCTGACTGACTAC-3′
		Reverse	5′-CTGCTCGAAGTCAAGAGCCA-3′
fads2d6	KJ576791	Forward	5′-CACAATGCCTCGGCACAAT-3′
		Reverse	5′-CCAGCCAGAGTTCTCCAGATTT-3′
fas	MF289408	Forward	5′-CAATGCCCAGCAGCATAAGG-3′
		Reverse	5′-TGGGCTTGTTGAACCTCGG-3′
elovl5	KF924199	Forward	5′-GGGCTGGCTGTATTTCCAGAT-3′
		Reverse	5′-GATGCCACCCCATTAGTGTGA-3′
elovl2	KR706498	Forward	5′-ATCAGTTTGGTCTGCCGGTT-3′
		Reverse	5′-CAGCACAATGAAGATGGTGTCC-3′

Water quality

Water quality parameters were monitored regularly. The Hach multimeter (HQ40D, USA) was used for the estimation of temperature, pH, conductivity, dissolved oxygen, and ammonia levels of water. The nitrite and nitrate contents were measured following the APHA (2017) method. Temperature, pH, and conductivity ranged from 9.0 to 19.2°C, 7.15 to 7.88, and 241.3 to 251.6 μS cm^{-1} respectively in different culture systems throughout the study period. The experiment was started October 8 and completed December 6, 2017. At the start of the experiment the water temperature was 19.2°C; in November the water temperature gradually reduced. The minimum 9.0°C was recorded in the first week of December, and the experiment was completed. Snow trout is a coldwater fish and survives well in temperatures ranging from 5 to 20°C (Barat et al. 2019). Dissolved oxygen, ammonia, nitrite, and nitrate levels of fish culture units ranged from 7.36 to 9.38, 0.06 to 0.8, 0.10 to 0.18, and 0.21 to 0.28 mg L^{-1} respectively throughout the experiment.

Statistical analysis

The growth parameters, digestive enzyme activities, and biochemical composition of snow trout were expressed as means ± SEM. Data were analyzed using ANOVA, one-way analysis of variance, and Duncan's multiple range test (DMR) (Montgomery 1984). Statistical significance was accepted at the $P < 0.05$ level.

Results

Survival and growth

In snow trout fed the control diet (D4), the survival rate was 93%; all fish survived (100%) in the other treatments. A significantly higher final average weight was recorded in snow trout fed the D3 diet (271.82 ± 4.90 mg) compared to fish fed the other diets. The weight gain was 1.45–2.76-fold higher in fish fed the experimental diets compared to the control. Significantly higher SGR (1.39 ± 0.03%) and lower FCR (1.55 ± 0.07) were observed in snow trout fed the D3 diet compared to the fish cultured in the other three different feeding regimes (Table 3).

Table 3. Survival, average weight, specific growth rate (SGR), feed conversion ratio (FCR), and digestive enzyme activities found in *S. richardsonii* fed with four different diets. Means (n = 4) sharing different letters in the same row are significantly ($P < 0.05$) different.

Parameter	Diet 1	Diet 2	Diet 3	Diet 4
Survival (%)	100 ± 0.00a	100 ± 0.00a	100 ± 0.00a	93 ± 2.50b
Average weight (mg)	198.72 ± 2.55b	201.22 ± 1.71b	271.82 ± 4.90a	173.74 ± 2.67 c
SGR (%)	0.86 ± 0.02b	0.87 ± 0.02b	1.39 ± 0.03a	0.64 ± 0.01 c
FCR	2.48 ± 0.03b	2.61 ± 0.16b	1.55 ± 0.07 c	3.26 ± 0.05a
Amylase ($mU\ \mu l^{-1} mg^{-1}$protein min^{-1})	19.93 ± 0.51b	21.46 ± 0.58b	33.28 ± 0.20a	12.09 ± 0.23 c
Total Protease (Fluorescence change $unit^{-1}$)	108.55 ± 4.07b	109.71 ± 1.83b	136.69 ± 3.0a	99.96 ± 2.54 c
Trypsin (μmol AMC mg^{-1}protein min^{-1})	141.66 ± 5.19b	150.87 ± 7.78b	254.63 ± 5.45a	72.86 ± 2.31 c
Chymotrypsin (μmol AMC mg_{-1}protein min^{-1})	74.80 ± 1.44b	80.34 ± 1.77b	162.35 ± 1.72a	59.83 ± 0.85 c
Lipase (μmol 4MU mg^{-1}protein min^{-1})	74.28 ± 2.06b	69.70 ± 1.66 c	157.68 ± 6.23a	78.68 ± 1.58b

Note. Diet 1, vitamin C 0.08%; Diet 2, *A. aspera* seeds 0.5%; Diet 3, vitamin C 0.08% and *A. aspera* seeds 0.5%; Diet 4, control.

Digestive enzyme activities

Significantly higher amylase, total protease, trypsin, chymotrypsin, and lipase activities were observed in snow trout fed the D3 diet compared to fish cultured in the other three different feeding regimes (Table 3). There was no significant difference in amylase, protease, trypsin, and chymotrypsin activities between snow trout fed the D1 and D2 diets. These enzyme activities were minimum in fish fed the control diet. Lipase activity was minimum in snow trout fed the D2 diet.

Proximate composition of muscle

The moisture contents ranged from 76.25% to 76.65% in different diets fed snow trout. The protein content was significantly higher in snow trout fed the D1 and D4 diets compared to the fish fed diets D2 and D3 (Table 4).

Table 4. Biochemical composition of muscle (wet weight) of *S. richardsonii* fed with four different diets. Means (n = 4) sharing different letters in the same row are significantly ($P < 0.05$) different.

Parameter (%)	Diet 1	Diet 2	Diet 3	Diet 4
Moisture	76.25 ± 0.36a	76.65 ± 0.54a	76.55 ± 0.17a	76.45 ± 0.35a
Protein	16.85 ± 0.10a	16.22 ± 0.27b	16.17 ± 0.16b	16.76 ± 0.20a
Lipid	3.20 ± 0.03b	3.32 ± 0.04b	3.54 ± 0.02a	3.11 ± 0.02 c
Ash	2.61 ± 0.05a	2.39 ± 0.02b	2.41 ± 0.01b	2.37 ± 0.02b

Note. Diet 1, vitamin C 0.08%; Diet 2, *A. aspera* seeds 0.5%; Diet 3, vitamin C 0.08% and *A. aspera* seeds 0.5%; Diet 4, control.

Table 5 (A) Amino acid profile (g 100 g^{-1} wet weight) of muscle of *S. richardsonii* fed with four different diets. Means ($n = 4$) sharing different letters in the same row are significantly ($P < 0.05$) different. Essential amino acids.

Amino acid	Diet 1			Diet 2			Diet 3			Diet 4		
Histidine (His)	0.395	±	0.001b	0.392	±	0.001b	0.405	±	0.003a	0.390	±	0.001b
Isoleucine (Ile)	0.630	±	0.002a	0.616	±	0.002b	0.632	±	0.001a	0.620	±	0.001b
Leucine (Lue)	1.126	±	0.004a	1.072	±	0.005 c	1.101	±	0.004b	1.109	±	0.005b
Lysine (Lys)	1.247	±	0.006a	1.210	±	0.005 c	1.247	±	0.002a	1.230	±	0.002b
Methionine (Met)	0.396	±	0.001a	0.289	±	0.001 c	0.298	±	0.003 c	0.388	±	0.006b
Phenylalanine (Phe)	0.641	±	0.010a	0.607	±	0.003 c	0.624	±	0.002b	0.628	±	0.016b
Threonine (Thr)	0.725	±	0.004a	0.686	±	0.006 c	0.701	±	0.001b	0.725	±	0.012a
Tryptophan (Trp)	0.219	±	0.002d	0.222	±	0.002 c	0.228	±	0.004a	0.226	±	0.010b
Valine (Val)	0.731	±	0.002a	0.715	±	0.004 c	0.733	±	0.006a	0.721	±	0.004b
∑ Essential amino acids	6.109	±	0.022a	5.808	±	0.029d	5.970	±	0.011 c	6.037	±	0.046b

Note. Diet 1, vitamin C 0.08%; Diet 2, *A. aspera* seeds 0.5%; Diet 3, vitamin C 0.08% and *A. aspera* seeds 0.5%; Diet 4, control.

Table 5 (B) Nonessential and free amino acids.

Amino acid	Diet-1			Diet-2			Diet-3			Diet-4		
Nonessential amino acids												
Alanine (Ala)	0.878	±	0.002a	0.845	±	0.005c	0.871	±	0.002a	0.865	±	0.003b
Arginine (Arg)	1.504	±	0.009b	1.352	±	0.211c	1.575	±	0.046a	1.574	±	0.002a
Asparatate (Asp)	1.500	±	0.003a	1.435	±	0.009d	1.465	±	0.002c	1.485	±	0.010b
Cysteine (Cys)	0.171	±	0.001a	0.154	±	0.003b	0.157	±	0.001b	0.170	±	0.001a
Glutamic Acid (Glu)	2.387	±	0.009a	2.293	±	0.006a	2.345	±	0.002a	2.375	±	0.016a
Glycine (Gly)	0.939	±	0.004a	0.901	±	0.008a	0.927	±	0.001a	0.926	±	0.003a
Proline (Pro)	0.526	±	0.003b	0.527	±	0.004b	0.541	±	0.004a	0.518	±	0.002 c
Serine (Ser)	0.697	±	0.001a	0.658	±	0.004c	0.671	±	0.002b	0.695	±	0.009a
Tyrosine (Tyr)	0.512	±	0.002a	0.485	±	0.001c	0.502	±	0.002b	0.501	±	0.004b
∑ Non-essential amino acids	9.113	±	0.033a	8.650	±	0.251b	9.054	±	0.062a	9.110	±	0.051a
Free amino acids												
Phosphoserine	0.034	±	0.002a	0.023	±	0.001b	0.025	±	0.001b	0.022	±	0.002b
Taurine	0.170	±	0.001b	0.209	±	0.001a	0.214	±	0.004a	0.164	±	0.001 c
Cystathionine	0.072	±	0.002a	0.065	±	0.001c	0.069	±	0.001b	0.073	±	0.001a
β-Amino isobutyric acid	0.266	±	0.040a	0.204	±	0.009b	0.207	±	0.004b	0.259	±	0.066a
γ-Amino-n-butyric acid	0.041	±	0.016a	0.032	±	0.002b	0.029	±	0.001b	0.043	±	0.002a
Ethanol amine	0.141	±	0.007a		nd			nd		0.135	±	0.013a
Hydroxylysine	0.144	±	0.005a	0.142	±	0.005a	0.147	±	0.001a	0.154	±	0.020a
Ornithine	0.034	±	0.003a	0.035	±	0.001a	0.039	±	0.001a	0.036	±	0.005a
3 Methylhistidine	0.013	±	0.000a	0.016	±	0.001a	0.017	±	0.001a	0.012	±	0.000a
Hydroxy proline	0.140	±	0.006b	0.104	±	0.007c	0.092	±	0.005c	0.156	±	0.004a
∑ Free amino acids	1.054	±	0.154a	0.832	±	0.030b	0.840	±	0.015b	1.057	±	0.213a

Note. Diet 1, vitamin C 0.08%; Diet 2, *A. aspera* seeds 0.5%; Diet 3, vitamin C 0.08% and *A. aspera* seeds 0.5%; Diet 4, control.

A significantly higher lipid level was observed in snow trout fed diet D3 compared to the other treatments. The ash content was significantly higher in snow trout fed the D1 diet compared to the others.

Amino acid profile of muscle

Significantly higher histidine and tryptophan contents were observed in snow trout fed the D3 diet compared to fish fed the other diets (Table 5). Significantly higher isoleucine, lysine, and valine contents were found in snow trout fed diets D1 and D3 compared to the others. Significantly higher leucine, methionine, and phenylalanine levels were observed in snow trout fed the D1 diet compared to fish fed the other diets. Significantly higher threonine levels were recorded in snow trout fed the D1 and D4 diets compared to the fish fed diets D2 and D3. The total essential amino acids content was significantly higher in snow trout fed the vitamin C-supplemented diet (D1) compared to the fish cultured in the other three feeding regimes.

Nonessential amino acid alanine content was significantly higher in snow trout fed the D1 and D3 diets, and arginine content was significantly higher in snow trout fed the D3 and D4 diets compared to the others. Aspartate and tyrosine levels were significantly higher in snow trout fed the D1 diet compared to the fish cultured in the other three different feeding regimes. Cysteine and serine contents were significantly higher in snow trout fed the D1 and D4 diets compared to the D2 and D3 treatments. In snow trout fed different diets, glutamic acid and glycine contents were not significantly different. Proline content was significantly higher in snow trout fed the D3 diet compared to fish fed the other diets. The total nonessential amino acids content was significantly lower in snow trout fed the D2 diet compared to the others.

The free amino acids phosphoserine level was significantly higher in snow trout fed the D1 diet and taurine content in the D2 and D3 diets compared to the others. Cystathionine, β-amino isobutyric acid, and γ-amino-n-butyric acid contents were significantly higher in snow trout fed the D1 and D4 diets compared to the others. The hydroxyproline content was significantly higher in snow trout fed the D4 diet compared to the others. Significantly higher total free amino acids were recorded in snow trout fed the D1 and D4 diets compared to the D2 and D3 diets.

Fatty acid profile of muscle

In the muscle of snow trout fed the different diets, various saturated (SFA), monounsaturated (MUFA), *n*-6 polyunsaturated (*n*-6 PUFA), and *n*-3 polyunsaturated (*n*-3 PUFA) fatty acids were observed. Significantly higher myristic acid and pentadecanoic acid was observed in snow trout fed the D2 and D3 diets compared to the other two treatments (Table 6). Palmitic acid was significantly higher in snow trout fed the D2 diet compared to fish fed the

Table 6. Fatty acid profile (mg 100 g^{-1} wet weight) of muscle of *S. richardsonii* fed with four different diets. Means (n = 4) sharing different letters in the same row are significantly (P < .05) different.

Fatty acids	Diet-1			Diet-2			Diet-3			Diet-4		
Saturated fatty acids												
14:0	269.39	±	1.00b	271.31	±	1.34a	274.02	±	1.67a	264.34	±	2.09c
15:0	33.06	±	4.94b	38.63	±	0.84a	38.44	±	1.88a	33.48	±	3.56b
16:0	714.18	±	9.04b	728.72	±	1.14a	719.19	±	14.72b	705.27	±	10.99b
18:0	88.35	±	3.70ab	92.17	±	0.68a	85.03	±	4.16b	91.49	±	2.11a
22:0	5.37	±	0.31a	5.53	±	0.33a	4.41	±	0.55b	5.23	±	0.15a
24:0	9.78	±	0.82a	7.97	±	0.19b	7.96	±	0.41b	9.97	±	0.57a
∑SFA	1120.14	±	18.80a	1144.34	±	2.38a	1129.04	±	18.82a	1109.78	±	14.45a
Monounsaturated fatty acids												
16:1 *n*-9	669.40	±	1.49b	674.17	±	2.76a	679.20	±	7.07a	657.01	±	5.98c
18:1 *n*-9	829.03	±	3.05a	838.09	±	2.16a	838.70	±	7.41a	816.74	±	7.96b
20:1 *n*-9	100.17	±	0.64a	101.27	±	0.43a	100.55	±	0.43a	98.76	±	1.06a
22:1 *n*-9	37.70	±	0.31a	38.06	±	0.35a	37.00	±	0.91a	37.41	±	1.49a
24:1	62.98	±	0.83a	62.22	±	0.70a	63.81	±	0.93a	59.40	±	0.97b
∑ MUFA	1699.27	±	4.04b	1713.82	±	5.45a	1719.25	±	14.15a	1669.31	±	17.14c
n-6 Polyunsaturated fatty acids												
18:2 *n*-6	347.89	±	1.50b	351.11	±	0.54a	352.29	±	3.08a	343.82	±	2.90b
18:3 *n*-6	14.59	±	0.82b	15.91	±	0.20a	14.27	±	0.23b	14.59	±	0.36b
20:2 *n*-6	17.55	±	0.44a	17.32	±	0.49a	17.81	±	0.46a	18.55	±	0.65a
20:3 *n*-6	112.49	±	0.32a	113.18	±	0.35a	112.84	±	0.89a	111.48	±	0.70a
20:4 *n*-6	8.10	±	0.25a	7.94	±	0.31a	8.01	±	0.24a	7.63	±	0.41a
∑ *n*-6 PUFA	500.63	±	1.94a	505.46	±	0.64a	505.22	±	3.76a	496.07	±	4.23a
n-3 Polyunsaturated fatty acids												
18:3 *n*-3	33.89	±	0.15a	34.33	±	0.39a	34.44	±	0.25a	33.57	±	0.64a
18:4 *n*-3	8.79	±	0.26a	8.76	±	0.09a	9.16	±	0.19a	8.76	±	0.19a
20:5 *n*-3	107.81	±	0.78b	110.33	±	0.94a	109.87	±	1.88a	105.72	±	1.67b
22:6 *n*-3	339.73	±	2.67b	338.94	±	0.68b	342.70	±	0.77a	335.39	±	2.37c
∑ *n*-3 PUFA	490.22	±	2.37b	492.36	±	1.28b	496.17	±	1.82a	483.44	±	3.02c
n-3/*n*-6	0.98	±	0.00a	0.97	±	0.00a	0.98	±	0.00a	0.97	±	0.01a

Diet-1, vitamin C 0.08%; Diet-2, *A. aspera* seeds 0.5%; Diet-3, vitamin C 0.08% and *A. aspera* seeds 0.5%; Diet-4, control.

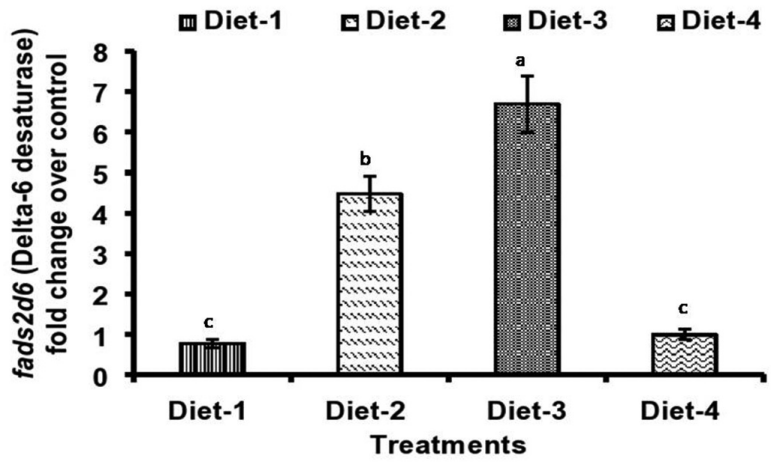


Figure 1. Relative mRNA expression of *fads2d6* gene over control, with reference to *β-actin*. Bars with different superscripts are significantly (P < 0.05) different (*n* = 4).

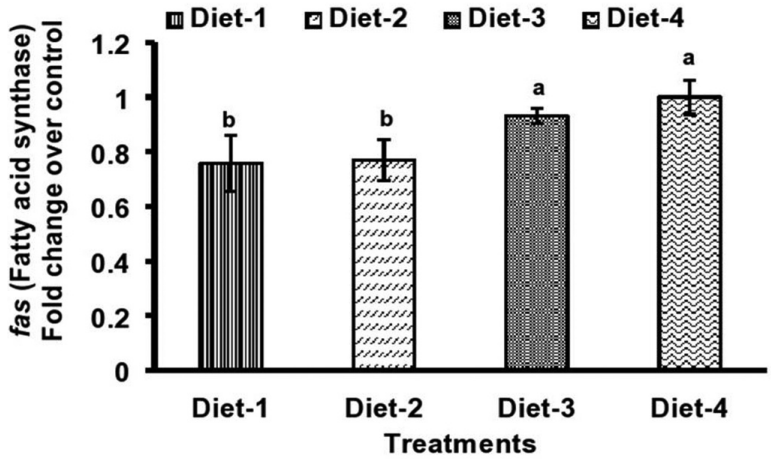


Figure 2. Relative mRNA expression of *fas* gene over control, with reference to *β-actin*. Bars with different superscripts are significantly (P < 0.05) different (*n* = 4).

other diets. In snow trout fed the D3 diet, significantly lower stearic acid and behenic acid were found compared to others. SFA content was minimum in snow trout fed the control diet (D4).

Among MUFAs, palmitoleic and oleic acids contents were significantly higher in snow trout fed the D2 and D3 diets compared to the others. The gondoic acid and erusic acid levels were not significantly different among fish fed different diets. The nervonic acid content was significantly lower in snow trout fed the D4 diet compared to the others. The total monounsaturated fatty

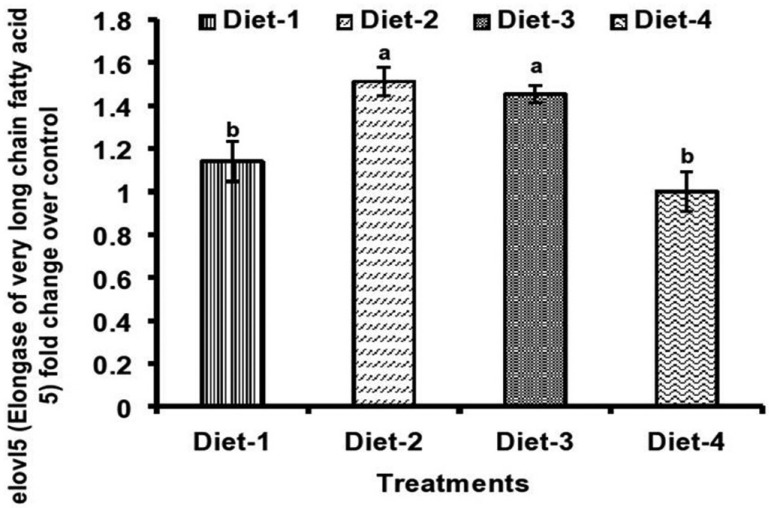


Figure 3.

Figure 3. Relative mRNA expression of *elovl5* gene over control, with reference to *β-actin*. Bars with different superscripts are significantly (P < 0.05) different (*n* = 4).

acid content was significantly higher in snow trout fed diets D2 and D3 compared to the others. In snow trout fed the D4 diet, the lowest MUFA content was observed.

Among *n*-6 PUFAs, the linoleic acid level was significantly higher in snow trout fed diets D2 and D3 and gamma linolenic acid in fish fed the D2 diet compared to the other diets. In snow trout fed the different diets, eicosadienoic, eicosatrienoic, and arachidonic acids contents were not significantly different.

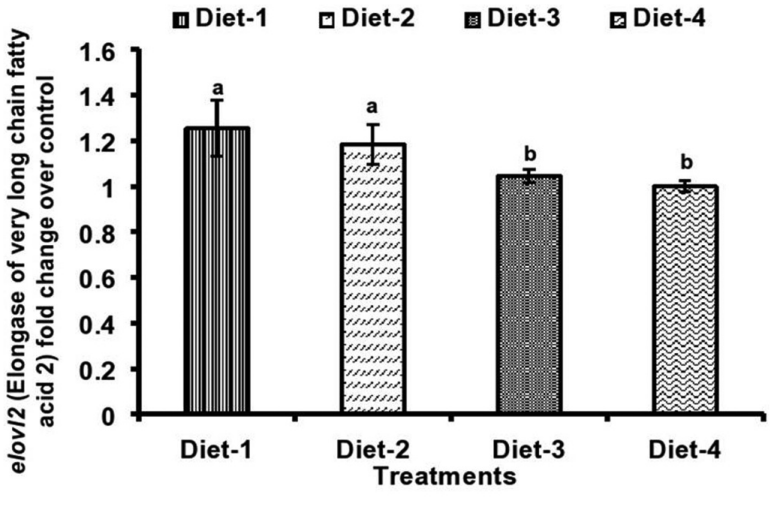


Figure 4.

Figure 4. Relative mRNA expression of *elovl2* gene over control, with reference to *β-actin*. Bars with different superscripts are significantly (P < 0.05) different (*n* = 4).

In snow trout fed the different diets, α-linolenic acid and stearidonic acid contents were not significantly different. A significantly higher eicosapentaenoic (EPA) acid level was observed in snow trout fed the D2 and D3 diets compared to the others. The docosahexaenoic acid (DHA) content was significantly higher in snow trout fed the D3 diet compared to other diets. Total *n*-3 PUFA content was significantly higher in snow trout fed the D3 diet compared to fish cultured in the other feeding regimes. A minimum *n*-3 PUFA content was found in snow trout fed the control diet (D4). The *n*-3/*n*-6 ranged from 0.97 to 0.98 in fish fed different diets.

Gene expression analysis

There was upregulation of delta-6 desaturase (*fads2d6*) in snow trout fed the D3 diet, and this feeding regime was followed by fish fed the D2 diet (Figure 1). Fatty acid synthase (*fas*) mRNA expression was downregulated in snow trout fed the enriched diets compared to the control (Figure 2). There were upregulations of *elovl5* in snow trout fed all enriched diets (Figure 3) and *elovl2* in snow trout fed the D1 and D2 diets compared to the others (Figure 4).

Discussion

The diets of snow trout fry were enriched with vitamin C, seeds of the plant *A. aspera*, and a combination of vitamin C and seeds. The positive effect of enriched diets was reflected on the survival rate of snow trout. The survival rate of snow trout was 93% in the control; in other three feeding regimes all fish survived (100%). The enhanced average weight was observed in snow trout fed the enriched diets compared to fish fed the control diet. Previous studies showed that the addition of vitamin C in diets enhanced the survival and growth and reduced the FCR of cultivable fishes such as rohu (Misra et al. 2007) and snow trout (Moses, Nilssen, and Chakrabarti 2017). In discus fish, supplementation of vitamin C and vitamin E resulted in higher SGR and lower FCR compared to the fish fed the control diet (Liu et al. 2018). Zou et al. (2019) recommended 71.46–150.26 mg kg^{-1} vitamin C for optimum growth of Chu's croaker *Nibea coibor*. Significantly higher average weight and SGR were found in snow trout fed both vitamin C and seeds enriched diet, D3 compared to the fish fed diets enriched with vitamin C/ seeds/control. In rohu, incorporation of 0.5% seeds of *A. aspera* increased the average weight and SGR (Sharma et al. 2019). The presence of ecdysterone in seeds increased the growth of fish (Chakrabarti et al. 2012). Ecdysterone increased protein synthesis in skeletal muscle (Goerlich-Feldmann et al. 2008) and thereby enhanced the growth of fish. The enrichment of diet with both vitamin C and seeds gave better results compared to the diet enriched with individual ingredients (vitamin C or seeds).

The digestibility of the consumed feed depends on the digestive enzyme activities of fish. All digestive enzymes activities assayed were significantly higher in snow trout fed the D3 diet. This resulted in efficient digestion of consumed diet and thereby the lowest FCR compared to fish fed other diets. All digestive enzyme activities (except lipase) were higher in fish fed diets with vitamin C and seeds compared to fish fed the control diet. Therefore, the highest FCR was observed in snow trout fed the control diet. In rohu fed seed-enriched diets, a lower FCR was recorded (Sharma et al. 2019; Singh et al. 2019).

Snow trout fed the D1 and D4 diets showed significantly higher protein contents compared to fish fed the D2 and D3 diets. There was no variation in the amino acid composition in snow trout cultured in four different feeding regimes; the amount of a specific amino acid varied. There were higher values of seven essential amino acids (isoleucine, leucine, lysine, methionine, phenylalanine, threonine, and valine) in snow trout fed the D1 diet and five amino acids (histidine, isoleucine, lysine, tryptophan, and valine) in fish fed the D3 diet compared to the fish fed the control diet, D4. Among nonessential amino acids, a higher value of alanine was found in snow trout fed both D1 and D3 diets compared to the fish fed diet D4; aspartate and tyrosine contents were maximum in fish fed the D1 diet and proline content was higher in D3 diets compared to snow trout fed the D4 diet. The free amino acid taurine content was significantly higher in fish fed the D2 and D3 diets. The enrichment of the diet with vitamin C and seeds increased the specific amino acid level in the muscles of fish. Several amino acids were found in the seeds of *A. aspera* (Goyal, Goyal, and Mehta 2007). Thus, feeding snow trout a seed-supplemented diet improved the amino acids contents of muscle. The enrichment of vitamin C in the diet of snow trout enhanced the protein synthesis and produced a higher amino acid profile. Vitamin C is a hydrogen donor and is associated with many hydroxylation reactions—in collagen maturation, catecholamine synthesis, and carnitine synthesis. Vitamin C helps in the synthesis of phenylalanine and tyrosine (Guillaume et al. 2001).

A significantly higher lipid content was found in fish fed the D3 diet. Enrichment of diets with seeds or vitamin C and seeds improved the fatty acid contents of the muscle of snow trout. Vitamin C showed no influence on the lipid and fatty acid contents of muscle of pacu (*Piaractus mesopotamicus*) or rainbow trout (*Oncorhynchus mykiss*) (Trattner et al. 2007; Trenzado, Morales, and de la Higuera 2008). Supplementation of vitamin C in the diets of grass carp (*Ctenopharyngodon idella*) improved the fatty acids (palmitoleic, oleic, linoleic, and linolenic acids and *n*-3 PUFA) contents in muscle compared to fish fed the control diet (Han et al. 2019). In snow trout fed the seed-enriched diet (D2), seven SFAs (myristic, pentadecanoic, palmitic, stearic, behenic acids, etc.) were found in highest proportion

compared to fish fed the other diets. In the muscle of snow trout fed the D1, D3, and D4 diets, 2 (behenic and lignoceric acids), 2 (myristic and pentadecanoic), and 3 (stearic, behenic, and lignoceric acids) SFAs were maximum respectively. Three MUFAs (palmitoleic, oleic, and nervonic acids) contents were maximum in snow trout fed the D2 and D3 diets. Among *n*-6 PUFA, the highest amount of linoleic acid was found in snow trout fed the D2 and D3 diets; γ-linolenic acid was maximum in fish fed the D2 diet. Among *n*-3 PUFAs, EPA content was significantly higher in fish fed the D2 and D3 diets; DHA content was highest in fish fed the D3 diet. Minimum EPA and DHA contents were observed in snow trout fed the D4 diet.

The present study showed that the supplementation of seeds in diets enhanced the fatty acids, especially DHA and EPA levels in the snow trout muscle. In common carp, linseed oil and sunflower oil induced substantial LC-PUFA production (Nguyen et al. 2019). Enhanced levels of EPA and DHA were found in snow trout fed the D2 and D3 diets in the present study. The presence of palmitic, oleic, and linolenic acids in seeds influenced the fatty acid compositions in snow trout fed the seed-supplemented diets. Moreover, the higher levels of EPA and DHA in snow trout fed the vitamin C and *A. aspera* seed-supplemented diet showed that the fish has the capacity to synthesize long-chain polyunsaturated fatty acids endogenously.

A significantly higher expression of *fads2d6* was found in fish fed the vitamin C and *A. aspera* seed-supplemented diet (D3). The expression was also higher in fish fed the *A. aspera* seed-supplemented diet (D2) compared to the fish fed the vitamin C-supplemented (D1) and control diets (D4). This showed that the presence of fatty acids (Chakrabarti et al. 2012) in the seed influenced the *fads2d6* expression. A higher expression of *delta 6 desaturase* was found in rainbow trout (*Onchorynchus mykiss*) fed a plant-based diet (Ve´ron et al. 2016). The upregulation of *fads-6a* was found in common carp fed sunflower oil (Nguyen et al. 2019). There was downregulation of *fas* in fish fed all enriched diets compared to the control. The downregulation of *fas* was found in grass carp fed a vitamin C-supplemented diet (Han et al. 2019). The mRNA expression of *elovl5* was higher in fish fed the D2 and D3 diets. The *elovl2* was upregulated in snow trout fed the D1 and D2 diets. There was no effect of plant-based oil in the expression of *elovl5* in common carp (Nguyen et al. 2019).

Conclusions

The enrichment of diets with vitamin C and *A. aspera* seeds enhanced the survival rate and growth of *S. richardsonii*. The enrichment also improved the flesh composition. Vitamin C improved the amino acids composition; seeds

improved the fatty acid composition of fish. Seeds enhanced the EPA and DHA contents in fish.

Acknowledgments

The authors are thankful to the Director, DCFR, ICAR, India for providing the experimental facility.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Indian Council of Medical Research (ICMR), Govt. of India [3/1/3/JRF-2017/HRD-LS/55340/01].

ORCID

JaiGopal Sharma http://orcid.org/0000-0002-5269-4310
Rina Chakrabarti http://orcid.org/0000-0002-2235-1077

Author contribution

RC, JGS, GK, and SC designed the experiment; GK and SC conducted the experiment; GK, RKG, NK, and AKS analyzed the samples; RC, JGS, and GK prepared the manuscript; GK and NK prepared the tables and figures.

Data availability statement

The data that support the finding of this study are available within the article.

Ethical clearance

The ethical clearance for conducting this experiment was approved by the animal ethics committee of the University of Delhi. Protocol No - DU/ZOOL/IAEC-R/02/2019.

References

AOAC (Association of Official Analytical Chemist). 2000. *Official methods of analysis*. 17th ed. Gaithersburg, MD, USA: Association of Official Analytical Chemist.

APHA (American Public Health Association). 2017. *Standard methods for the examination of water and waste water*. 23rd ed. Washington DC: American Public Health Association, American Water Works Association, Water Environment Federation.

Barat, A., P. K. Sahoo, R. Kumar, C. Goel, C. Siva, and S. Ali. 2019. Data on solute carrier transporter genes of a threatened Himalayan fish. *species-Schizothorax richardsonii. Data in Brief* 23:103712. doi:10.1016/j.dib.2019.103712.

Bradford, M. M. 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry* 72 (1–2):248–54. doi:10.1016/0003-2697(76)90527-3.

Buzzi, M., R. J. Henderson, and J. R. Sargent. 1997. Biosynthesis of docosahexaenoic acid in trout hepatocytes proceeds via 24-carbon intermediates. *Comparative Biochemistry and Physiology - Part B* 116 (2):263–67. doi:10.1016/S0305-0491(96)00210-6.

Cao, M. J., K. Osatomi, M. Suzuki, K. Hara, K. Tachibana, and T. Ishihara. 2000. Purification and characterization of two anionic trypsins from the hepatopancreas of carp. *Fisheries Science* 66 (6):1172–79. doi:10.1046/j.1444-2906.2000.00185.x.

Chakrabarti, R., W. D. Clark, J. G. Sharma, R. K. Goswami, A. K. Shrivastav, and D. R. Tocher. 2018. Mass production of *Lemna minor* and its amino acid and fatty acid profiles. *Frontiers in Chemistry* 6:479. doi:10.3389/fchem.2018.00479.

Chakrabarti, R., and P. K. Srivastava. 2012. Effect of dietary supplementation with *Achyranthes aspera* seed on larval rohu labeo rohita challenged with *Aeromonas hydrophila. Journal of Aquatic Animal Health* 24 (4):213–18. doi:10.1080/08997659.2012.694834.

Chakrabarti, R., P. K. Srivastava, K. Kundu, R. S. Khare, and S. Banerjee. 2012. Evaluation of immunostimulatory and growth promoting effect of seed fractions of *Achyranthes aspera* in common carp *Cyprinus carpio* and identification of active constituents. *Fish & Shellfish Immunology* 32 (5):839–43. doi:10.1016/j.fsi.2012.02.006.

Chakrabarti, R., P. K. Srivastava, N. Verma, and J. Sharma. 2014. Effect of seeds of *Achyranthes aspera* on the immune responses and expression of some immune-related genes in carp *Catla catla*. *Fish & Shellfish Immunology* 41 (1):64–69. doi:10.1016/j.fsi.2014.03.007.

Christie, W. W. 2003. *Lipid analysis*. 3rd ed. Bridgewater, UK: Oily Press.

Cook, H. W. 1996. Fatty acid desaturation and chain elongation in eukaryote. In *Biochemistry of lipids, lipoproteins and membranes*, ed. D. E. Vance and J. E. Vance, 129–52. Amsterdam, Netherlands: Elsevier.

Dong, G. F., Q. Zou, H. Wang, F. Huang, X. C. Liu, L. Chen, C. Y. Yang, and Y. O. Yang. 2014. Conjugated linoleic acid differentially modulates growth, tissue lipid deposition, and gene expression involved in the lipid metabolism of grass carp. *Aquaculture* 432:181–91. doi:10.1016/j.aquaculture.2014.05.008.

Folch, J., M. Lees, and G. H. Sloane-Stanley. 1957. *Journal of Biological Chemistry* 226 (1):497–509. doi:10.1016/S0021-9258(18)64849-5.

Glencross, B. D. 2009. Exploring the nutritional demand for essential fatty acids by aquaculture species. *Reviews in Aquaculture* 1:71–124. doi:10.1111/j.1753-5131.2009.01006.x.

Gorelick-Feldman, J., D. MacLean, N. Ilic, A. Poulev, M. A. Lila, D. Cheng, and I. Raskin. 2008. Phytoecdysteroids increase protein synthesis in skeletal muscle cells. *Journal of Agricultural and Food Chemistry* 56 (10):3532–37. doi:10.1021/jf073059z.

Goyal, B. R., R. K. Goyal, and A. A. Mehta. 2007. Plant review phyto-pharmacology of *Achyranthes aspera*: A review. *Pharmacognosy Reviews* 1:143–50.

Guillaume, J., S. Kaushik, P. Bergot, and R. Metallier. 2001. *Nutrition and feeding of fish and crustaceans*. 1st ed. Chichester: UK: Springer in Association with Praxis Publishing.

Han, Q., H. Fan, J. Peng, L. Zhou, and L. Gan. 2019. Pleiotropic function of vitamin C on fatty acids in liver and muscle of juvenile grass carp (*Ctenopharyngodon idella*). *Aquaculture* 512:734352. doi:10.1016/j.aquaculture.2019.734352.

Huang, F., M. Jiang, H. Wen, F. Wu, W. Liu, J. Tian, and H. Shao. 2016. Dietary vitamin C requirement of genetically improved farmed Tilapia, *Oreochromis niloticus. Aquaculture Research* 47 (3):689–97. doi:10.1111/are.12527.

Imanpoor, M., M. R. Imanpoor, and Z. Roohi. 2017. Effects of dietary vitamin C on skeleton abnormalities, blood biochemical factors, haematocrit, growth, survival and stress response of *Cyprinus carpio* fry. *Aquaculture International* 25 (2):793–03. doi:10.1007/s10499-016-0080-3.

Joshi, K. D. 2004. Artificial breeding and rearing of *Schizothorax richardsonii* (Gray). *The Indian Journal of Fisheries* 51:233–37.

Kamalam, B. S., J. Mahija, P. Barat, A. Pandey, M. S. Akhtar, A. Ciji, and M. Rajesh. 2019. Temperature and oxygen related ecophysiological traits of snow trout (*Schizothorax richardsonii*) are sensitive to seasonal changes in a Himalayan stream environment. *Journal of Thermal Biology* 83:22–29. doi:10.1016/j.jtherbio.2019.04.014.

Kolkovski, K., C. Czesny, Y. Yackey, M. Moreau, C. Cihla, M. Mahan, and D. Dabrowski. 2000. The effect of vitamins C and E in (n-3) highly unsaturated fatty acids-enriched *Artemia* nauplii on growth, survival, and stress resistance of fresh water walleye *Stizostedion vitreum* larvae. *Aquaculture Nutrition* 6 (3):199–206. doi:10.1046/j.1365-2095.2000.00112.x.

Kord, M. I., T. M. Srour, E. A. Omar, A. A. Farag, A. A. M. Nour, and H. S. Khalil. 2021. The immunostimulatory effects of commercial feed additives on growth performance, non-specific immune response, antioxidants assay, and intestinal morphometry of Nile tilapia *Oreochromis niloticus*. *Frontiers in Physiology* 12:627499. doi:10.3389/fphys.2021.627499.

Kumar, N., J. G. Sharma, S. P. Singh, A. Singh, V. H. Krishna, and R. Chakrabarti. 2019. Validation of growth enhancing, immunostimulatory and disease resistance properties of *Achyranthes aspera* in *Labeo rohita* fry in pond conditions. *Heliyon* 5 (2):e01246. doi:10.1016/j.heliyon.2019.e01246.

Lall, S. P., and G. Olivier. 1993. Role of micronutrients in immune response and disease resistance of fish. In *Fish nutrition in practice*, ed. S. J. Kaushik and P. Luquet, 101–18. Paris, France: National Institute for Agricultural Research.

Liu, H. P., B. Wen, Z. Z. Chen, J. Z. Gao, Y. Liu, Y. C. Zhang, and Y. Peng. 2018. Effects of dietary vitamin C and vitamin E on the growth, antioxidant defence and digestive enzyme activities of juvenile discus fish (*Symphysodon haraldi*). *Aquaculture Nutrition* 25 (1):176–83. doi:10.1111/anu.12841.

Liu, Y., L. Chi, L. Feng, J. Jiang, W. D. Jiang, K. Hu, and X. Q. Zhou. 2011. Effects of graded levels of dietary vitamin C on the growth, digestive capacity and intestinal microflora of juvenile Jian carp (*Cyprinus carpio* var. Jian). *Aquaculture Research* 42 (4):534–48. doi:10.1111/j.1365-2109.2010.02649.x.

Livak, K. J., and T. D. Schmittgen. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* 25 (4):402–08. doi:10.1006/meth.2001.1262.

Misra, C. K., B. K. Das, S. C. Mukherjee, and J. Pradhan. 2007. Effects of dietary vitamin C on immunity, growth and survival of Indian major carp *Labeo rohita*, fingerlings. *Aquaculture Nutrition* 13 (1):35–44. doi:10.1111/j.1365-2095.2007.00451.x.

Montgomery, D. C. 1984. *Design and analysis of experiments*. New York, USA: John Wiley.

Moses, N. R., K. J. Nilssen, and R. Chakrabarti. 2017. Effect of dietary supplementation of vitamin C and seeds of *Achyranthes aspera* on growth, digestive enzyme activities, immune system and lipid peroxidation of snow trout *Schizothorax richardsonii*. *Madridge Journal of Aquaculture Research and Development* 1 (1):24–30. doi:10.18689/mjard-1000105.

Nguyen, T. M., S. N. Mandiki, T. N. T. Tran, Y. Larondelle, J. Mellery, E. Mignolet, and P. Kestemont. 2019. Growth performance and immune status in common carp *Cyprinus carpio* as affected by plant oil-based diets complemented with β-glucan. *Fish & Shellfish Immunology* 92:288–99. doi:10.1016/j.fsi.2019.06.011.

Nsonga, A. R., J. Kang'Ombe, W. Mfitilodze, C. K. Soko, and A. H. Mtethiwa. 2009. Effect of varying levels of dietary vitamin C (ascorbic acid) on growth, survival and hematology of juvenile tilapia, *Oreochromis karongae* (Trewavas 1941) reared in aquaria. *Brazilian Journal of Aquatic Science and Technology* 13 (2):17–23. doi:10.14210/bjast.v13n2.p17-23.

Petr, T. 2002. Cold water fish and fisheries in countries of the high mountain arc of Asia (Hindu Kush-Pamir-Karakoram-Himalayas). A review. In *Cold water fisheries in the trans-himalayan countries*, FAO Fisheries Technical Paper No. 431, ed. T. Petr and D. B. Swar, 1–38. Rome, Italy: Food and Agriculture Organization of the United Nations.

Rao, Y. V., and R. Chakrabarti. 2005. Dietary incorporation of *Achyranthes aspera* seed influences the immunity of common carp. *Cyprinus carpio. Indian Journal of Animal Sciences* 75:1097–102.

Rao, Y. V., M. Romesh, A. Singh, and R. Chakrabarti. 2004. Potentiation of antibody production in Indian major carp *Labeo rohita*, rohu, by *Achyranthes aspera* as a herbal feed ingredient. *Aquaculture* 238 (1–4):67–73. doi:10.1016/j.aquaculture.2004.04.029.

Roberts, I. M. 1985. Hydrolysis of 4-methylumbelliferyl butyrate: A convenient and sensitive fluorescent assay for lipase activity. *Lipids* 20 (4):243–47. doi:10.1007/BF02534195.

Sarma, D., M. S. Akhtar, P. C. Mahanta, and D. Debnath. 2013. Nutritional quality in terms of amino acid and fatty acid of five coldwater fish species: Implications to human health. *National Academy Science Letters* 36 (4):385–91. doi:10.1007/s40009-013-0151-1.

Sato, P., M. Nishikimi, and S. Udenfriend. 1976. Is gulonolactone-oxidase the only enzyme missing in animals subject to scurvy? *Biochemical and Biophysical Research Communications* 71 (1):293–99. doi:10.1016/0006-291X(76)90281-3.

Sharma, J. G., N. Kumar, S. P. Singh, A. Singh, V. HariKrishna, and R. Chakrabarti. 2019. Evaluation of immunostimulatory properties of prickly chaff flower *Achyranthes aspera* in *rohu Labeo rohita* fry in pond conditions. *Aquaculture* 505:183–89. doi:10.1016/j.aquaculture.2019.02.065.

Singh, A., J. G. Sharma, M. Paichha, and R. Chakrabarti. 2019. *Achyranthes aspera* (Prickly chaff flower) leaves- and seeds-supplemented diets regulate growth, innate immunity, and oxidative stress in *Aeromonas hydrophila* -challenged *Labeo rohita. Journal of Applied Aquaculture* 32:250–67. doi:10.1080/10454438.2019.1615594.

Singh, M. K., J. G. Sharma, and R. Chakrabarti. 2013a. Effect of UV-B radiation on the defence system of *Labeo rohita* (Actinopterygii: Cypriniformes: Cyprinidae) Larvae and its modulation by seed of Devil's Horsewhip, *Achyranthes aspera. Acta Ichthyologica et Piscatoria* 43 (2):119–26. doi:10.3750/AIP2013.43.2.04.

Singh, M. K., J. G. Sharma, and R. Chakrabarti. 2013b. Impact of UV-B radiation on the physiology of freshwater carp *Labeo rohita* larvae and evaluation of UV-B protective properties of seeds of *Achyranthes aspera* and vitamin C. *Agricultural Research* 2 (2):166–71. doi:10.1007/s40003-013-0060-z.

Tocher, D. R., M. Agaba, N. Hastings, J. G. Bell, J. R. Dick, and A. J. Teale. 2001. Nutritional regulation of hepatocyte fatty acid desaturation and polyunsaturated fatty acid composition in zebrafish (*Danio rerio*) and tilapia (*Oreochromis niloticus*). *Fish Physiology and Biochemistry* 24 (4):309–20. doi:10.1023/A:1015022406790.

Tocher, D. R., K. Dabrowski, and R. Hardy. 2010. Fatty acid requirements in ontogeny of marine and freshwater fish. *Aquaculture Research* 41 (5):717–32. doi:10.1111/j.1365-2109.2008.02150.x.

Tocher, D. R., and D. G. Harvie. 1988. Fatty acid compositions of the major phosphoglycerides from fish neural tissues; (n–3) and (n–6) polyunsaturated fatty acids in rainbow trout (*Salmo gairdneri*) and cod (*Gadus morhua*) brains and retinas. *Fish Physiology and Biochemistry* 5 (4):229–39. doi:10.1007/BF01874800.

Torstensen, B. E., and D. R. Tocher. 2010. The effects of fish oil replacement on lipid metabolism of fish. In *Fish oil replacement and alternative lipid sources in aquaculture feeds*, ed. G. M. Turchini, W. K. Ng, and D. R. Tocher, 405–37. Boca Raton, Florida: Taylor & Francis (CRC Press).

Trattner, S., J. Pickova, K. H. Park, J. Rinchard, and K. Dabrowski. 2007. Effects of α-lipoic and ascorbic acid on the muscle and brain fatty acids and antioxidant profile of the South American pacu *Piaractus mesopotamicus*. *Aquaculture* 273 (1):158–64. doi:10.1016/0006-291X(76)90281-3.

Trenzado, C. E., A. E. Morales, and M. de la Higuera. 2008. Physiological changes in rainbow trout held under crowded conditions and fed diets with different levels of vitamins E and C and highly unsaturated fatty acids (HUFA). *Aquaculture* 277 (3–4):293–302. doi:10.1016/j.aquaculture.2008.02.031.

Ueberschär, B. 1988. Determination of the nutritional condition of individual marine fish larvae by analysing their proteolytic enzyme activities with a highly sensitive fluorescence technique. *Meeresforschung-Reports on Marine Research* 32:144–54.

Véron, V., S. Panserat, R. Le Boucher, L. Labbé, E. Quillet, M. Dupont-Nivet, and F. Médale. 2016. Long-term feeding a plant-based diet devoid of marine ingredients strongly affects certain key metabolic enzymes in the rainbow trout liver. *Fish Physiology and Biochemistry* 42 (2):771–85. doi:10.1007/s10695-015-0174-2.

Zou, W., Z. Lin, Y. Huang, S. M. Limbu, H. Rong, C. Yu, and X. Wen. 2019. Effect of dietary vitamin C on growth performance, body composition and biochemical parameters of juvenile Chu's croaker (*Nibea coibor*). *Aquaculture Nutrition* 26 (1):60–73. doi:10.1111/anu.12967.

Thermal-Electrochemical Modeling and Analysis of Different Cathode-Anode Pairs for Lithium-ion Battery

Snigdha SHARMA[1], Amrish K. PANWAR[2], Madan Mohan TRIPATHI[1]
[1]*Department of Electrical Engineering, Delhi Technological University, New Delhi, India*
[2]*Department of Applied Physics, Delhi Technological University, New Delhi, India*
mmtripathi@dce.ac.in

***Abstract*—This article represents a computational approach for the estimation of the characteristics of lithium-ion batteries for a 2D electrochemical model of cylindrical type lithium-ion battery using COMSOL software. Various input parameters used in the model formulation have been obtained from experimental data as well as published work. For the evaluation of battery performance, an electrochemical-thermal mathematical model has been designed to determine the electrochemical parameters such as rate capability, charge/ discharge characteristic, ragone curve. Further, the electrochemical characteristics have been characterized with the help of cyclic voltammetry (CV) test which indicates the uniform electrochemical reactivity and transport properties of the designed battery. Two types of electrode pairs in which $LiMn_2O_4$ as cathode materials form $LiMn_2O_4$ (LMO)/Graphite and LMO/Carbon (MCMB) are designed and these electrode pairs are simulated and evaluated to determine the characteristics of battery material using charge and discharge curve at different current loads (C-rates) as well as ragone curve. The simulation results validate the electrochemical behavior of the proposed 2D model. Also, the LMO/graphite electrode pair shows better discharging and charging capacity, high energy, and power curve as compared to the LMO/Carbon (MCMB) electrode pair.**

***Index Terms*—anode, batteries, cathode, electrochemical, electrodes.**

I. INTRODUCTION

Nowadays, Electrochemical Energy Storage (EES) devices are becoming important due to their good life cycle, high energy density, reliability, stability, etc. [1-5].

Rechargeable batteries are the most useful and popular EES device as it gives the best comprehensive performance in the commercialized segments such as portable electronic devices, electric vehicles for transportation, communications and many more [5-7].

Among different types of rechargeable batteries, Lithium-ion batteries (LIBs) give the best comprehensive performance but, due to the physical limitation of the electrode material still, the application of these batteries is limited [7-8]. The ionic and electronic conduction, as well as transport phenomenon of the LIBs, are the major focus area of research as the ionic and electronic conduction of LIBs are strongly related to the rate capability of the battery [9].

This work was supported by LIBT Lab, Department of Applied Physics, Delhi Technological University, New Delhi to carry out this research work.

A. Li-ion Transport Mechanism

In LIBs, both cathode and anode electrode materials together reveal the ionic and electronic conduction properties of the battery [10]. It is very difficult to improve the ionic conduction of the material because it is strongly related to the ionic transport species that directly represents the composition and structure of the material [11-12]. To improve these properties, certain technical methods are used which help to enhance the intrinsic electronic conductivity of the material through adding a conduction additive or carbon coating, etc. [13]. There are various factors related to ionic transport such as electric field gradient, concentration gradient, and chemical potential gradient [14]. The ionic transport mechanism is generally associated with macroscopic diffusion or related to the movement of Li ions when the external electric potential was applied to the battery. This diffusion coefficient relationship can be explained with the help of the Nernst-Einstein equation 1.

$$\frac{\sigma_i}{D_i} = \frac{F^2 C_i Z_i^2}{RT} \tag{1}$$

where σ_i is the ionic conductivity of species i, D_i is the chemical diffusion coefficient of species i, Z_i is the charge of species i, F is faraday's constant, T is the absolute temperature, R is the gas constant [1].

Further, the mobility of ions u_i is linked to the diffusion coefficient of ions D_i as shown in equation 2.

$$D_i = Tu_i k_b \tag{2}$$

where k_b is the Boltzmann's constant and T is the temperature in Kelvin [10].

The diffusion in solids takes place through the movement of ions when an adjacent site of the atom is empty or displacement of atoms causing lattice distortion and have sufficient energy to break bonds with its neighbor atoms [10-15]. But in the case of material used in LIB, the diffusion mechanism is more complex which is analyzed by various factors such as the motion of atoms, thermal energy, vacancy at the adjacent site, sufficient energy to break the bond, and many more [10], [15-16]. Therefore, the electronic conduction mechanism of LIBs becomes the focus area of concern.

B. Electronic Conduction Mechanism

Electron transport in the solid electrode is one of the most important factors in the evaluation of the rate performance of the battery [12] [14]. In LIBs, when the charging and

Digital Object Identifier 10.4316/AECE.2021.03007

1582-7445 © 2021 AECE

2021 8th International Conference on Signal Processing and Integrated Networks (SPIN) | 978-1-6654-3564-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SPIN52536.2021.9565994

Threshold Type Binary Memristor Emulator based on DVCC

Juli Kumari Roy
Department of Electronics & Communication Engineering
Delhi Technological University
Delhi,India
julikumariroy@gmail.com

Neeta Pandey
Department of Electronics & Communication Engineering
Delhi Technological University
Delhi, India
neetapandey@dce.ac.in

***Abstract*— This paper proposes a Threshold Type Binary Memristor Emulator based on Differential Voltage Current Conveyor (DVCC). Additionally, it uses an analog multiplier, two diodes, nine grounded resistors and one grounded capacitor. This threshold sensitive behavior is embedded through anti parallel configuration of diode. Further, threshold voltage is adjusted by resistor ratio. The emulator uses an integrator which ensures the dependence of memductance on history state. The non-volatility and bistability characteristics of the memristor emulator are provided by the bistable circuit. The workability of the proposed emulator circuit is verified through PSPICE simulations.**

Keywords*— DVCC, Memristor, Diode, Bistable circuit, Threshold voltage

I. INTRODUCTION

Memristor, the fourth passive element of electrical circuits, was first postulated by Chua in 1971[1]. It has characteristics which shows relation between charge and flux. The wide range of memristor applications namely combinational and sequential circuits [2], memory [3], neuro network [4], chaotic systems [5] and analog circuits [6] has attracted lots of research interest. The memristor is not commercially available due to difficulty in its fabrication at nanoscale. Therefore, analog active blocks [7-9], are used to emulate its behaviour. The second-generation current conveyor (CCII) [10], a multiplier and op-amp [11] have been used to demonstrate floating-type memristor functionality. Electronically adjustable memristor emulator comprised of CCII and operational transconductance amplifier (OTA) have a relatively restricted frequency response [12]. A differential voltage current conveyor transconductance (DVCCTA) is used to develop a charged controlled memristor emulator in [13]. Recently there is an interest in binary memristor design. The available designs are based on low pass filter [14] and thresholding concept [15]. Thus, there is limited literature available for threshold type memristor emulator. This paper presents a differential voltage current conveyor (DVCC) based binary memristor emulator. The proposed threshold type memristor emulator has the ability of threshold type switching which ensure the basic logic operations essential for non-volatile storage property, on which all its applications are based.

The contents of the paper are divided into six sections. Section II and III provide a quick overview about practical model and basic of characteristics DVCC respectively. Section IV describes the proposed memristor emulator circuit. The functional verification is done in simulation section V. The paper is concluded in section VI.

II. PRACTICAL MODEL

The behaviour of threshold type memristor [15] is briefly reviewed in this section. The current i through voltage controlled memristor is product of memconductance G and voltage v where G is function of state variable x, voltage v and time (t), and is expressed by (1).

$$i = G(x, v, t)v \tag{1}$$

Additionally, the following equation for derivative of state variable x also needs be satisfied.

$$\frac{dx}{dt} = f(x, v, t) \tag{2}$$

The threshold phenomenon [16] is modelled by (3)

$$\frac{dx}{dt} = \beta v + 0.5(\alpha-\beta)\,[|v + v_{th}| - |v - v_{th}|] \tag{3}$$

Here coefficients α and β characterize the rate of change in state variable and v_{th} corresponds to threshold voltage. Further, the binary memristor has low resistance and high resistance states which are abbreviated as LRS and HRS. Assuming threshold voltages as Y1 and Y2 the value of resistance is LRS if state variable is less than Y1 and HRS if state variable is greater than Y2. In case the state variable lies in between Y1 and Y2, the memristor holds the state.

III. THE DVCC

The Differential Voltage Current Conveyor (DVCC) is capable of processing difference of voltages and also possesses current conveying capability. The DVCC finds wide applications in analog signal processing circuits namely filters [17-18], oscillator [19-20], multivibrator [21] etc.

The symbolic representation of DVCC is depicted in Fig. 1. It has four terminals namely Y_1, Y_2, X and Z. Its principle of operation can be explained by the set of characteristics equations as given in (4)

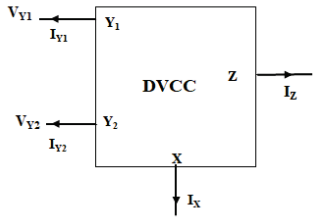


Fig. 1. Symbolic diagram of DVCC [22]

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:44:17 UTC from IEEE Xplore. Restrictions apply.

I_{Y1}=0

I_{Y2}=0

$V_X = V_{Y1} - V_{Y2}$

$I_Z = I_X$ (4)

From equations (4), it can be concluded that the difference of voltages applied at Y_1 & Y_2 terminals is conveyed to voltage at X terminal while the injected current at X terminal is the conveyed to terminal Z. Since input terminals Y_1 & Y_2, have high input impedances, hence no current flows through these terminals. The Fig. 2 depicts the realization of DVCC based on IC AD844.

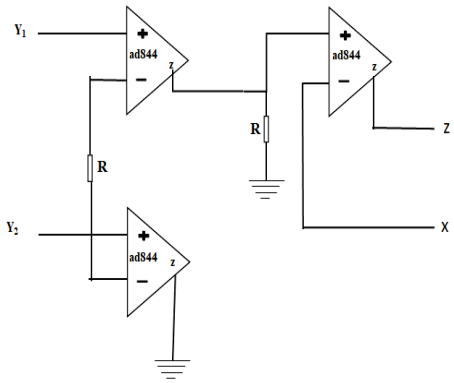
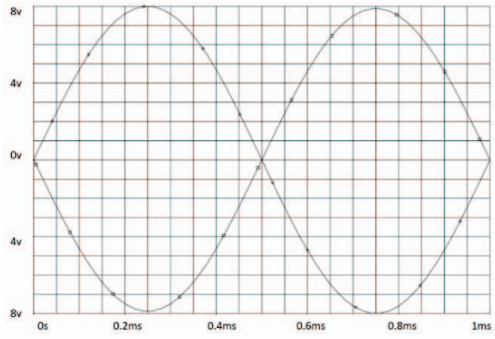
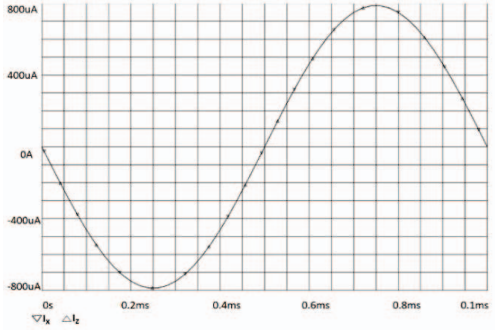


Fig.2. DVCC Realization based on IC AD844 [22]

The operation of the DVCC is examined through PSPICE simulations. The power supply of V_{DD} = - V_{SS}=10V and R =10kΩ are used. An input Voltage V= 8 sin(2000πt) is applied at Y_2 terminal and Y_1 is connected to ground in the circuit shown in Fig. 2. The output was taken at the Z and X terminal of DVCC. The simulation waveforms are depicted in Fig. 3 which are in tune with basic characteristics of DVCC.



(a)



(b)

Fig. 3 Time domain waveform a) V_X and $V_{Y2}$and b) I_Z and I_X.

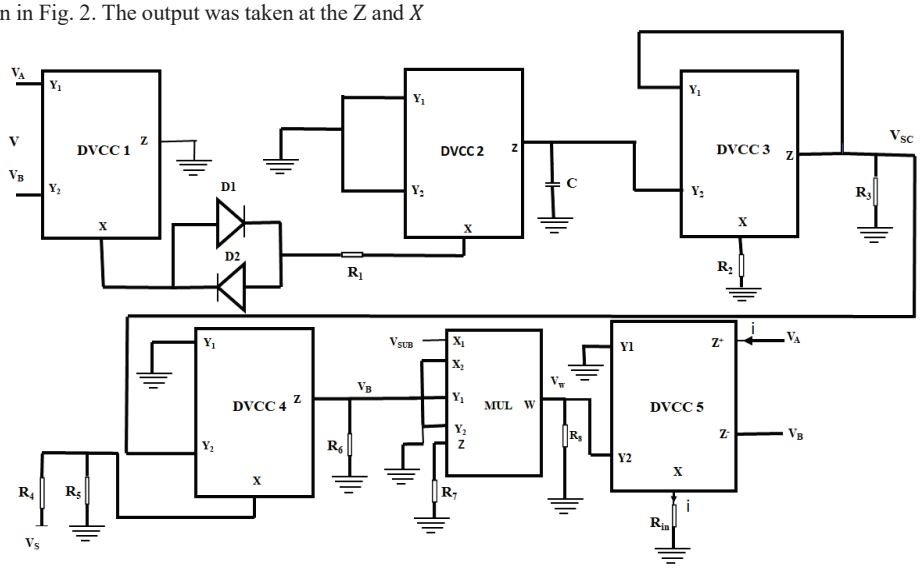


Fig.4. DVCC based proposed threshold binary type memristor emulator

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:44:17 UTC from IEEE Xplore. Restrictions apply.

IV. PROPOSED MEMRISTOR EMULATOR

The proposed floating emulator circuits consists of subtractor, integrator, bistable circuit, inverting summing amplifier, multiplier and voltage into current to ensures floating characteristic of the emulator. The subtractor senses the memristor voltage, therefore a voltage subtractor circuit is required DVCC 1 serve this purpose the output V_{SUB} is given by (5)

$$V_{SUB} = (V_A - V_B)$$
$$V = V_A - V_B \quad (5)$$

The diode D1 and D2 in anti-parallel configuration provide threshold sensitive behaviour of the emulator which may be adjusted by resistor ratio. Assuming cut in voltage of both the diodes as v_{th}, in equation (6).

$$V_d' = V - 0.5[|V + v_{th}| - |V - v_{th}|] \quad (6)$$

It may be noted that for input voltage ≤ |vth| the output is zero. The threshold voltage remains same as diode cut in voltage, however, by introducing block with gain (G_1) before integrator formed by DVCC2. In presence of gain, (6) modifies to

$$V_d' = V - 0.5[|V + G_1 v_{th}| - |V - G_1 v_{th}|] \quad (7)$$

It may be noted that the threshold voltage increases when G_1 is less than unity while a decreasing trend is observed for G_1 is greater than unity.

An integrator circuit is needed next to add the dependence of memductance on history state. The DVCC 2 provides the functionality of integrator circuit.

The product of RC performs the operating frequency of memristor emulator, for higher value will produce lower frequency and a lower value results in higher frequency.

For bistable and non- volatility properties to the memristor emulator, a bistable circuit is needed. The DVCC 3 deliver the functionality of bistable circuit. Its output is given by (8)

$$\text{Vsc} = \begin{cases} \text{V}_{\text{ss}} & \text{V}_{\text{TH}} < \text{V}_{\text{int}} \\ \text{hold} & \text{V}_{\text{TL}} \le \text{V}_{\text{int}} \le \text{V}_{\text{TH}} \\ \text{V}_{\text{DD}} & \text{V}_{\text{int}} < \text{V}_{\text{TL}} \end{cases} \quad (8)$$

Other symbols in (8) have their usual meaning

The bistable circuit employing a single DVCC and two resistors which shown in Fig. 5. The connection Y_1 to Z provides the positive feedback of bistable circuit. Loop gain of the positive feedback is greater than unity.

The voltage at output node of the bistable circuit varies from high to low and low to high. When the input voltage given by equation (9).

$$V_{in} = V_{Y1} - V_X \quad (9)$$

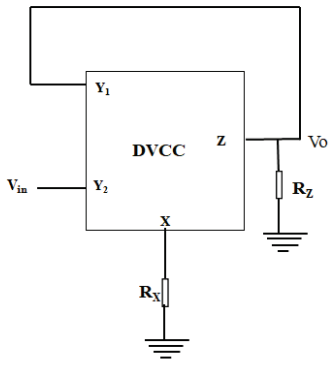


Fig. 5. DVCC based Bistable circuit [23]

Where V_{TH} and V_{TL} related to upper and lower threshold voltage of bistable circuit given by (10) and (11) respectively.

$$V_{TH} = V_{DD} - V_{XDD} \quad (10)$$
$$V_{TL} = V_{SS} - V_{XSS} \quad (11)$$

The output of bistable circuit passes through an inverting summing amplifier formed by DVCC 4 so that output voltage V_S (V_S is used for shifting the voltage level).

A multiplier is necessary to provide voltage proportional to memristor current. AD633 is used for this purpose. It multiplies signals V_{SUB} and V_S and uses resistors R_7 and R_8 to adjust multiplication coefficient.

$$V_W = R_{in}\, i \quad (12)$$

The DVCC 5 is used to convert the voltage V_W to memristor current i. The current flowing through R_{in} is given by

$$i = \frac{V_w}{R_{in}} \quad (13)$$

The proposed circuits, according to above analysis, emulates a threshold type memristor.

The proposed emulator uses 5 DVCC, two diodes, 9 resistors, one capacitor, and one multiplier whereas the existing counterpart [15] uses 2 CFOAs, two diodes 7 opamps, 17 resistors, one capacitor and one multiplier.

V. SIMULATION RESULTS

OrCAD software version 16.6 is to examined the proposed design. The power supply of V_{DD} = - V_{SS}=10V is taken. An input Voltage V= 10 sin(2000πt) is applied to the circuit shown in Fig. 4.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:44:17 UTC from IEEE Xplore. Restrictions apply.

Fig. 6 shows that the transient waveforms between input and output voltage of bistable circuit. R_X =10 kΩ and R_Z = 100 kΩ are used as per shown in the Fig. 5.

TABLE I COMPONENT SETTING

component	*value*
$R_1 = R_2 = R_4 = R_5 = R_6 = R_7 = R_8 = R_{in}$	10 kΩ
C	10nF
R_3	100 kΩ

The functionality of complete circuit is verified for G_1=1 and V_S =2.1V. The component settings are listed in Table I and for diode D1 and D2 D1N4001 is used.

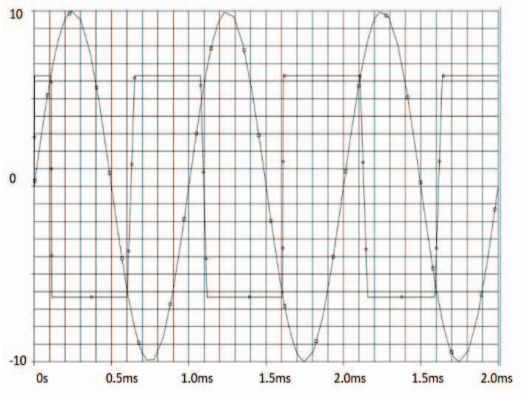
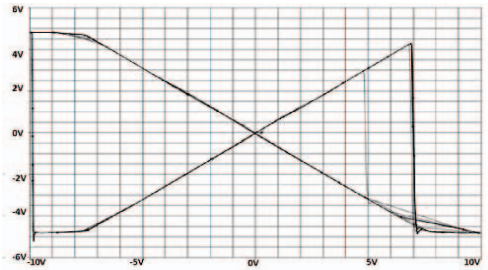


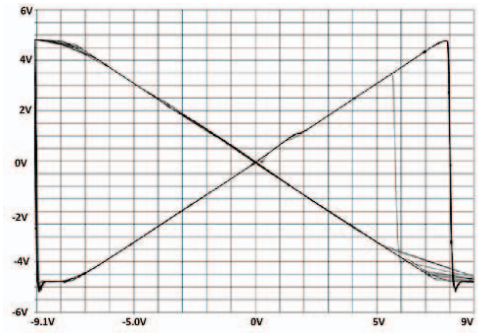
Fig.6. Transient waveforms of bistable circuit

The plot of *V* versus V_W are shown in Fig. 7 for G_1=1 and while keeping another component setting same as given in Table I. The plots correspond to pinched hysteresis loop confirming the functionality of memristor.

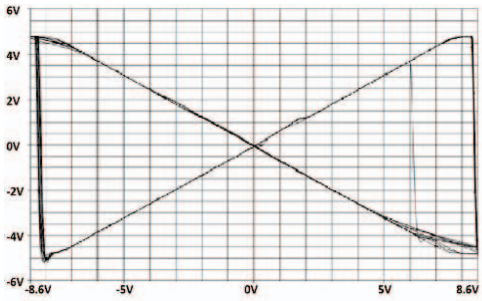
simulations are done to examine the frequency-dependent pinched hysteresis loop of the proposed design for different frequencies. The input voltage (V) versus V_W characteristics of the proposed emulator at different input frequencies, i.e.,7kHz, 9kHz, 10kHz and 15kHz are observed and are shown in Fig. 7. It is noted that the area of the hysteresis lobe is dependent on the frequency of the input signal. The enclosed area of the pinched hysteresis loop decreases monotonically with the increase in frequency of the input.



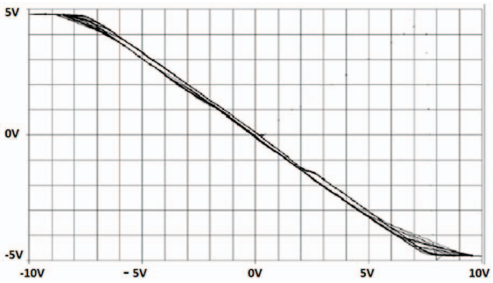
(a)



(b)



(c)



(d)

Fig.7. V versus V_W characteristics of the emulator (a) at f=7kHz (b) at f=9kHz (C)at f=10kHz (d) at f= 15kHz

The power is noted to be 2.08W. Power supply is 10% variation is study proposed circuit. It found that ± 0.2 W change in power consumption.

In fig. 7, it is concluded that the enclosed area of the Pinched hysteresis loop decreases as frequency increases.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:44:17 UTC from IEEE Xplore. Restrictions apply.

VI. CONCLUSION

A DVCC based Threshold Type Memristor Emulator is proposed in this paper. Lesser number of resistors are used in the proposed memristor. Due to its adjustable threshold characteristics, it can be used in applications based on logical operations. The components and features of the proposed memristor can help us for further study, simulations and designs in this particular area.

REFERENCES

[1] L. O. Chua, "Memristor-the missing circuit element," IEEE Trans. Circuit Theory, vol. CT-18, no. 5, pp. 507–519, Sep. 1971.

[2] I. Vourkas and G. C. Sirakoulis, "Emerging memristor-based logic circuit design approaches: a review," IEEE Circuits and Systems Magazine, vol. 16, no. 3, pp. 15-30, Aug. 2016.

[3] J. Y. Seok, S. J. Song, J. H. Yoon, et al., "A review of three_dimensional resistive switching cross-bar array memories from the integration and materials property points of view," Advanced Functional Materials, vol. 24, pp. 5316-5339, 2014.

[4] W. Yi, K. K. Tsang, S. K. Lam, et al., "Biological plausibility and stochasticity in scalable VO2 active memristor neurons," Nature Communications, vol. 9, pp. 4661, Nov. 2018.

[5] J. Sun, Y. Wu, G. Cui, et al., "Finite-time real combination synchronization of three complex-variable chaotic systems with unknown parameters via sliding mode control," Nonlinear Dynamics, vol. 88, no. 3, pp. 1677-1690, 2017.

[6] P. Mazumder, S. M. Kang, and R. Waser, "Memristors: devices, models, and applications," Proceedings of the IEEE, vol. 100, no. 6, pp. 1911-1919, June 2012.

[7] Singh D, Pandey N. Realization of Single CCCDTA based incremental/decremental type Memconductance Emulator. In2019 International Symposium on Advanced Electrical and Communication Technologies (ISAECT) 2019 Nov 27 (pp. 1-4).

[8] Ayten UE, Minaei S, Sağbaş M. Memristor emulator circuits using single CBTA. AEU-International Journal of Electronics and Communications. 2017 Dec 1;82:109-18.

[9] Ranjan RK, Rani N, Pal R, Paul SK, Kanyal G. Single CCTA based high frequency floating and grounded type of incremental/decremental memristor emulator and its application. Microelectronics Journal. 2017 Feb 1;60:119-28.

[10] D. Yu, H. H.-C. Iu, A. L. Fitch, and Y. Liang, "A floating memristor emulator based relaxation oscillator," IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 61, no. 10, pp. 2888–2896, Oct. 2014.

[11] H. Kim, M. P. Sah, C. Yang, S. Cho, and L. O. Chua, "Memristor emulator for memristor circuit applications," IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 59, no. 10, pp. 2422–2431, Oct. 2012.

[12] H. Sözen and U. Çam, "Electronically tunable memristor emulator circuit," Analog Integr. Circuits Signal Process., vol. 89, no. 3, pp. 655–663, Dec. 2016.

[13] Raj N, Ranjan RK, Khateb F. Flux-controlled memristor emulator and its experimental results. IEEE Transactions on Very Large Scale Integration (VLSI) Systems. 2020 Jan 30;28(4):1050-61.

[14] Li D, Zhang J, Yu D, Xu R, Iu HH, Fernando T, Wang X. A Family of Binary Memristor-Based Low-Pass Filters With Controllable Cut-Off Frequency. IEEE Access. 2020 Mar 24;8:60199-209.

[15] Liang Y, Lu Z, Wang G, Yu D, Iu HH. Threshold-Type Binary Memristor Emulator Circuit. IEEE Access. 2019 Dec 3;7:180181-93.

[16] I. Vourkas, A. Batsos, G. C. Sirakoulis, "SPICE modeling of nonlinear memristive behavior," International Journal of Circuit Theory and Applications, vol. 43, no. 5, pp. 553-565, May 2015.

[17] S. Minaei, E. Yuce, Novel voltage-mode all-pass filter based on using DVCCs. Circuits Syst. Signal Process. 29(3), 391–402 (2010).

[18] Mishra SK, Gupta M, Upadhyay DK. Active realization of fractional order Butterworth lowpass filter using DVCC. Journal of King Saud University-Engineering Sciences. 2020 Feb 1;32(2):158-65.

[19] Daoud H, Bensalem S, Lahiani S, Bensalem C, Loulou M. NanoCMOS optimized DVCC-based quadrature voltage controlled oscillator performances prediction through bisquare-weights method. Analog Integrated Circuits and Signal Processing. 2019 Sep;100(3):547-63.

[20] Yesil A, Babacan Y, Kacar F. An electronically controllable, fully floating memristor based on active elements: DO-OTA and DVCC. AEU-International Journal of Electronics and Communications. 2020 Aug 1;123:153315.

[21] Chien, Y.-K. Lo, Design and implementation of monostable multivibrators employing differential voltage current conveyors. Microelectron. J. 42(10), 1107–1115 (2011).

[22] Vista J, Pheiroijam NM, Pamu H, Tarunkumar H, Ranjan A. A Novel Differential Voltage Current Conveyor (DVCC) Based Instrumentation Amplifier. IETE Journal of Research. 2019 Apr 13:1-2.

[23] Minaei S, Yuce E. A simple Schmitt trigger circuit with grounded passive elements and its application to square/triangular wave generator. Circuits, Systems, and Signal Processing. 2012 Jun;31(3):877-88.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:44:17 UTC from IEEE Xplore. Restrictions apply.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/355425839

TWO-WAY METRIC LEARNING WITH MAJORITY AND MINORITY SUBSETS FOR CLASSIFICATION OF LARGE EXTREMELY IMBALANCED FACE DATASET

Article *in* Jordanian Journal of Computers and Information Technology (JJCIT) · October 2021

DOI: 10.5455/jjcit.71-1626417940

CITATIONS
0

READS
54

2 authors:



Ashu Kaushik
Delhi Technological University
3 PUBLICATIONS **3** CITATIONS

SEE PROFILE



Seba Susan
Delhi Technological University
104 PUBLICATIONS **719** CITATIONS

SEE PROFILE

All content following this page was uploaded by Seba Susan on 20 October 2021.

The user has requested enhancement of the downloaded file.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 07, No. 04, December 2021.

TWO-WAY METRIC LEARNING WITH MAJORITY AND MINORITY SUBSETS FOR CLASSIFICATION OF LARGE EXTREMELY IMBALANCED FACE DATASET

Ashu Kaushik[1] and Seba Susan[2]

(Received: 16-Jul.-2021, Revised: 12-Sep.-2021, Accepted: 27-Sep.-2021)

ABSTRACT

This paper proposes a new learning methodology involving deep features and two-way metric learning for large, extremely imbalanced face datasets where the number of minority classes and the imbalance ratio are both very high. The problem arises because the faces of some celebrities, being more popular, are readily available in social media and the internet, while the faces of some relatively lesser-known personalities are fewer in number. Resampling being impractical in this scenario, we propose metric learning as the tool for mitigating the class-imbalance problem prior to the classification stage. To reduce the computational overhead associated with metric learning, we separately conduct weakly supervized metric learning with majority and minority class subsets, a process that we call two-way metric learning. Transformation matrices learnt from the majority and minority subsets are used to transform the entire input space twice. The test sample in the transformed space is assigned the class of its nearest neighbor in the training set of the twice-transformed input space. Deep features derived from the state-of-the-art pre-trained deep network VGG-Face form the input space and the aggregate cosine similarity measure is used to find the closest neighbor in the training set of the twice-transformed input space. Experiments on the benchmark LFW face database having 1680 classes of celebrity faces prove that the proposed methodology is more effective than existing methods for the classification of large, extremely imbalanced face datasets. The classification accuracies of the minority classes are especially found to be boosted which is a rare accomplishment among existing methods for imbalanced learning in deep frameworks.

KEYWORDS

Face recognition, Metric learning, VGG-Face, Deep learning, Imbalanced learning, Extremely imbalanced dataset.

1. INTRODUCTION

In this computer and mobile frenzy era, almost everything is getting digitized. The large amount of data that is being generated by the use of such digital devices is creating a havoc and needs to be analyzed, sorted and stored properly and judiciously. Social media platforms, like Facebook, Instagram, Twitter and Zoom, create a large amount of data and logs based upon the usage of the account holder. Face recognition plays a crucial role in detecting and tagging the identity of a person in social media. The users who are very active on social media have a large amount of data associated with them, including images that reveal the identity of the individual. Such users constitute the majority class in the learning framework. On the other hand, users who are less active contribute to lesser data and fewer images that make automated face recognition a difficult task; such users constitute the minority class. Face recognition from such imbalanced datasets, where the difference between the volume of data between the majority and minority classes is very high, is indeed a difficult task [1]-[2]. Learning from imbalanced data is a well-researched problem in data mining [3]-[4] with various solutions proposed ranging from resampling [5] and metric learning [6] to cost-sensitive learning [7]. Selective pruning of majority and minority samples is found helpful, especially when some amount of overlapping is there between the majority and minority classes [8]. Most of the proposed solutions are effective for the binary classification problem, but classification in the multi-class scenario with a highly imbalanced class distribution is still an open research problem [9]. Resampling techniques might work on small toy datasets popular in data mining, but while dealing with a very large dataset comprising of faces derived from social media, consisting of more than a

1. A. Kaushik is a Postgraduate Student in the Department of Information Technology, Delhi Technological University, Delhi, India. Email: ashukaushik8395@gmail.com
2. S. Susan is a Professor in the Department of Information Technology, Delhi Technological University, Delhi, India. Email: seba_406@yahoo.in, ORCID:0000-0002-6709-6591.

thousand classes and with a very high class-imbalance ratio (ratio of majority to minority population), it is not considered a feasible solution [10].

Novel learning methodologies need to be devised for extremely imbalanced large face databases in order to meet the computational overhead and at the same time improve the classification accuracy, especially for the minority classes having inadequate number of samples to learn from. This is the problem tackled in this paper and we choose metric learning [11] as the tool for transforming the entire input space in order to reduce intra-class differences and increase the inter-class differences. This is achieved by identifying two smaller subsets of the large imbalanced face dataset as the majority and minority classes and performing metric learning using these two subsets. Metric learning would be done in a weakly supervized fashion for both majority and minority subsets to learn the distance metric which can be used to transform the entire input space prior to the classification stage. The contributions made by this paper can be summarized as follows:

1. Metric learning with deep features is introduced as a viable tool for large extremely imbalanced facial datasets having more than a thousand minority classes, for which resampling is not feasible.

2. To reduce the computational overhead associated with metric learning, a weakly supervized learning scheme is devised, for which smaller-sized majority and minority class subsets are identified.

3. The entire input space is then transformed twice, once using the transformation matrix learnt from the majority class subset and likewise from the minority class subset.

4. The aggregate cosine similarity measure is eventually used for the classification of the transformed test sample by finding its closest neighbor in the training set of the twice-transformed input space.

5. Experiments on the large, extremely imbalanced LFW face database having 1680 classes, with large disparity in class populations, yield effective classification, especially for the minority classes, a rare accomplishment among existing methods for imbalanced learning in deep frameworks. The methods proposed so far mostly concentrate on the performance of majority classes only and exclude the minority classes in the learning process.

The further sections of the paper are organized as follows. Section 2 describes the motivation for our work and the proposed methodology. Section 3 analyzes the results of the experimentation and Section 4 outlines the conclusions and the future work.

2. PROPOSED METHOD

2.1 Motivation and Brief Background

Deep neural networks have been used to classify large image datasets, such as ImageNet, and have achieved excellent results [12]. However, they are computationally costly; it would take a long time to setup and train the network for accurate predictions. Pre-trained deep networks trained on large databases and fine-tuned on smaller datasets have been used successfully for complex computer vision tasks, such as face recognition and age estimation [13]. Deep neural networks have, by themselves, some inherent property of improving the scores of minority classes [14]. Pruning of the insignificant features while passing them into the deep networks is a solution to ease out on the computation part [15]. Resampling strategies prevalent in data mining are infeasible for very large, extremely imbalanced image datasets due to the high computational complexity, as is the case in our current work. We therefore, propose to use metric learning using sparse samples for transforming the input space of the large, extremely imbalanced face dataset. Our work is motivated by prior works [2], [16]-[17], [18] that have applied metric learning to mitigate class imbalance for toy datasets. Application of metric learning for large imbalanced datasets, however, requires a lot of computations and very few works have addressed the problem. In our earlier work, which is a precursor of the current work [2], we identified a majority subset of top-186 classes and learned the distance metric using a few samples of each class of the majority subset. The result was an improvement in the performance of majority classes. However, the improvement in the performance of the minority classes was only marginal. In the current work, emphasis is on improving the performance of minority classes by devising a metric learning scheme that would concentrate on the minority classes as well. Some of the earlier techniques propose oversampling of the minority class for improving the performance [19]-[20]. However, this solution is impractical in our case due to the presence of more than a thousand classes and the large size of the dataset. The process pipeline for our method is described in the following sub-sections.

2.2 Deep Feature Extraction from VGG-Face Deep Pre-trained Network

VGG-Face [22], FaceNet [24], DeepFace [25] and OpenFace [39] are a few state-of-the-art deep pre-trained networks customized for face recognition. The Convolutional Neural Network (CNN) [26] is the core neural network of all these models. Because of the large number of hidden layers in their architecture, they are referred to as deep networks. Parkhi *et al.* introduced the pre-trained network VGG-Face in 2014 [22]. It is based on the VGG-16 [27] architecture which consists of 16 convolutional layers followed by a series of pooling and activation layers. The pre-trained network originally trained on two million images is used to generate a 2622x1 feature embedding for each image in our dataset, as shown in the VGG-Face model process flow recreated in Figure 1. The 2622-dimensional feature vectors are further used, in our work, for metric learning and subsequent classification by a suitable classifier.

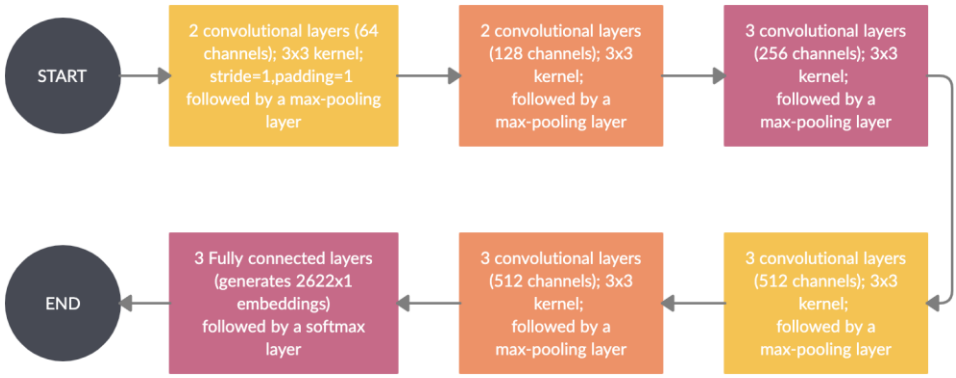


Figure 1. VGG-Face model.

2.3 Metric Learning for Transformation of the Input Space

Various distance metric learning schemes have been proposed, in the past, that improved the classification performance of imbalanced datasets. The aim is to transform the input space so as to bring the samples of a class closer and push samples from different classes farther apart. Some of the most prominent distance metric learning algorithms are based on the Mahalanobis distance that is shown in Equation (1) for two feature vectors (**x, y**).

$$\boldsymbol{dm}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^T \boldsymbol{M} (\boldsymbol{x}-\boldsymbol{y})} \tag{1}$$

Here, $\boldsymbol{M}$ represents the positive semidefinite matrix that is to be estimated. It is similar to the Euclidean distance in a different space or a linear projection of the distance between two points. One of the most popular algorithms using the Mahalanobis distance metric is Large Margin Nearest Neighbor (LMNN). Weinberger *et al.* developed LMNN in 2009 [23] and it has since become one of the most widely used data space modification algorithms. It is a supervized metric learning algorithm that may be used before the classification stage. Optimization problem involved is convex and simple to solve. The cost function shown in Equation (2) is the one that must be minimized.

$$\boldsymbol{costf} = (\delta)\boldsymbol{fu}_{push}(\boldsymbol{L}) + (1-\delta)\boldsymbol{fu}_{pull}(\boldsymbol{L}) \tag{2}$$

The loss function has two terms: one relates to the force that pulls samples from the same class closer together, while the other refers to the force that pushes samples from other classes apart. The cost function in (2) is a weighted sum of push and pull functions. The value of δ lies between 0 and 1. This function's transformation matrix limits the margin between k-similar samples to a minimum and maximizes the margin between samples of different classes. When the number of classes is considerable, direct application of metric learning is not recommended due to the computational complexity involved.

LMNN is based on the principle of bringing the samples belonging to the same class closer and samples belonging to different classes are moved further apart, as illustrated in Figure 2. LMNN thus brings about a global linear transformation of the input space that improves the classification of distance-based classifiers, such as *k*NN. We use the cosine similarity measure in the classification stage that follows the metric learning phase in the process pipeline. The cosine similarity measure

between two feature vectors (**x, y**) is shown in Equation (3).

$$sim(x, y) = \frac{\sum_i x_i \times y_i}{\sqrt{\sum_i x_i^2} \times \sqrt{\sum_i y_i^2}} \tag{3}$$

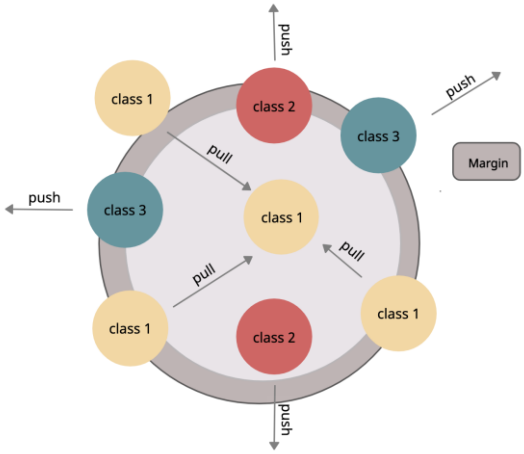


Figure 2. Diagrammatic representation of how LMNN attempts to bring similar & nearest k samples closer and moves dissimilar samples further apart (k=3, as in original paper [23]).

Other examples of metric learning are Neighborhood Component Analysis (NCA) [28], Metric Learning for Kernel Regression (MLKR) [29], Information Theoretic Metric Learning (ITML) [30], Least Squares Metric Learning (LSML) [31] and Sparse Discriminant Metric Learning (SDML) [32]. It was proved in a recent work that LMNN, NCA and MLKR yield the best performance for the *k*NN classification of toy datasets having high imbalance ratio.

2.4 Methodology

The basic outline of the methodology used in our experiments is described next. We segregate the majority and minority classes based upon the number of samples that each class contains. Figure 3 shows the class populations of the Labeled Faces in the Wild (LFW) face dataset [21] used in our experiments, that range from 530 to 2. The graph shows an extremely uneven population distribution.

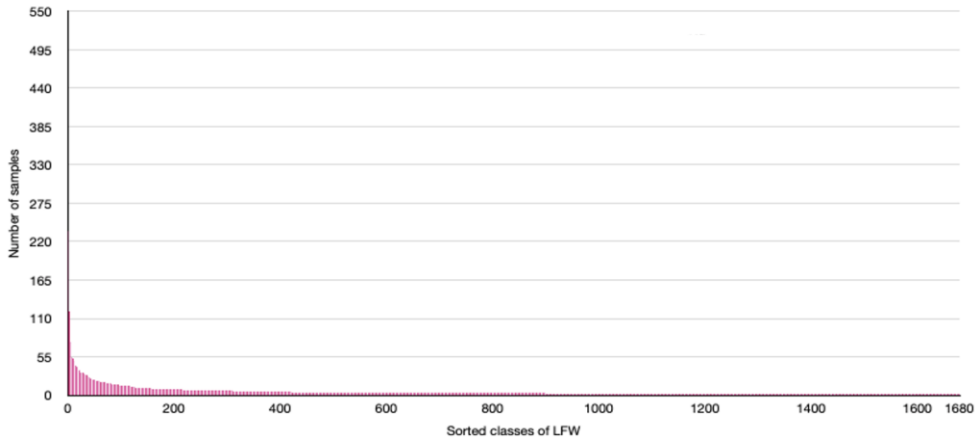
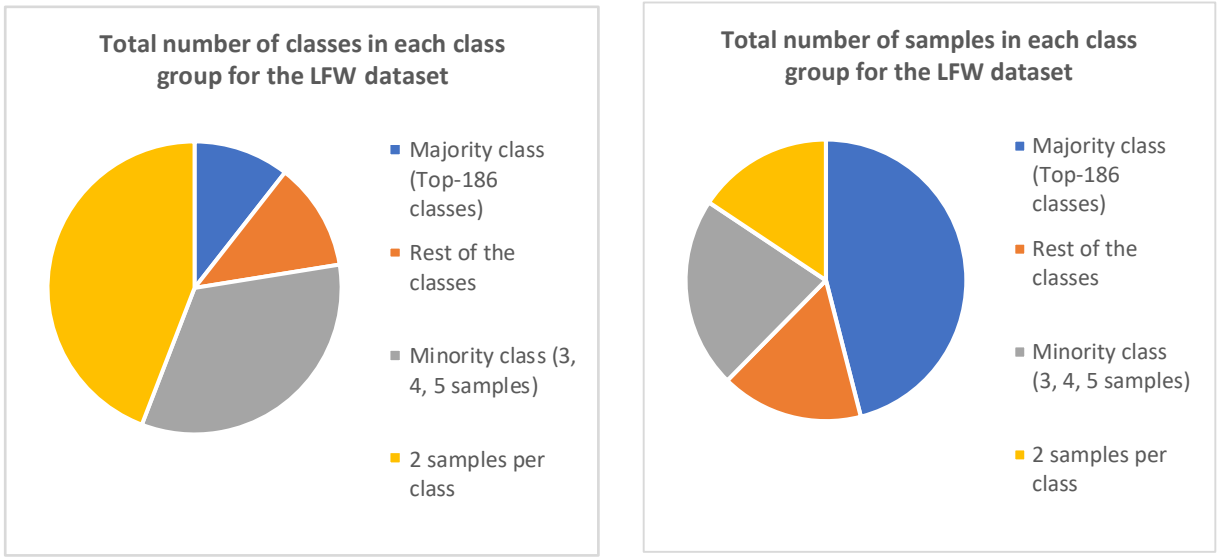


Figure 3. Sorted class populations of the LFW dataset containing 1680 classes of celebrity faces.

As the number of minority classes is very high as compared to the majority classes in the extremely imbalanced LFW dataset, we select a majority class subset and a minority class subset to perform metric learning twice, one from the perspective of the majority class and the other from the perspective of the minority class. This is the primary contribution of our work. We divided the classes based on the class populations, as shown in Figure 4. We define the group of top-186 classes as the majority class and the group of classes with 3, 4 or 5 samples per class as the minority class. To derive

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 07, No. 04, December 2021.

the top-186 classes, as per the procedure in our previous work [2], the class populations are sorted in the decreasing order of their populations and a sum-based partitioning of the sorted class populations yields the threshold as 186 as the lower boundary of the majority class in the LFW dataset. The class-wise and sample-wise groupings are shown in the pie charts in Figure 4 (a) and Figure 4 (b), respectively. It is noted from the pie chart in Figure 4 (a) that the number of minority classes having less than or equal to 5 samples is more than 1300 out of the total available 1680 classes.



(a) (b)

Figure 4. Grouping of the 1680 classes of LFW dataset into majority and minority classes (a) class-wise distribution (b) sample-wise distribution.

Due to the computational complexity involved in metric learning, we have excluded the rest of the classes having samples in the range of 6 to 8 and those having 2 samples from the metric learning computations, since the number of such classes is large and inclusion of these two groups would render metric learning computationally infeasible and impractical. The input space transformation after learning the distance metric is, however, applied to the entire training space. Also, only a few samples from each class are taken into consideration while performing metric learning to reduce the computational expense. Our learning framework is thus an instance of weakly supervized learning. The block diagram for our learning model is shown in Figure 5.

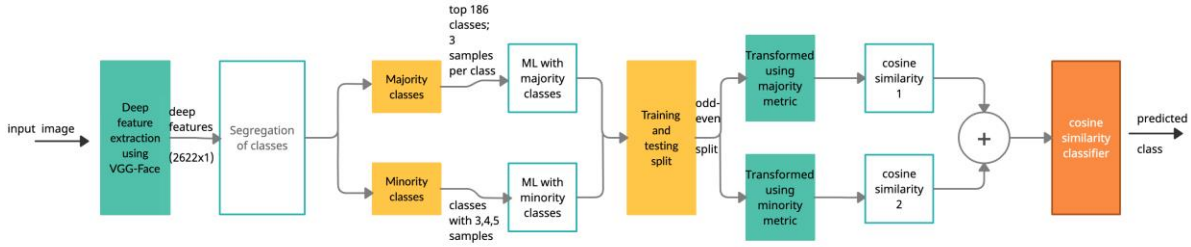


Figure 5. Proposed model.

A deep neural network VGG-Face [22] that is pre-trained on two-million facial images is used to generate the feature embeddings for the LFW face dataset. The gray input images of dimension 64x64 were given as input to the VGG-Face model and vector embeddings of dimension 2622x1 were extracted as per the guidelines in the original paper of VGG-Face [22]. We perform metric learning for both majority class and minority class subsets and get two different transformation metrics. We considered only 3 samples per class from both subsets to reduce the computational cost, since a large number of inter- and intra-class distances need to be calculated. The entire input space is divided into two parts; i.e., training and testing based on alternate sampling and after that the entire training subset

is transformed using both the minority class metric and majority class metric. We use Large Margin Nearest Neighbor (LMNN) as the metric learning technique and it is based on the nearest neighbor rule. The number of nearest neighbors is fixed as k=3, which is the same as given in the original paper of LMNN proposed by Weinberger *et al*. in 2009 [23]. For the minority classes having fewer than 3 samples in the training set, all the samples in the training set are included. The number of neighbors is to be kept small, since a large number of distances, both inter- and intra-class, need to be calculated, which is computationally expensive. The final step would be the classification stage in which the class label of the test sample has to be determined. For each test sample, we transform it using both the distance metrics and in each case, compute the cosine similarity with each sample in the training set of the transformed input space. The two cosine similarity vectors are summed up and the training sample corresponding to the maximum aggregate cosine similarity is selected as the closest neighbor in the training space; its class label is assigned to the test sample.

3. RESULTS

The experiments were performed on the publicly available dataset Labeled Faces in the Wild (LFW) developed in 2007 by Huang *et al.* [21]. It is today a benchmark in the field of facial recognition that is used for training several state-of-the-art pre-trained networks for face recognition. It is a highly imbalanced dataset consisting of 1680 celebrity classes with George W. Bush having the maximum number of samples (=530) and Michel Duclos having the minimum number of samples (=2). Only those celebs were selected who have two or more than two samples and the celebs with only one sample were discarded from our experiments. Out of the 1680 selected celebs, 1369 celebs have <=5 samples, as verified from the pie charts in Figure 4, which proves that the minority classes outnumber the majority classes in the LFW dataset.

We extracted the deep features using the pre-trained VGG-Face model, as discussed in Section 2. The gray-scale images were resized to dimensions 64x64. The pre-trained model generated the 2622-dimensional feature embeddings which were further fed to the learning module. The dataset was divided into majority and minority class subsets as explained in Section 2 and two-way metric learning was performed using these two subsets. The transformation matrices generated were used to transform the entire input space twice, separately. The cosine similarity measure was used to find the closeness of the test sample to a training sample in both the transformed spaces; this was followed by a simple summation of the cosine similarity measures.

The dataset was divided into training and testing sets by alternate sampling. In case of odd number of samples n, the training set contained (n+1)/2 samples and the test set contained (n-1)/2 samples. Cross-validation is done by swapping the training and test sets. The results - Accuracy, F1-score and AUC scores obtained from ROC curves, are compiled in Table 1 for both Validation (V) and Cross-Validation (CV). We compared the performance of our method with that of existing methods: HOG + SVM [33], HOG+ Cosine similarity [34], HOG + Metric learning with majority class [2], VGG-Face + SVM [37], VGG-Face + Cosine similarity [38] and VGG-Face + Metric learning with majority class [36]. The proposed method outperformed all existing methods in terms of accuracy, F1-score and AUC scores as observed from Table 1. The scores are overall on the lower side due to the inclusion of the entire set of 1680 classes including the 779 minority classes with only 2 samples per class of which one sample is used for training and one sample for testing. Most of the earlier experiments on LFW dataset report results only for the majority classes that have at least 10 samples following the face verification protocol in the LFW technical report in [21]. The minority classes are excluded from previous works, since they contribute to class imbalance and deteriorate the overall performance. The state-of-the-art deep networks need a large number of training samples per class for efficient classification [35].

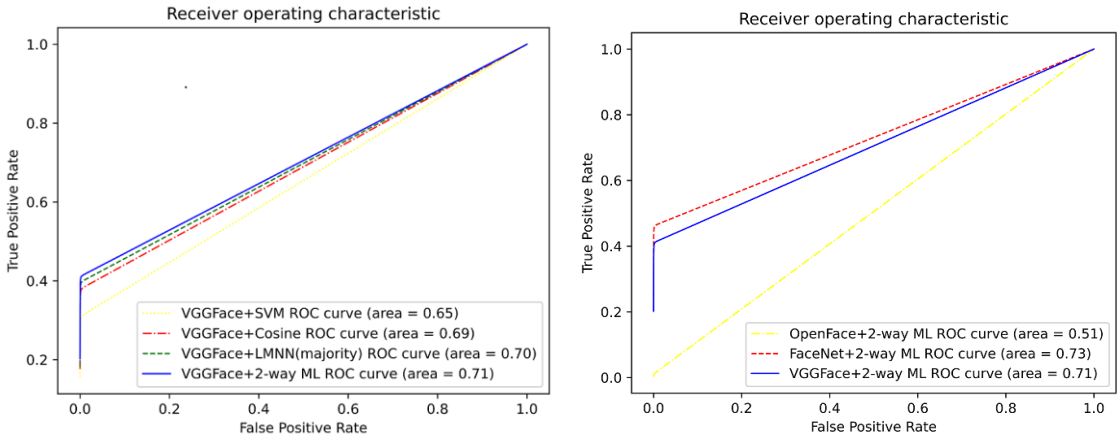
We also compared the current results with that of the weakly supervized metric learning scheme using majority classes [36] by substituting the traditional HOG features in [2] with VGG-Face deep features. We have used an Intel i5 dual core processor clocked at 2.7 GHz and python 3.7 software platform to perform the experiments. The system took half an hour each to learn the distance metrics for both majority and minority subsets. The classification took only a few minutes to execute. The deep features were found to outperform the HOG features, as observed from the classification scores in Table 1. The corresponding ROC curves are shown in Figure 6. A comparison of models that classify

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 07, No. 04, December 2021.

VGG-Face deep features from the ROC graph in Figure 6 (a) reveals that two-way metric learning prior to classification improves the performance scores. Figure 6 (b) compares our method on two other deep features other than VGG-Face; i.e., FaceNet [24] and OpenFace [39]. The FaceNet model performs better than VGG-Face for the proposed two-way metric learning scheme.

Table 1. Performance comparison of various methods on the LFW dataset.

Method	AUC		F1-score		Accuracy	
	V	CV	V	CV	V	CV
HOG + SVM [33]	0.528	0.523	0.059	0.053	28.7%	27.4%
HOG + Cosine Similarity [34]	0.544	0.541	0.078	0.076	21.5%	19.1%
HOG + LMNN metric learning with majority subset [2]	0.556	0.554	0.1006	0.097	26.8%	24.6%
VGG-Face + SVM [37]	0.654	0.635	0.287	0.263	55.4%	51.1%
VGG-Face + Cosine similarity [38]	0.689	0.675	0.342	0.329	55.4%	51.9%
VGG-Face + LMNN metric learning with majority subset [36]	0.697	0.681	0.355	0.339	57%	53.6%
VGG-Face + Two-way metric learning (proposed)	0.705	0.689	0.372	0.356	58.5%	54.8%

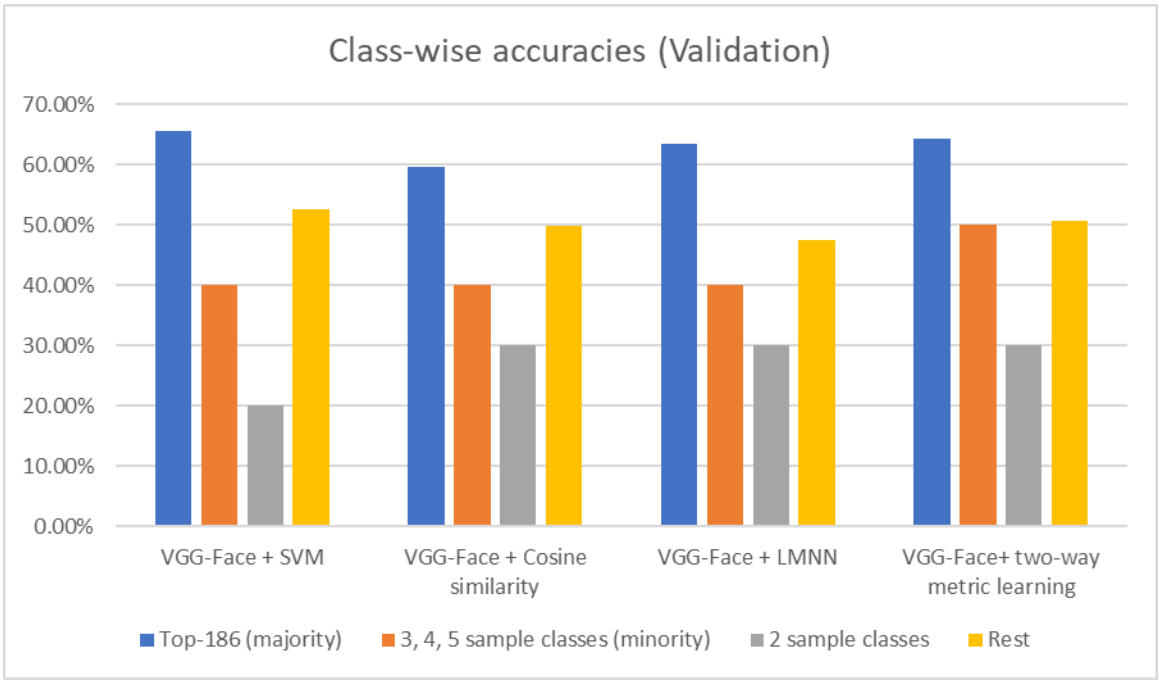


(a) (b)

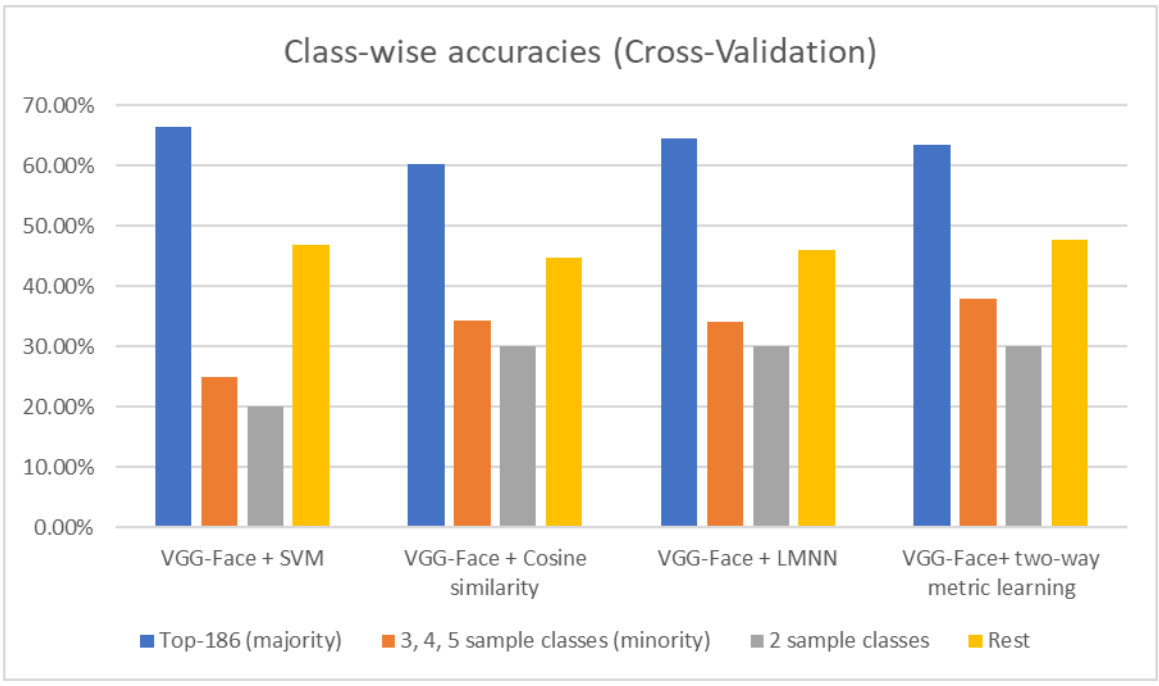
Figure 6. ROC graphs for (a) classification based on VGG-Face deep features by various methods, (b) classification based on VGG-Face, FaceNet and OpenFace deep features for the proposed two-way metric learning scheme.

Some of the class-wise accuracies are shown in Figure 7 for the majority and minority classes to understand the impact of our two-way metric learning scheme as opposed to a scenario where there is no metric learning and the deep features extracted from VGG-Face are learned directly by the classification framework. The cosine similarity measure is the classifier. As observed, Figure 7 shows a consistent performance for all majority classes for the two-way metric learning scheme which is at par with VGG-Face + SVM. However, for the minority classes with 3, 4, 5 samples per class, a significant improvement in accuracy was recorded, with the accuracies jumping from 0% to 50% and above, for most of the minority classes.

The 2-sample classes showed a higher performance than VGG-Face+SVM. The category of classes tagged as “Rest” contain about 6 to 8 samples each. The performance of this set of classes was found improved over VGG-Face+Cosine similarity and VGG-Face+LMNN, though the performance was marginally lower than that of VGG-Face+SVM.



(a)



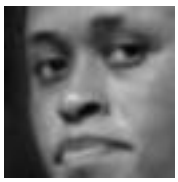
(b)

Figure 7. Performance comparison of majority class (top-186 classes), minority class (3, 4, 5 samples), 2-sample classes, rest of the classes (with samples in the range 6 to 8), with and without metric learning for (a) validation and (b) cross-validation experiments. (*LMNN was used as metric learning method).

Some success cases and failure cases for the proposed method are shown in Figure 8. The success cases shown are examples when metric learning proved to be useful for the classification. The failure cases are those which were not classified by our method. Figure 9 shows the comparison between NCA, LMNN and MLKR metric learning schemes for the proposed methodology of two-way metric learning with deep features. The classification accuracies achieved for the top-10 majority classes are shown for all three metric learning schemes.

It is noted that LMNN significantly outperforms NCA and MLKR in terms of classification accuracies. LMNN involves minimization of the distance between each training sample and its k

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 07, No. 04, December 2021.

	Ai Sugiyama (5 samples)	Allyson Felix (5 samples)	Chanda Rubin (5 samples)	Ricky Santorum (4 samples)
Training sample				
Testing sample				

(a)

	Michel Duclos (2 samples)	Jerry Falwell (2 samples)	Chuck Yeager (2 samples)	Michelle_Rodriguez (2 samples)
Training sample				
Testing sample				

(b)

Figure 8. (a) Some success cases of the proposed method, where VGG-Face features without metric learning could not classify the faces and (b) Some failure cases of the proposed method.

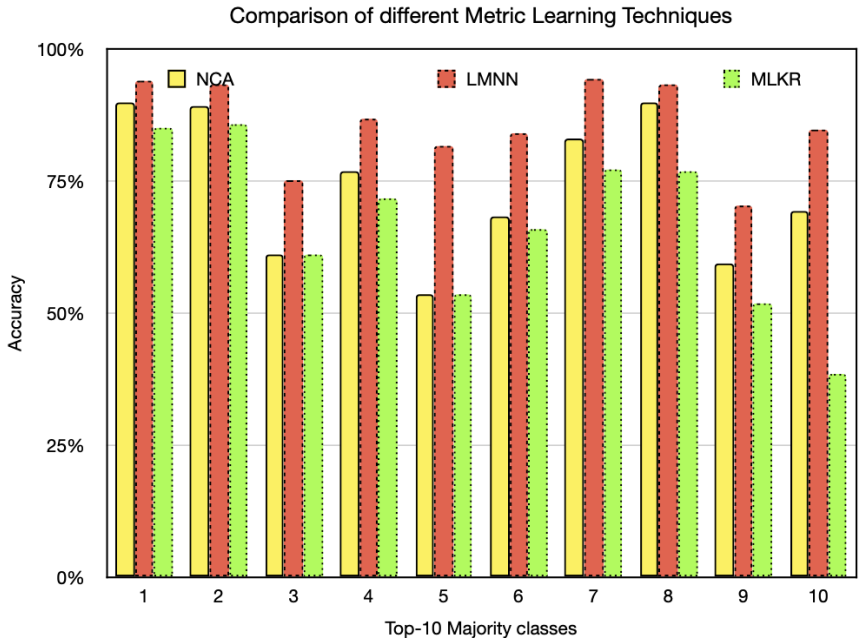


Figure 9. Performance comparison of top-10 majority classes of LFW for different metric learning techniques NCA, MLKR and LMNN.

nearest neighbors belonging to the same class while pushing the differently labelled samples farther apart. LMNN thus projects the input space into metric space in such a way that the inter-class similarities could be measured more accurately.

The primary contribution of our work as compared to our previous work and other works in literature is the improvement of accuracy of the minority classes. The challenge, here, was the existence of more than a thousand minority classes containing sparse samples, rendering metric learning a difficult task. On comparison with other methods, especially VGG-Face+SVM, we observe that though the accuracy of the majority class was comparable, the accuracy of the minority class significantly improved over all existing methods.

4. CONCLUSIONS

A novel learning methodology for large, extremely imbalanced face databases is proposed in this paper that involves deep features and two-way metric learning. LMNN is the metric learning scheme used. Deep features are extracted from the VGG-Face pre-trained model that is trained on two-million facial images. Majority and minority class subsets are identified based on the class population. Metric learning is applied twice, once for the majority subset and the second time for the minority subset. The closeness of the test sample from each training sample in the twice-transformed input space is measured using the sum of the cosine similarities computed in the two cases. The class of the closest training sample, in both the transformed spaces taken together, is assigned as the class of the test sample. Metric learning is known to transform the input space to bring samples of a class closer together. Two-way metric learning introduced in our scheme aims to improve the classification scores, especially for the minority classes, since it brings the few samples in the minority class closer together. Experiments were conducted on the LFW face dataset containing more than a thousand minority classes and the classification scores achieved indicate that the proposed learning technique is more effective than the existing methods for the classification of large, extremely imbalanced face datasets. The LFW dataset, to the best of our knowledge, is the largest extremely imbalanced dataset available for face recognition which is the central theme of this paper. Our method can be easily adapted to other datasets having different scales of imbalance.

REFERENCES

[1] C. Huang, Y. Li, C. C. Loy and X. Tang, "Deep Imbalanced Learning for Face Recognition and Attribute Prediction," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 11, pp. 2781-2794, 2019.

[2] S. Susan and Ashu Kaushik, "Weakly Supervized Metric Learning with Majority Classes for Large Imbalanced Image Dataset," Proceedings of the 4th International Conference on Big Data and Internet of Things, pp. 16-19, DOI: 10.1145/3421537.3421549, 2020.

[3] H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, 2009.

[4] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang and C. Jiang, "Random Forest for Credit Card Fraud Detection," Proc. of the 15th IEEE International Conference on Networking, Sensing and Control (ICNSC), pp. 1-6, Zhuhai, China, 2018.

[5] F. Zhang, G. Liu, Z. Li, C. Yan and C. Jiang, "GMM-based Undersampling and Its Application for Credit Card Fraud Detection," Proc. of the IEEE International Joint Conference on Neural Networks (IJCNN), pp. 1-8, Budapest, Hungary, 2019.

[6] S. Susan and A. Kumar, "DST-ML-EkNN: Data Space Transformation with Metric Learning and Elite K-nearest Neighbor Cluster Formation for Classification of Imbalanced Datasets," Proc. of Advances in Artificial Intelligence and Data Engineering, Part of the Advances in Intelligent Systems and Computing Book Series (AISC), vol. 1133, pp. 319-328, Springer, Singapore, 2021.

[7] H. Zhu, G. Liu, M. Zhou, Y. Xie, A. Abusorrah and Q. Kang, "Optimizing Weighted Extreme Learning Machines for Imbalanced Classification and Application to Credit Card Fraud Detection," Neurocomputing, vol. 407, pp. 50-62, DOI: 10.1016/j.neucom.2020.04.078, 2020.

[8] Z. Li, M. Huang, G. Liu and C. Jiang, "A Hybrid Method with Dynamic Weighted Entropy for Handling the Problem of Class Imbalance with Overlap in Credit Card Fraud Detection," Expert Systems with Applications, vol. 175, pp. 114750, DOI: 10.1016/j.eswa.2021.114750, 2021.

[9] S. Wang and X. Yao, "Multiclass Imbalance Problems: Analysis and Potential Solutions," IEEE Trans. on Systems, Man and Cybernetics, Part B (Cybernetics), vol. 42, no. 4, pp. 1119-1130, 2012.

[10] T. Hasanin, T. M. Khoshgoftaar, J. L. Leevy and R. A. Bauder, "Severely Imbalanced Big Data Challenges: Investigating Data Sampling Approaches," J. of Big Data, vol. 6, no. 1, pp. 1-25, 2019.

[11] B. Kulis, "Metric Learning: A Survey," Foundations and Trends in Machine Learning, vol. 5, no. 4, pp. 287-364, 2012.

[12] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems, vol. 25, pp. 1097-1105, 2012.

[13] A. Al-Shannaq and L. Elrefaei, "Age Estimation Using Specific Domain Transfer Learning," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 6, no. 2, pp. 122-139, 2020.

[14] J. M. Johnson and T. M. Khoshgoftaar, "Survey on Deep Learning with Class Imbalance," Journal of Big Data, vol. 6, no. 1, pp. 1-54, 2019.

[15] R.-C. Chen and C.-Y. Liao, "Deep Learning to Predict User Rating in Imbalance Classification Data Incorporating Ensemble Methods," Proc. of the IEEE International Conference on Applied System Invention (ICASI), pp. 200-203, Chiba, Japan, 2018.

[16] N. Wang, X. Zhao, Y. Jiang and Y. Gao, "Iterative Metric Learning for Imbalance Data Classification," Proc. of the 27th International Joint Conference on Artificial Intelligence (IJCAI-18), pp. 2805-2811, [Online], available: https://www.ijcai.org/proceedings/2018/0389.pdf, 2018.

[17] L. Gautheron, A. Habrard, E. Morvant and M. Sebban, "Metric Learning from Imbalanced Data with Generalization Guarantees," Pattern Recognition Letters, vol. 133, pp. 298-304, 2020.

[18] S. Susan and A. Kumar, "Learning Data Space Transformation Matrix from Pruned Imbalanced Datasets for Nearest Neighbor Classification," Proc. of the IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), pp. 2831-2838, Zhangjiajie, China, 2019.

[19] S. Barua, Md. M. Islam, X. Yao and K. Murase, "MWMOTE - Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 2, pp. 405-425, 2012.

[20] V. Ganganwar, "An Overview of Classification Algorithms for Imbalanced Datasets," International Journal of Emerging Technology and Advanced Engineering, vol. 2, no. 4, pp. 42-47, 2012.

[21] G. B. Huang, M. Mattar, T. Berg and E. Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," Technical Report in Workshop on Faces in'Real-Life'Images: Detection, Alignment and Recognition, [Online], Available: http://vis-www.cs.umass.edu/lfw/lfw.pdf, 2008.

[22] O. M. Parkhi, A. Vedaldi and A. Zisserman, "Deep Face Recognition," Proc. of the British Machine Vision Conference (BMVC), pp. 41.1-41.12, [Online], Available: https://www.robots.ox.ac.uk/~vgg/ publications/2015/Parkhi15/parkhi15.pdf, Sep. 2015.

[23] K. Q. Weinberger and L. K. Saul, "Distance Metric Learning for Large Margin Nearest Neighbor Classification," Journal of Machine Learning Research, vol. 10, no. 2, pp. 207-244, 2009.

[24] F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815-823, Boston, MA, USA, 2015.

[25] Y. Taigman, M. Yang, M. A. Ranzato and L. Wolf, "DeepFace: Closing the Gap to Human-level Performance in Face Verification," Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1701-1708, Columbus, OH, USA, 2014.

[26] Y. LeCun and Y. Bengio, "Convolutional Networks for Images, Speech and Time Series," The Handbook of Brain Theory and Neural Networks, vol. 3361, no. 10, pp. 1-14, 1995.

[27] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition," Proc. of ICLR 2015, arXiv preprint arXiv:1409.1556, 2014.

[28] J. Goldberger, G. E. Hinton, S. T. Roweis and R. R. Salakhutdinov, "Neighborhood Components Analysis," Advances in Neural Information Processing Systems, pp. 513-520, [Online], Available: https://www.cs.toronto.edu/~hinton/absps/nca.pdf, 2005.

[29] K. Q. Weinberger and G. Tesauro, "Metric Learning for Kernel Regression," Artificial Intelligence and Statistics, pp. 612-619, Ava.: http://proceedings.mlr.press/v2/weinberger07a/Weinberger07a.pdf, 2007.

[30] J. V. Davis, B. Kulis, P. Jain, S. Sra and I. S. Dhillon, "Information-theoretic Metric Learning," Proc. of the 24th Int. Conf. on Machine Learning, pp. 209-216, DOI: 10.1145/1273496.1273523, ACM, 2007.

[31] E. P. Xing, M. I. Jordan, S. J. Russell and A. Y. Ng, "Distance Metric Learning with Application to Clustering with Side-information," Proc. of the 15th International Conference on Neural Information Processing Systems (NIPS'02), pp. 521-528, 2003.

[32] G.-J. Qi, J. Tang, Z.-J. Zha, T.-S. Chua and H.-J. Zhang, "An Efficient Sparse Metric Learning in High-dimensional Space *via* l_1-penalized Log-determinant Regularization," Proc. of the 26th Annual Int. Conf. on Machine Learning, pp. 841-848, DOI: 10.1145/1553374.1553482, ACM, 2009.

[33] H. S. Dadi and G. K. M. Pillutla, "Improved Face Recognition Rate Using HOG Features and SVM Classifier," IOSR Journal of Electronics and Communication Eng., vol. 11, no. 04, pp. 34-44, 2016.

[34] D. Chen, X. Cao, F. Wen and J. Sun, "Blessing of Dimensionality: High-dimensional Feature and Its Efficient Compression for Face Verification," Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3025-3032, Portland, OR, USA, 2013.

[35] Y. C. Wong, L. J. Choi, R. S. Sarban Singh, H. Zhang and A. R. Syafeeza, "Deep Learning-based Racing Bib Number Detection and Recognition," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 5, no. 3, pp. 181-194, 2019.

[36] A. Kaushik and S. Susan, "Metric Learning with Deep Features for Highly Imbalanced Face Dataset," Proc. of the International Conference on Innovative Computing and Communications, Part of the Advances in Intelligent Systems and Computing Book Series, vol. 1394, pp. 639-646, 2022.

[37] B. Knyazev, R. Shvetsov, N. Efremova and A. Kuharenko, "Leveraging Large Face Recognition Data for Emotion Classification," Proc. of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 692-696, Xi'an, China, 2018.

[38] S. Karahan, M. K. Yildirum, K. Kirtac, F. S. Rende, G. Butun and H. K. Ekenel, "How Image Degradations Affect Deep CNN-based Face Recognition?," Proc. of the International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1-5, Darmstadt, Germany, 2016.

[39] T. Baltrušaitis, P. Robinson and L.-P. Morency, "OpenFace: An Open Source Facial Behavior Analysis Toolkit," Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1-10, Lake Placid, NY, USA, 2016.

ملخص البحث:

تقترح هذه الورقة منهجية تعلّم جديدة تتضمّن سِماتٍ عميقةٍ والتّعلُّم ذا الطّريقيْن لمجموعات البيانات المتعلّقة بتمييز الوجوه التي تتميّز بالضّخامة وشدّة عدم الاتّزان حيث عدد الأصناف الأقلية ومعدّل عدم الاتّزان كبيران جداً. وتبرز المشكلة من أنّ بعض الوجوه الأكثر شيوعاً متوافرة في وسائلالتّواصل الاجتماعي والإنترنت، بينما وجوه الشّخصيات المعروفة بدرجةٍ أقلّ هي أقلّ عدداً. ولأنّ إعادة تشكيل العيّنات في ظلّ هذا السيناريو أمرٌ غير عمليّ، فإننا نقترح التّعلّم القياسي أداةً للتخفيف من مشكلة عدم الاتّزان في الأصناف قبل مرحلة التّصنيف. وللتقليل من تكلفة الحسابات المرتبطة بالتّعلّم القياسي، نقوم –بشكل منفصل- بإجراء التّعلّم القياسي المراقب بشكلٍ ضعيف على مجموعات البيانات الفرعية الأغلبية والأقلية، في عمليةٍ نسميها التّعلّم القياسي ذا الطريقيْن. وقد أثبتت التجارب على مجموعة البيانات المرجعية لتمييز الوجه LFW التي تحوي (1680) صنفاً من الوجوه أنّ المنهجية المقترحة كانت أكثر فاعلية مقارنةً بالطرق القائمة. حيث انّ دقّة التّصنيف للأصناف الأقليّة ارتفعت على نحوٍ خاص، وهو إنجاز نادر في طرق تصنيف مجموعات البيانات المعتمدة على التّعلّم غير المتوازن العميق.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

View publication stats

2021 8th International Conference on Signal Processing and Integrated Networks (SPIN) | 978-1-6654-3564-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SPIN52536.2021.9566106

Unidirectional Onboard Battery Charging for Electric Vehicle using Interleaved Luo Converter

Chanchal Chaudhary
Electrical Engineering Department
Delhi Technological University
Shahbad Daulatpur, Main Bawana Road, Delhi NCR-110042
Delhi, India
chanchalchaudhary94@gmail.com

Saurabh Mishra
Electrical Engineering Department
Delhi Technological University
Shahbad Daulatpur, Main Bawana Road, Delhi NCR-110042
Delhi, India
saurabhmishra@dtu.ac.in

Uma Nangia
Electrical Engineering Department
Delhi Technological University
Shahbad Daulatpur, Main Bawana Road, Delhi NCR-110042
Delhi, India
uma_nangia@rediffmail.com

Abstract*— The intent of paper is to deal with the design of onboard battery charger by interleaved Luo converter topology cascaded with flyback converter to charge a battery. As per this configuration two Luo converters are interleaved and connected in parallel so that the current at input gets divided into them to ensure reliable operation of charger in discontinuous conduction mode. The charger efficiency gets improved because of the reduced stresses of current in the system moreover system works well due to inclusion of less ripple in current at input and output due to the cancellation in ripple within the circuit. Control technique involves closed loop control of cascaded PI in flyback to gain better and controlled results. Distortions in the circuit are reduced to a larger extent and THD is reduced within IEC 61000-3-2 standards. Prototype simulation is validated in MATLAB SIMULINK.

Keywords—: Interleaved Luo converter, Power Factor Correction, Total Harmonic Distortion (THD), Electric vehicle (EV), Discontinuous conduction mode (DCM), Flyback converter,

I. INTRODUCTION

Global demands for clean fuel energy are rising rapidly as fuel shortage and pollution are the major concerns and consequently it results in pushing researchers towards more and more to go for findings in electric vehicle (EV) area. EV is one of the wide areas of research as they put forward some realistic and reliable solution in the field of energy that came in existence after the concerns of global warming has arisen. Also, with energy modernization and globalization, electrification in vehicles is a need of time. That is why government has also promoted EV on a large scale and has decided to expand the EV industry on a large scale. With development in EV main concerns related to it is battery and charging of its battery, its efficiency and fastness of charging. In view these, two types of charging are discussed broadly i.e., onboard charging and offboard charging. Onboard charging has become popular because of the shortage of availability of charging stations while the power quality should also be maintained along with these issues because deterioration in power quality as given in [1] can raise problems in the charging phenomenon. So, while addressing EV apart from high battery cost and slow charging concerns, power factor and THD should also be maintained within limits so that charger can work efficiently. The main aim is to make EV accessible, lower in cost and energy efficient with less losses along with reliable charging system equipped in it. Deterioration in power factor less than unity affects the system and the connected components working efficiency negatively.

In conventional chargers generally Diode bridge rectifier (DBR) is used, for the conversion of AC-DC and power electronics converters especially basic rectifiers introduces harmonics in huge amount. So, usage of DBR introduces a low power quality problem in the charger which leads to rise in THD to larger extent as well as deteriorating the power factor much less than unity. Consequently, efficiency of the charger is at stake making it less economical to use. In conventional chargers a power factor correction (PFC) unit is used which includes any of the DC-DC converters like Boost, Buck, Buck-Boost [2] etc., a general conventional charger topology can be seen in [3] and [4]. In general buck and boost topologies are not frequently used in PFC because variation in duty cycle is having a certain limited range only so buck-boost is a better choice over both of them as per [5]. Apart from buck-boost nowadays new converters like Cuk, SEPIC, Zeta, Luo, landsman etc. converters are used as PFC unit due to their added advantages of improved current profile, less ripples in output voltage and current as mentioned in [6]. Although Cuk as well as SEPIC are good choices among PFC units yet SEPIC has problems regarding large ripples in output voltage while Zeta and Luo converters have some added advantages of good regulation of light load, low charging current ripples etc. as per [7]. Nowadays, we can use new configurations like Bridgeless, isolated [8] and interleaved converters [9] for some added and improved Power factor correction advantages, some of them are mentioned in [10]. These new and combinational converters are having edge over conventional converters both in terms of power factor correction at input as well as reduction in THD also resulting in good performance of the charger thereby enhancing its efficiency.

This paper illustrates an interleaved Luo converter topology, as discussed above it is having better power factor correction features making it almost unity. THD reductions are also very high so it becomes excellent for a low cost and efficient charging device to use an interleaved Luo converter as PFC unit. This configuration also includes a flyback converter whose design is mentioned in [11]. Flyback converter provides isolation in the circuit and it is used here due to its simpler design and working. Flyback converter's working includes CC-CV charging. Both the converters are made to work in DCM mode to ensure reliable operation of system. Topology for control used here is PI however PID technique of control is discussed in [12]. interleaved Luo

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:39:37 UTC from IEEE Xplore. Restrictions apply.

converter is having a special feature of ripple cancellation within the circuit due to its connection design consequently the ripple at the output decreases to much larger extent making it apt for onboard charging. So overall interleaved Luo converter in combination with flyback converter is charging the battery efficiently

II. Configuration and Working Operation

The configuration of charger is shown in Fig.1. which illustrates the charging of a battery of 48V,80AH with the combination of an interleaved Luo converter and a flyback converter. DC voltage generated at interleaved Luo converter is including DBR. DC link generated before flyback converter is expected to be 300V while at output of flyback converter it is 65V. This configuration of interleaved converter is made up by interleaving two converter cells into one and connecting them into parallel combination. The operation of these two are simultaneous but having a shift of 180°. The converters working are ensured in DCM mode. So overall it works in six modes as described further.

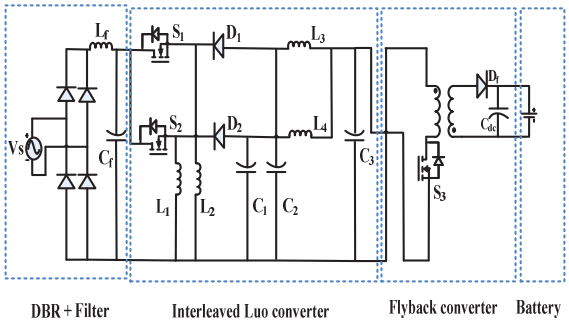


Fig.1. Proposed Configuration

A. Modes of Operation

- *Mode1* In Fig.2. (a) it can be seen that as switch S_1 is working in on mode, the inductor L_2 starts charging and energy is stored from supply mains while inductor L_1 is discharging via diode D_2 (forward biased) by transferring the energy to it. Capacitor C_1 is getting charged and resulting in increase of voltage across it while capacitor C_2 is discharging through L_3.
- *Mode 2* Operation of this mode as demonstrated in Fig.2. (b) where S_1 is the only switch which is conducting and, in this way, L_2 is storing energy via supply mains. Inductor L_1 is completely discharged in this whole process and it enters in DCM mode. Diode D_1 is reverse biased now whereas C_1 and C_2 starts discharging in this mode.
- *Mode 3* S_1 and S_2 both switches are acting in off condition as illustrated in Fig.2. (c). Inductor L_1 is still in DCM mode while L_2 starts discharging via diode D_1 as it is conducting. Immediate charging of capacitor C_2 starts meanwhile the voltage across capacitor C_1 starts decreasing as it is discharging via L_4.
- *Mode 4* In Fig.2. (d) the moment when the switch S_2 is turned on, voltage across C_2 starts rising and diode D_1 and inductor L_2. whereas inductor L_1 is gaining energy via supply mains meanwhile C_1 discharges through L_4 and current is supplied to the battery.
- *Mode 5* In this mode diode D_1 and switch S_1 are non-conducting in nature which shows that inductor L_2 energy is completely discharged and current starts decreasing through it while switch S_2 is still in on mode as per Fig.2. (e). Current in L_1 keeps on increasing as it is charging. Both C_1 and C_2 are getting discharged through L_3 and L_4.
- *Mode 6* Switches S_1 and S_2 as well as diode D_1 remains off in this mode as per the illustration in Fig.3. (f). Inductor L_2 is completely discharged as it enters in DCM mode while diode D_2 starts conducting gradually. Capacitor C_1 is charging via L_4 and voltage starts building up across the capacitor.

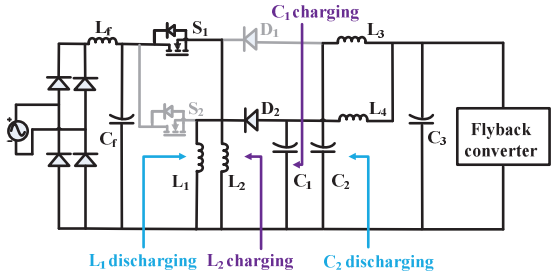


Fig.2. (a) Mode 1

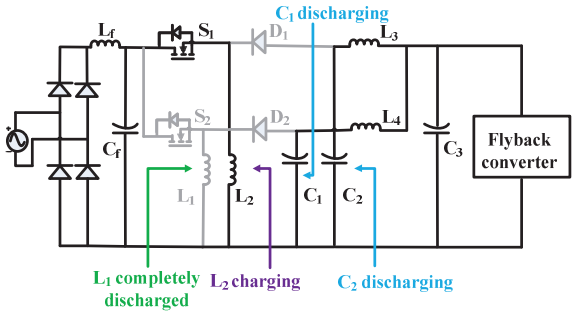


Fig.2. (b) Mode 2

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:39:37 UTC from IEEE Xplore. Restrictions apply.

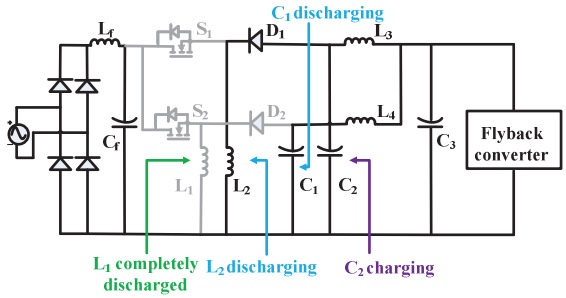


Fig.2. (c) Mode

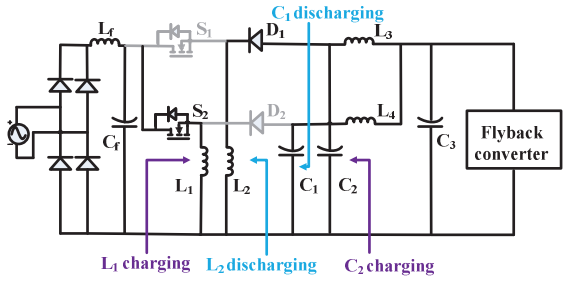


Fig.2. (d) Mode 4

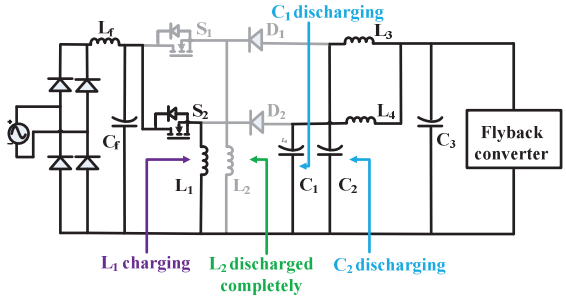


Fig.2. (e) Mode 5

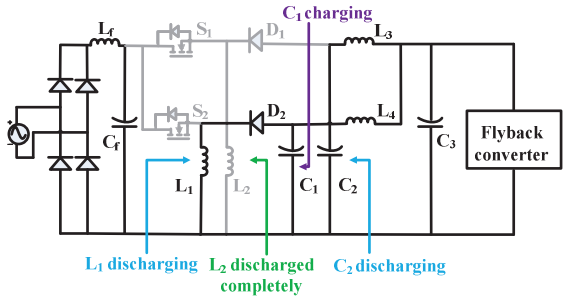


Fig.2. (f) Mode 6

III. DESIGN CONSIDERATIONS

This charging system design includes an interleaved Luo converter cascaded with flyback converter which in turn is charging the battery at the end. The AC-DC conversion is carried out by DBR followed by an EMI filter. The value considered here is average value because of the two converters in which switching frequency is chosen quite higher than frequency of supply. All the calculated chosen values of various parameters are shown in Table 1.

A. Interleaved Luo Converter Design

Interleaved Luo converter is fed by DC voltage created at the output of DBR and EMI filter i.e., average output is denoted as V_{in} and is given by:

$$V_{in} = \frac{2\sqrt{2}V_s}{\pi} \tag{1}$$

where V_s is the supply voltage. The output voltage and input voltage are interconnected as given by equation (2)

$$D = \frac{V_a}{V_a + V_{in}} \tag{2}$$

where V_a is the voltage generated at DC link of Luo converter and D is the duty cycle of interleaved converter. It varies between a minimum value of 0.56 to maximum value of 0.66. Converter has two inductors L_1 and L_2 at the input, they are made to work in DCM mode. So, the critical value of L_1, L_2 is calculated as follows:

$$L_{1,2} = \frac{D_{\min}(1 - D_{\min})}{2 f_s I_{bat}} \tag{3}$$

where f_s is the Interleaved Luo converter switching frequency while I_b is the current which can be calculated as per equation (4).

$$I_b = \frac{P}{V_a} \tag{4}$$

The critical value of mid capacitors C_1 and C_2 is given by:

$$C_{1,2} = \frac{V_a D_{\max}}{2 f_s (\delta/2) R_L} \tag{5}$$

where δ is the capacitor voltage ripple and R_L is the load resistance.

The output capacitors of converter i.e., L_3 and L_4 have the critical values as given by:

$$L_{3,4} = \frac{I_{bat} D_{\max}}{16 f_s^2 C_{1,2} (\sigma/2)} \tag{6}$$

where σ is considered as ripple in current of inductor. Output capacitor C_3 of Luo converter i.e., DC link generated at output of Interleaved Luo converter is given by:

$$C_3 = \frac{I_{bat}}{2\omega \Delta V_3} \tag{7}$$

where ΔV_3 is the ripple voltage at C_3.

B. Flyback Converter Design

Flyback converter design consists of a transformer providing isolation, a diode D_f, a capacitor C_{DC} at output and a switch S_3. The selection of magnetizing inductance L_m plays an important role in high frequency transformer operation as well as flyback converter operation. It is chosen to be very less than the calculated critical value. Value of L_m is given as:

$$L_m = \frac{(V_a D_{fb})^2}{2 f V_{bat} I_{bat}} \tag{8}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:39:37 UTC from IEEE Xplore. Restrictions apply.

Duty cycle D_{fb} for flyback converter operation is given by:

$$D_{fb} = \frac{V_a}{(n_s/n_p)V_{bat} + V_a} \quad (9)$$

Where (n_s/n_p) is secondary to primary turn ratio and is considered as $(1/3)$ here. For the minimal ripple in the output current, an output DC link capacitor C_{DC} is employed. It can be estimated as:

$$C_{DC} = \frac{D_{fb}V_{bat}}{f({V_{bat}}^2/P)\mu} \quad (10)$$

where μ represents the battery voltage ripple.

TABLE I. CHARGER SPECIFICATIONS

S.No.	*Parameters*	*Values chosen*
1.	Supply voltage V_s	220V
2.	Average output voltage V_{in}	198V
3.	Duty ratio (Interleaved Luo) D	0.56-0.66
4.	Switching frequency f_s	20kHz
5.	Input inductors $L_{1,2}$	150µF
6.	Mid capacitors $C_{1,2}$	0.44µF
7.	Output inductors $L_{3,4}$	2.35mH
8.	DC link capacitor C_3	800µF
9.	Magnetizing inductance L_m	130µH
10.	Switching frequency f	50kHz
11.	Duty ratio (Flyback) D_{fb}	0.394
12.	Output DC link capacitor C_{DC}	1500µF

IV. CONTROL STRATEGY

The proposed system deals with a balanced control topology which provides pulses for Interleaved Luo converter at 180° apart in phase. DC link voltage produced at output of interleaved Luo converter is used to compare with reference DC voltage controlling and fed to voltage PI controller as per Fig.3. The switching of flyback converter involves a dual loop control of PI in which battery voltage is being detected and in comparison, to a fixed voltage taken as reference and the compared value is fed to voltage PI controller and cascaded with current PI controller followed by PWM generator which in turn generates controlled pulses for flyback converter switch as shown in Fig.4.

For tuning of PI, various methods can be used. In this Ziegler Nicholas method is used for the same. First transfer function is calculated using steady state analysis along with state space modelling then with the help of Routh Hurwitz criteria value of critical constant K_{cr} is calculated which contributes to the auxiliary equation calculation and in finding out T_{cr}. The values of K_P and K_I can be find out as follows:

$$K_P = 0.45K_{cr} \quad (11)$$

$$K_I = \frac{1.2K_P}{T_{cr}} \quad (12)$$

$$T_{cr} = \frac{2\pi}{\omega_{cr}} \quad (13)$$

Selected values of K_P, K_I are 0.1 and 0.073 respectively.

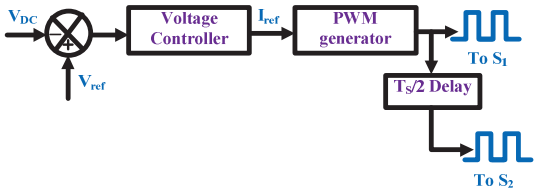


Fig.3. Control of Interleaved Luo switch

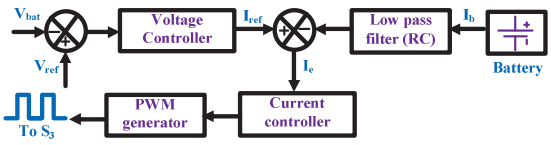


Fig.4. Control of Flyback converter switch

V. RESULTS

The system depicts the values of various parameters of a Li-ion battery of 48V,80AH which is getting charged by proposed system. Fig.5. (a)-Fig.5. (e) represents Input voltage, Input current, Battery voltage, SOC%, battery current respectively. It is clear to see from Fig.5. (a) and Fig.5. (b) that both input voltage and current are in phase which shows power factor correction in the circuit. Increasing SOC% in Fig.5. (d) shows the battery charging in steady manner. Fig.6. shows the variation in THD which is less than 5% in the system. Battery is getting charged up to a constant voltage of 52V with a constant current of almost 15A. Negative value of current shows that battery is getting charged.

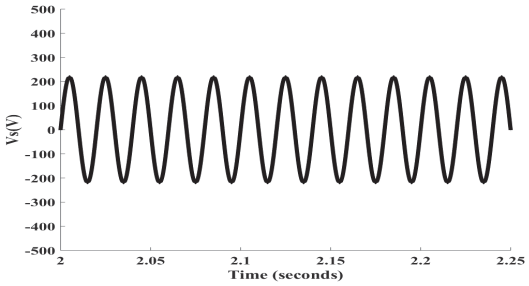


Fig.5. (a) Input Voltage(V)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:39:37 UTC from IEEE Xplore. Restrictions apply.

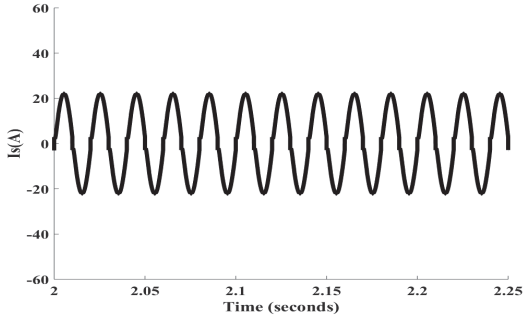


Fig.5. (b) Input Current(A)

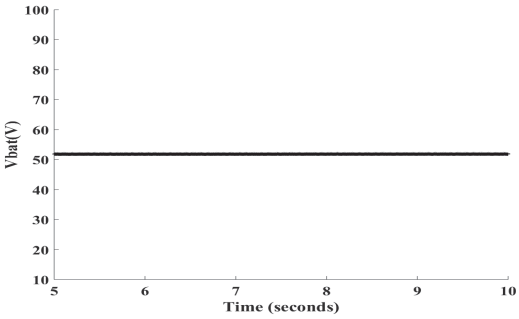


Fig.5. (c) Battery Voltage(V)

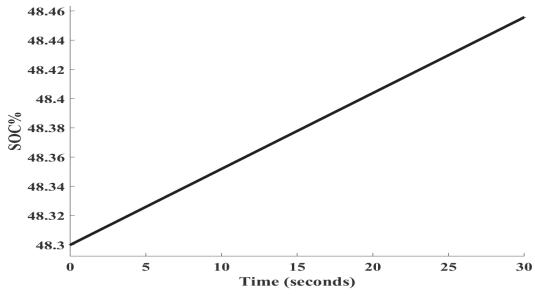


Fig.5 (e) SOC%

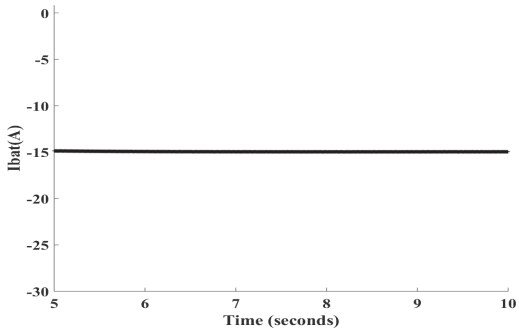


Fig.5. (d) Battery Current(A)

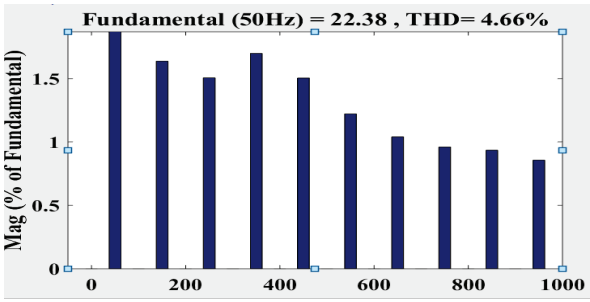


Fig.6. Input Current THD

In Fig.7. the variation of THD among various topologies related to Luo can be seen. It represents how Interleaved converter cuts down the THD of input current effectively making the system energy efficient. Fig.8. illustrates that how power factor is varying with variation in load and with different converter topologies. It may be seen that with interleaved converter almost unity power factor is getting achieved and importance of power factor correction unit can also be clearly detected.

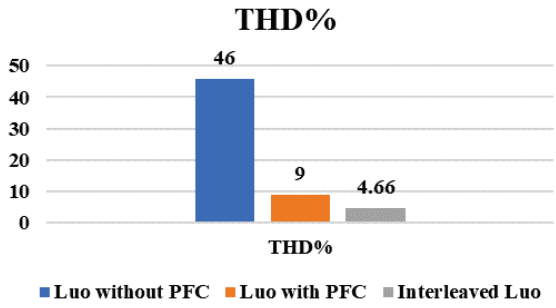


Fig.7. THD comparison

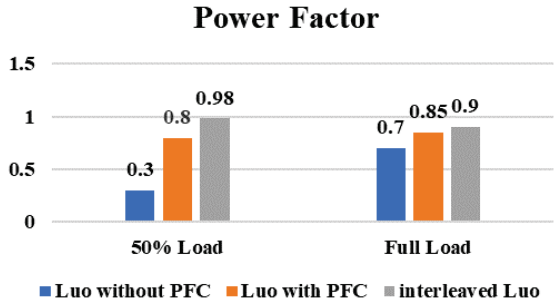


Fig.8. Power Factor comparison

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:39:37 UTC from IEEE Xplore. Restrictions apply.

VI. Conclusion

Proposed system which includes an interleaved Luo converter (combination of two parallelly connected Luo converters) is low cost and efficient topology as compared to other conventional converters. The advantage of cancellation of ripples within the circuit makes it much more reliable to use as a PFC unit. Power factor correction properties of interleaved Luo converter are excellent and THD has also been reduced within 5% as per standards of IEC 61000-3-2.

Acknowledgment

The author thanks to Electrical Engineering Department (EED) of Delhi Technological University, SPIN for providing the platform to present my research learnings and findings.

References

[1] B. Singh, B. N. Singh, A. Chandra, K. Al-Haddad, A. Pandey and D. P. Kothari, "A review of single-phase improved power quality AC-DC converters," in *IEEE Transactions on Industrial Electronics*, vol. 50, no. 5, pp. 962-981, Oct. 2003.

[2] R. Philip and C. Sreeja, "Single phase PFC using Buck-Boost converter "*Annual International Conference on Emerging Research Areas: Magnetics, Machines and Drives (AICERA/iCMMD)"* pp. 1-5, July 2014.

[3] I. Subotic and E. Levi, "A review of single-phase on-board integrated battery charging topologies for electric vehicles,"*IEEE Workshop on Electrical Machines Design, Control and Diagnosis (WEMDCD*), pp. 136-145, march 2015.

[4] S. J. Thomson, P. Thomas, A. R. and E. Rajan, "Design and Prototype Modelling of a CC/CV Electric Vehicle Battery Charging Circuit," 2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET), pp. 1-5, Dec 2018

[5] S. Singh, B. Singh, G. Bhuvaneswari, Vashist Bhist, "Improved power quality switched mode power supply using buck-boost converter", *IEEE Trans. on industry application, International Conference on Power Electronics, Drives and Energy Systems (PEDES),* pp.1-6, Dec 2014.

[6] A. Jha and B. Singh, "Portable Battery Charger for Electric Vehicles," *International Conference on Sustainable Energy and Future Electric Transportation (SEFET),* pp. 1-6, Jan 2021

[7] K. Mishra, M. K. Pathak and S. Das, "Isolated converter topologies for power factor correction — A comparison*," International Conference on Energy, Automation and Signal*, Bhubaneswar, pp. 1-6, 2011

[8] R. Kushwaha, B. Singh, "Design and Development of Modified BL Luo Converter for PQ Improvement in EV Charger," *in IEEE Transactions on Industry Applications*, vol. 56, no. 4, pp. 3976-3984, July-Aug 2020.

[9] R. Kushwaha and B. Singh, "Interleaved Landsman Converter Fed EV Battery Charger with Power Factor Correction*," IEEE 8th Power India International Conference (PIICON*), Dec 2018.

[10] B. Singh and R. Kushwaha, "Power Factor Pre regulation in Interleaved Luo Converter-Fed Electric Vehicle Battery Charger," in *IEEE Transactions on Industry Applications*, vol. 57, no. 3, pp. 2870-2882, May-June 2021.

[11] Nagesha C, A. Sreedevi and M. Gopal, "Simulation and hardware implementation of 24 watt multiple output Flyback converter," *International Conference on Power and Advanced Control Engineering (ICPACE)"*, Bengaluru, pp. 366-370, August 2015.

[12] N. Tiwari and A. N. Tiwari, "Performance Analysis of Unidirectional and Bidirectional Buck-Boost Converter Using PID Controller," 2nd *International Conference on Electronics, Materials Engineering & Nano-Technology*, pp. 1-6, may 2018.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 09,2021 at 11:39:37 UTC from IEEE Xplore. Restrictions apply.

2021 Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 28-29, 2021

WCA Optimized Fuzzy PD controller for Stabilizing the Two Wheel Self-Balancing Robot

Likhitha Kakinada
Dept. of Electrical Engineering
Delhi Technological University
Delhi, India
vidyalikhitha.k@gmail.com

Kuldeep Singh
Dept. of Electrical Engineering
Delhi Technological University
Delhi, India
kuldeepsingh@dtu.ac.in

***Abstract*— An inverted pendulum inspired Two wheeled self-balancing robot (TWSBR) finds several applications in real life. It can also be used to test the working of different control algorithms. In this thesis work the two-wheeled system's transfer function is modelled. In this thesis work the two-wheeled system's transfer function is modelled. Further, for performing the simulation for stabilizing the system three different controllers namely PID, Fractional order PID (FOPID) and Fuzzy PD (FzPD) have been employed. It is common knowledge that controllers give better performance with proper values of the several parameters. These parameters can be adjusted to improve the controller's response. In this work water cycle algorithm (WCA) has been used to tune the parameters of all three controllers. Comparative analysis among different controllers shows that FzPD controller outperforms the PID and FOPID controller in terms of stabilizing the system. Different graphs have been shown in the thesis work to demonstrates the working of the simulation.**

Keywords—Fuzzy, FzPD, MATLAB, TWSBR, MFs

I. INTRODUCTION

The TWSBR has developed a warm subject in proving numerous control concepts, which mostly regulate the system during the unbalanced dynamic as well as nonlinearities conditions. Along with that, it has other benefits for self-governing mobile such as insignificant size, simple construction as well as flexibility in operation. As a result of this concept, it may be appropriate for some labor that takes place in a confined space or is hazardous. In general, a TWSBR is an excellent resource to test a variety of control algorithms in both civilian and military applications. [1-3].

A brief about work done by several other researchers have been provided here. Fang et al. (2014) used LQR controller optimized with PSO algorithm for TWSBR [4]. Bin et al. (2010) described the kinematics architecture of a TWSBR [5]. Sadeghian et al. (2016) presented a dynamic structure as well as control approaches for a self-balancing robot, which is estimated by the help of Newtonian method [6]. Azimi et al. (2013) demonstrated a model predictive control (MPC) for a TWSBR, which is also called as a three-degree-of-freedom movement robot along with multivariable dynamics [7]. P. frankovský et al. (2017) presented the prototype of a self-balancing system operated with the help of geared DC motors [8].

In this work a TWSBR was designed, modeled and simulated with the help of MATAB Simulink. Fuzzy PD (FzPD) controller was used to stabilize the TWSBR by keeping vertical angle to zero. Further, proposed FzPD controller was also analyzed with PID and fractional order PID (FOPID) controllers for performance comparison. The parameters of all the controllers employed in this investigation were likewise tuned using the Water cycle algorithm.

Further, the remaining part of the work is systemized as illustrated: Module II extends modeling of TWSBR and also the transfer function of the system. Module III describes the FzPD controller used in this work in detail. Module IV gives the brief description of the optimization algorithm. Module V lists out the simulation results and Module VI winds up the paper with possible direction of future work.

II. SYSTEM MODELING

This section describes the steps taken into modeling of the plant under observation.

A. Mathematical Description of the Plant

A TWSBR is similar to an inverted pendulum in terms of design. In the absence of a balancing force, the inverted pendulum will fall. As a result, it is possible to conclude that the suggested system is unbalanced. The position of the pendulum, on the other hand, is normal or not inverted, indicating the robust nature.

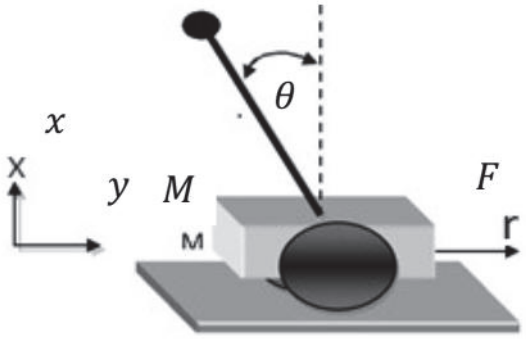


Fig. 1. Inverted pendulum's model.

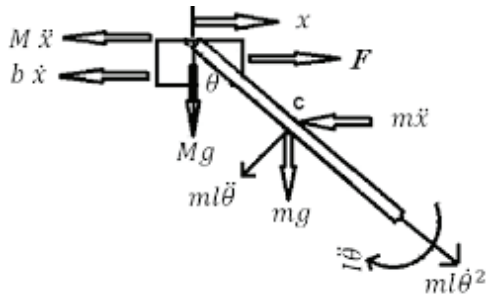


Fig. 2. Free body representation of the Inverted pendulum.

978-1-7281-8402-9/21/$31.00 ©2021 IEEE

2021 Asian Conference on Innovation in Technology (ASIANCON) | 978-1-7281-8402-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ASIANCON51346.2021.9544711

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:44:00 UTC from IEEE Xplore. Restrictions apply.

Fig 1 describes the structure of an inverted pendulum along with tilt angle θ. F signifies the force injected to the designed system. M symbolizes the mass of the model, x and y are the vertical as well as horizontal axis, correspondingly. During the practical execution, DC motors will afford the proper velocity, which helps the robot to remain stable. Deprived of the proper velocity, the proposed robot will be dropped.

The Newton II law is used to implement the dynamic equation for TWSBR. Fig 2 displays the Free Body Representation of the proposed system. The pendulum's slope has various angles with the two components along the horizontal as well as vertical directions. The mathematical approach for modeling self-balancing robots is seen in Fig 2 and is described as follows [9].

$$\sum f_x = 0 \#(1)$$

$$F - b\dot{x} - M\dot{x} - m\ddot{x} - ml\ddot{\theta}cos\theta + ml\dot{\theta}^2 sin\theta = 0$$

$$\sum f_y = 0 \ \#(2)$$

$$Mg + mg - ml\ddot{\theta}sin\theta + ml\dot{\theta}^2 cos\theta = 0$$

$$\sum M_A = 0 \ \#(3)$$

$$Mglsin\theta + I\ddot{\theta} - ml^2\ddot{\theta} + ml\ddot{x}cos\theta = 0$$

By combining all the above equations, the equation for the force is derived in both horizontal as well as vertical directions.

$$(I + ml^2)\theta + mglsin\theta = -ml\ddot{x}cos\theta \quad (4)$$

$$(M + m)\ddot{x} + b\dot{x} + ml\ddot{\theta}cos\theta - ml\dot{\theta}^2 sin\theta = F \quad (5)$$

Eq (4) and Eq (5) represent the linearized equations of the transfer function, where $q = \pi$. With considerations $\theta = \pi + \emptyset$,

$$cos\theta = -1 \quad (6)$$

$$sin\theta = -\emptyset \quad (7)$$

$$\frac{d^2}{dt^2} = 0 \quad (8)$$

After executing the above equations to non-linear, we find two discrepancies of motion equations. The U signifies the system's input.

$$(I + ml)^2\ddot{\emptyset} - mgl\emptyset = ml\ddot{x} \quad (9)$$

$$(M + m)\ddot{x} + b\dot{x} + ml\ddot{\emptyset} = U \quad (10)$$

B. Transfer Function of the TWSBR

On implementing Laplace transform on eq (9) and eq (10),

$$(I + ml^2)\ddot{\emptyset}(s)s^2 - mgl\emptyset(s) = mlX(s)s^2 \quad (11)$$

$$(M + m)X(s)s^2 + bX(s)s + ml\emptyset(s)s^2 = U(s) \quad (12)$$

TABLE I. PHYSICAL PARAMETER VALUES OF TWSBR

Variable	*Table column subhead*	*Subhead*
M	Cart's mass	0.3 Kg
m	Pendulum rod's mass	0.2 Kg
b	Viscous friction Co-efficient	0.1 N/m/sec
l	Pendulum's one half rod length	0.15m
I	Moment of inertia for the pendulum rod	0.0015 Kg.m^2
F	Applied force to the cart	$Kg.m/s^2$
g	Gravity	m/s^2
θ	Pendulum angle	rad

We generated the transfer function of TWSBR based on the equations (11) and (12) and by using the values given by Table I.

$$\frac{\emptyset(s)}{U(s)} = \frac{\frac{ml}{q}s}{s^3 + \frac{b(I + ml^2)}{q}s^2 - \frac{(M + m)mgl}{q}S - \frac{bmgl}{q}} \quad (13)$$

Where $q = [(M + m)(I + ml^2) - (ml)^2]$. The constraints given in Table I are replaced in eq (13), to get the below transfer function of TWSBR.

$$\frac{\emptyset(s)}{U(s)} = \frac{14.286s}{s^3 + 0.286s^2 - 70s - 0.0294} \quad (14)$$

III. FZPD CONTROLLER

In this work FzPD controller was used to stabilize the TWSBR. This section describes structure of FzPD controller in detail.

Fuzzy logic control mostly relies on the rules designed with the help of Linguistic variables. The control methodologies of fuzzy logic do not necessitate compound analytical computations, such as alternative approaches. It merely uses modest arithmetical computations to regulate the process model. Regardless of depending on fundamental arithmetic investigation, it performs well to regulate the control system. Henceforward, this technique is the utmost approaches accessible as well as also simpler one to regulate a plant. According to the fuzzy set concept, every component has a degree of membership that is associated to a specific set. Fuzzy based Controllers are mostly employed when accuracy needed is modest. Also, the process plant must be empty to accomplish the sophisticated mathematical investigation [10-12]. Fig 3 represents the block schema of fuzzy logic based PD controller. Fig. 4 represents the FzPD controller graphically.

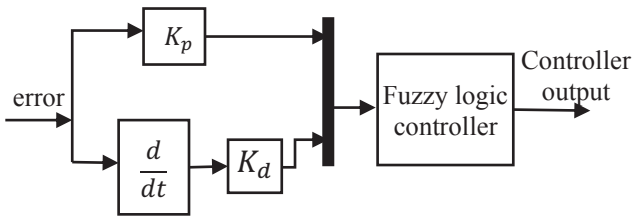


Fig. 3. Block design of the FzPD controller.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:44:00 UTC from IEEE Xplore. Restrictions apply.

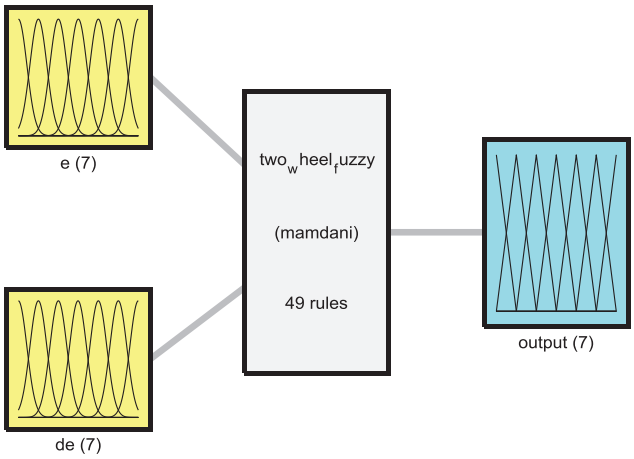


Fig. 4. Inside View of FzPD controller.

The Fuzzy Logic controller's three foremost elements are

- Fuzzification
- Fuzzy Rule base and Interfacing engine
- Defuzzification

A. *Fuzzification*

During fuzzification the numerical value of the input variable is converted into the fuzzy linguistic variable in which it is represented as the member of different membership function with a certain degree. The variable in Fuzzy-PD controller which are going to be fuzzified are error and derivative of the error or the area of the error. The MFs are the pictorial representation of the belonging of the input variable. Generally used MFs are trapezoidal, triangular, gaussian or most famous bell shaped. In this paper gaussian membership functions has been used for the input error and derivative of error.

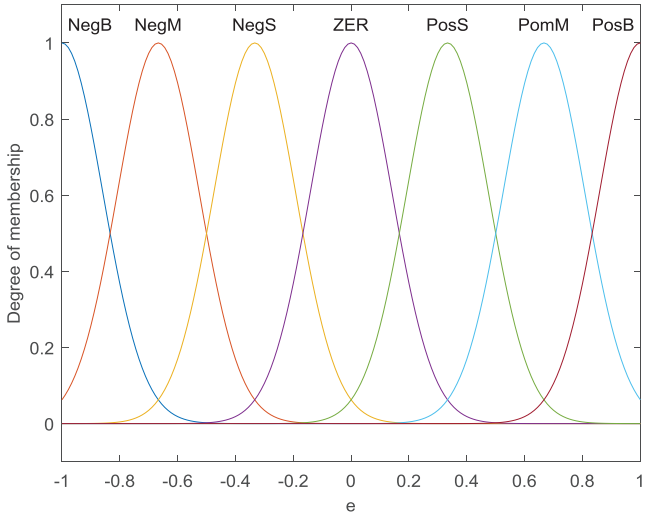


Fig. 5. Membership function of 1st input error.

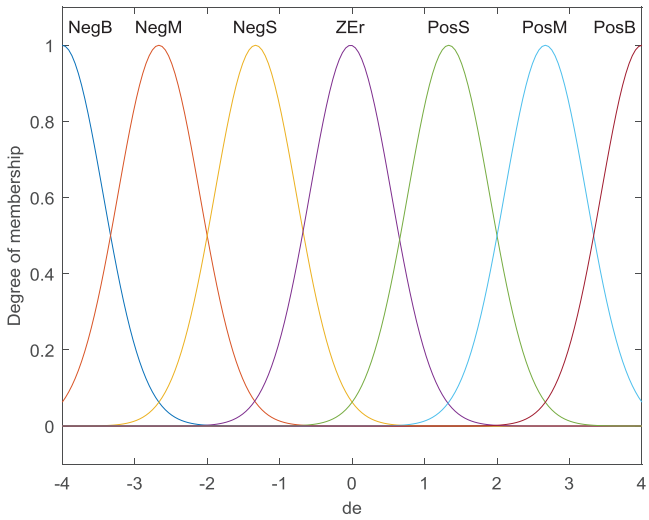


Fig. 6. Membership function of 2nd input change in error.

B. *Fuzzy Inference or Rule Base*

Rule base for proposed inference mechanism is shown by Table II. Total 49 rules were used in this thesis work. Figure 3-7 show the rule surface for the proposed FzPD system. Fig. 7 shows the surface view of rules used in the FzPD system.

TABLE II. RULE BASE

Sr. No	*e*	*de*	*Output*
1	NEGB	NEGB	NEGB
2		NEGM	NEGB
3		NEGS	NEGB
4		ZER	NEGB
5		POSS	NEGM
6		POSM	NEGS
7		POSB	ZER
8	NEGM	NEGB	NEGB
9		NEGM	NEGB
10		NEGS	NEGB
11		ZER	NEGM
12		POSS	NEGS
13		POSM	ZER
14		POSB	POSS
15	NEGS	NEGB	NEGB
16		NEGM	NEGB
17		NEGS	NEGM
18		ZER	NEGS
19		POSS	ZER
20		POSM	POSS
21		POSB	POSM
22	ZER	NEGB	NEGB
23		NEGM	NEGM
24		NEGS	NEGS
25		ZER	ZER
26		POSS	POSS
27		POSM	POSM
28		POSB	POSB
29	POSS	NEGB	NEGM
30		NEGM	NEGS
31		NEGS	ZER
32		ZER	POSS
33		POSS	POSM
34		POSM	POSB
35		POSB	POSB
36	POSM	NEGB	NEGS
37		NEGM	ZER
38		NEGS	POSS
39		ZER	POSM

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:44:00 UTC from IEEE Xplore. Restrictions apply.

40		POSS	POSB
41		POSM	POSB
42		POSB	POSB
43	POSB	NEGB	ZER
44		NEGM	POSS
45		NEGS	POSM
46		ZER	POSB
47		POSS	POSB
48		POSM	POSB
49		POSB	POSB

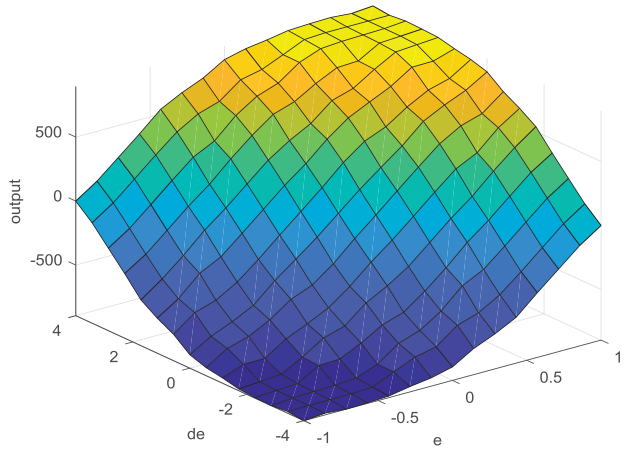


Fig. 7. Surface view of rules.

C. Defuzzification

Defuzzification is basically reversal of Fuzzification process. The Fuzzy Logic Controller generates a linguistic variable as the outcome. These variables must be translated to crisp output as described by legit fundamentals. There are several techniques available to do this task. In this work, centre of gravity approach has been used for the Defuzzification purpose. It obtains the centre of gravity of a section associated with fuzzy set. Fig. 8 shows the membership functions used in the output for defuzzification.

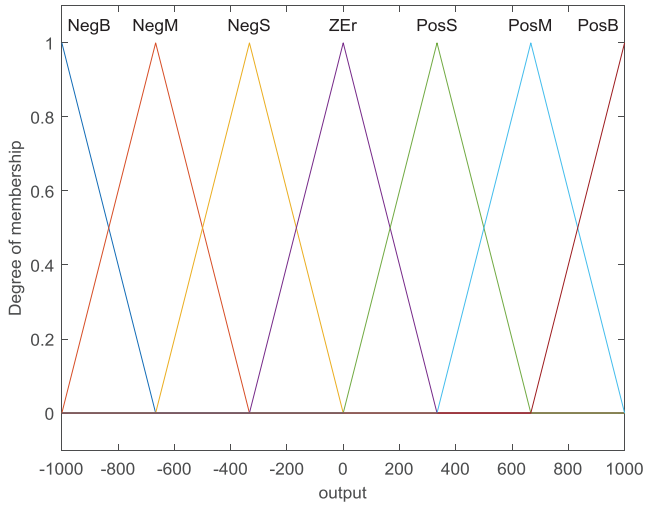


Fig. 8. Output MFs.

IV. Controller Parameters' Tuning

Controller output depends upon the gain parameter value. There are several methods by which these parameters can be tuned. In this work WCA optimization technique is used for tuning the controller's parameters. This section describes the parameter tuning method in detail.

A. WCA Optimization

The WCA imitators the movement of rivers as well as streams in the direction of the sea and observed by water cycle procedure. An initial streams' population is arbitrarily formed succeeding the raining procedure. The best stream with least objective function, is selected as the sea [13]. Many reliable streams are selected as rivers, in the mean while the others move to the rivers and sea. A stream of an array of $1 \times N$ is well-interpreted as follows:

$$steam\ Candidate = [x_i]_{1\times N}, \qquad \forall \in N \qquad (15)$$

In the above equation N signifies the design control variable quantity; and x_i is considered as the values of decision variable.

To begin the optimization process, an initial population having streams of size $N_{pop} \times N$ is produced. N_{pop} is the over-all number of populations. The matrix having initial population, which is produced arbitrarily, is specified using Eq. (16).

$$Total\ Population = \left[x_{ij}\right]_{N_{pop}\times N} \quad \forall \in N_{pop}, \forall j \in N \qquad (16)$$

Every (x_j) can be signified as decimal numbers or real values as an undefined set for continuous as well as discrete issues, correspondingly. The stream's cost is found by the estimation of objective function (C) shown as follows:

First process, N_{pop} steams are formed. A quantity of $N_{sr} = N_r + 1$; N_r signifies the quantity of rivers along with one sea; out of the finest, individuals with least numbers are collected as the sea and rivers. The sea is supposed to be the stream with the lowest value among the others. The streams that go into rivers or right into the sea are computed using (17).

$$N_{streams} = N_{pop} - N_{sr} \qquad (17)$$

Eq. (18) represents the number of streams which move to the rivers or sea. Definitely, (17) is portion of (15).

$$population\ of\ stream = \left[x_{ij}\right]_{N_{stream}\times N} \quad \forall \in N_{stream}, \forall j \in N \qquad (18)$$

The intensity of selected streams for every river as well as sea are estimated using:

$$C_n = Cost_n - Cost_{Nsr+1}, \forall n \in N_{sr} \qquad (19)$$

$$NS_n = round\left\{\left|\frac{C_n}{\sum_{i=1}^{N_{sr}} C_i}\right| \times N_{stream}\right\} \quad \forall n \in N_{sr} \qquad (20)$$

In the above equation NS_n signifies the quantity of streams that move to the particular rivers as well as sea.

Let's consider there are N_{pop} streams from which $N_{sr} - 1$ are picked-out as rivers as well as one is chosen as the sea. The displace X within the stream as well as the river may be arbitrarily updated employing the formula represented in Eq. (21)

$$X \in (0,\ C \times d), \qquad C > 1 \qquad (21)$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:44:00 UTC from IEEE Xplore. Restrictions apply.

In the above equation$1 < C < 2$; d represents the present distance within river and stream.

X agrees to an arbitrary value varies within 0 to $(C \times d)$. By assuming $C > 1$ permits streams to move in dissimilar orientation inclined towards rivers and after that rivers into the sea. Consequently, as the exploitation stage, the new locations for streams as well as rivers are defined as revealed in Eq. (22) - (24).

$$\vec{X}_{stream}^{t+1} = \vec{X}_{stream}^{t} + r(\dots) \times C \times \left(\vec{X}_{river}^{t} - \vec{X}_{stream}^{t}\right) \quad (22)$$

$$X_{stream}^{t+1} = \vec{X}_{stream}^{t} + r(\dots) \times C \times \left(\vec{X}_{sea}^{t} - \vec{X}_{stream}^{t}\right) \quad (23)$$

$$\vec{X}_{river}^{t+1} = \vec{X}_{river}^{t} + r(\dots) \times C \times \left(\vec{X}_{sea}^{t} - \vec{X}_{river}^{t}\right) \quad (24)$$

In the above equation, $r(\dots)$ signifies the random number $\in [0, 1]$. Position of river and stream is switched if stream's fitness function is better than its' connecting river. A related interchange is accomplished within a river and the sea. Mainly, sea water is formed due to evaporation and then it evaporated as rivers or streams move to the sea. In this algorithm, evaporation procedure is accountable for the exploration stage in the algorithm. For that resolution, the following condition is employed for evaporation situation:

$$if \ \left\|X_{sea}^{i} - X_{river}^{i}\right\| < d_{max} \ or \ rand(\dots) < 0.1, \forall i \in (N_{sr} - 1) \quad (25)$$

The newly produced sub-population, the finest stream flows like a new river. Another streams flow in the direction of the new river.

$$X_{stream}^{new} = L_B + r(..) \times (H_B - L_B) \quad (26)$$

In the above equation L_B and H_B are lesser as well as higher bounds well-defined by the specified issues, correspondingly. In the similar way, the finest newly generated stream is assumed to be a river moving into the sea while the rest are considered to move into the rivers or straight to the sea. A great value for d_{max} avoids additional searches as well as insignificant values inspire the severity of search closer to the sea. Consequently, d_{max} regulates the intensity of search closer to the sea such as finest obtained solution. The value of d_{max}adaptively reduces by using Eq. (27).

$$d_{max}^{i+1} = d_{max}^{i} - \frac{d_{max}^{i}}{Max_{iter}} \quad (27)$$

B. Objective Function

Objective function is the function which needs to be minimized during optimization procedure. Integral time absolute error (ITAE) has been used as an objective function in this work. Mathematical description of the same is given by (28).

$$ITAE = \int_0^t t * |e| \, dt \, \#(28)$$

V. Results & Discussion

This section presents the results obtained in this paper work. Fig. 9 shows the Simulink model of the FzPD controller used in this work. Transfer function output is the angle of TWSBR from vertical and as per the requirements, this needs to be zero. Hence, reference point for the system is set to zero. Further, disturbance of 0.1radian was added to the robot vertical angle as illustrated in Fig. 9

A. FzPD Output

Fig. 10 shows the convergence graph of WCA optimization algorithm for tuning the parameters of the FzPD controller. Optimization was run for 30 iterations with 50 population size. Lower bound and upper bound for the variables to be optimized were 0 and 100 respectively. Fig. 11 depicts the FzPD controller's response for disturbance rejection. It is clear from Fig. 11 that FzPD controller successfully stabilized the robot within the reasonable amount of time.

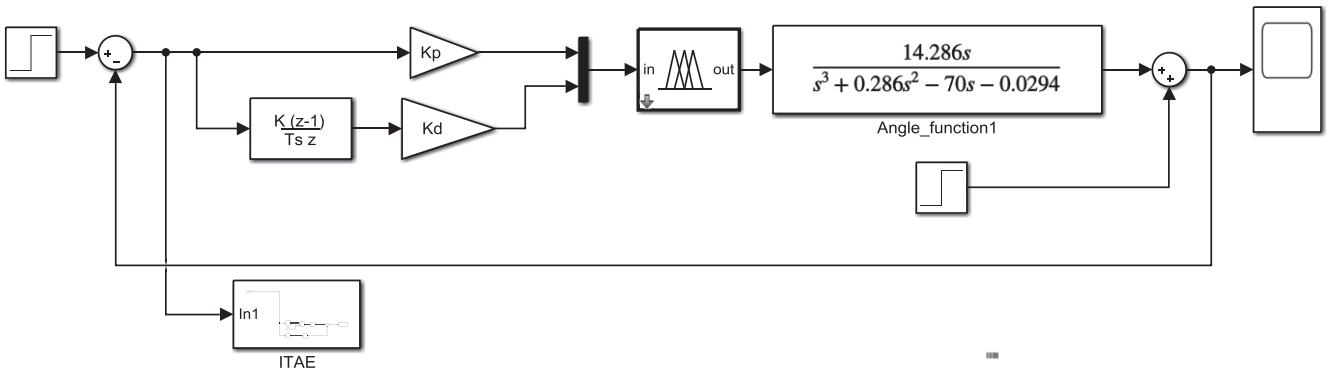


Fig. 9. Simulink model of FzPD controller.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:44:00 UTC from IEEE Xplore. Restrictions apply.

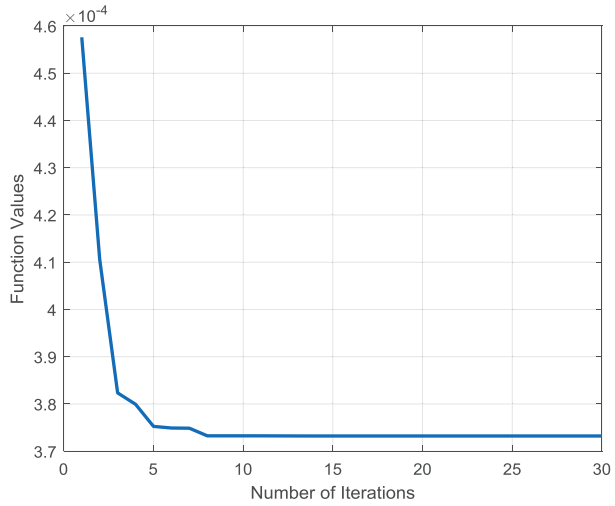


Fig. 10. WCA convergence curve for FzPD.

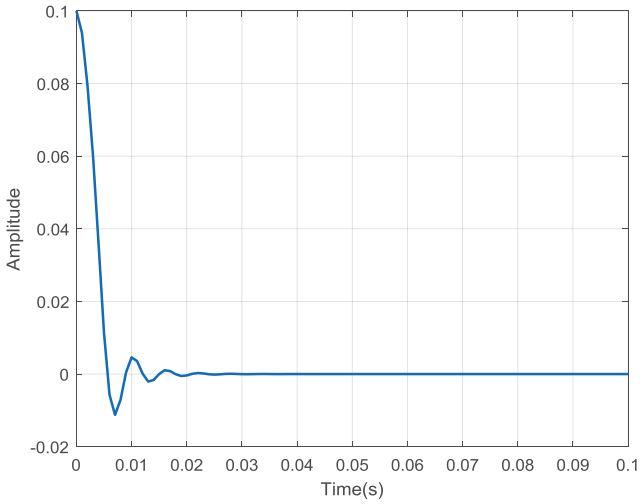


Fig. 11. System response with FzPD controller.

B. Comparative Analysis

In this work, FzPD controller output has been compared with two different controllers named as PID and fractional order PID (FOPID). Fig. 12 shows the comparative results obtained from this study. From figure it is clear that FzPD controller outperforms the other two controllers in terms of stabilizing the TWSBR.

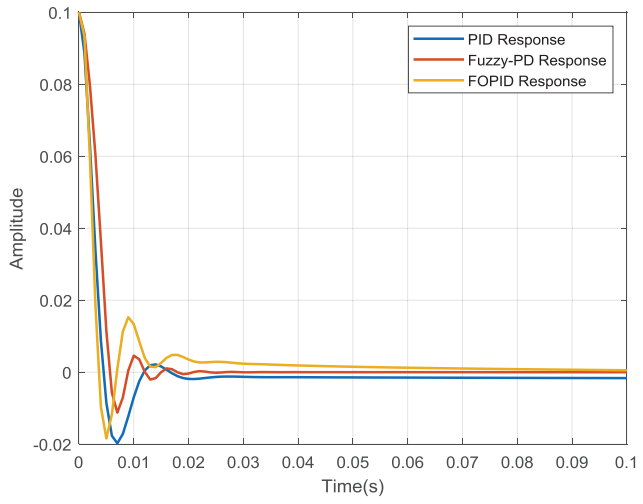


Fig. 12. Comparative analysis for different controllers.

TABLE III. OPTIMIZED CONTROLLER PARAMETERS

Parameter Name	PID	FOPID	FzPD
K_p	249.8	500	31.73
K_i	500	456	-
K_d	37.78	36	0.1718
N	500	-	-
λ	-	1.403	-
μ	-	1.08	-
ITAE	0.00047	0.00045	0.00037

VI. CONCLUSION

In this work, the TWSBR has been designed and simulated with the help of MATLAB 2018. Three different control algorithms, namely PID, FOPID and FzPD were used for controlling the system. WCA optimization algorithm was utilized for tuning the different parameters of the proposed controllers. Comparative study between these controllers shows that the FzPD controller outperform the other two controllers is stabilizing the self-balancing robot. In the future work, further study will be done to implement the proposed controllers on real hardware.

REFERENCES

[1] L. Sun & J. Gan "Researching of two-wheeled self-balancing robot base on LQR combined with PID", *2nd International Workshop on Intelligent Systems and Applications*, pp. 1-5, 2010.

[2] NGM Thao, DH Nghia, NH Phuc, "A PID backstepping controller for two-wheeled self-balancing robot", In *International Forum on Strategic Technology,* pp. 76-81, 2010.

[3] W. An, Y. Li "Simulation and control of a two-wheeled self-balancing robot", *International Conference on Robotics and Biomimetics (ROBIO)*, pp. 456-461, 2013.

[4] J. Fang, "The LQR controller design of two-wheeled self-balancing robot based on the particle swarm optimization algorithm", *Mathematical Problems in Engineering*, 2014.

[5] H. Bin, L. W. Zhen, L. Hai Feng, "The kinematics model of a two-wheeled self-balancing autonomous mobile robot and its simulation" *In 2010 Second International Conference on Computer Engineering and Applications*, vol. 2, pp. 64-68. 2010.

[6] R. Sadeghian & M. T. Masoule, "An experimental study on the PID and Fuzzy-PID controllers on a designed two-wheeled self-balancing autonomous robot", *In 2016 4th International Conference on Control, Instrumentation, and Automation (ICCIA)*, pp. 313-318, 2016.

[7] M. M. Azimi; H. R. Koofigar, "Model predictive control for a two wheeled self-balancing robot", *In 2013 First RSI/ISM International Conference on Robotics and Mechatronics (ICRoM)*, pp. 152-157, 2013.

[8] P. frankovský, L. dominik, A. gmiterko, I. virgala, P. kurylo, O. perminova, "Modeling of two-wheeled self-balancing robot driven by DC gearmotors", *International Journal of Applied Mechanics and Engineering*, vol.22, no.3, pp.739-747, 2017.

[9] N. Razmjooy, M. Ramezani, "Optimal control of two-wheeled self-balancing robot with interval uncertainties using Chebyshev inclusion method", *Majlesi Journal of Electrical Engineering*, vol.12, no.1, pp.13-21, 2018.

[10] C Qiu, Y Huang, "The design of fuzzy adaptive PID controller of two-wheeled self-balancing robot", *International Journal of Information and Electronics Engineering*, vol.*5, no.*3, pp.193,2015.

[11] R. Sadeghian, MT. Masoule, "An experimental study on the PID and Fuzzy-PID controllers on a designed two-wheeled self-balancing autonomous robot", In *2016 4th International Conference on Control, Instrumentation, and Automation (ICCIA)*, pp. 313-318, 2016.

[12] J. Wu & W. Zhang, "Design of fuzzy logic controller for two-wheeled self-balancing robot", *In Proceedings of 2011 6th international forum on strategic technology*, vol. 2, pp. 1266-1270, 2011.

[13] H Eskandar, A Sadollah, A Bahreininejad & M. Hamdi, "Water cycle algorithm–A novel metaheuristic optimization method for solving constrained engineering optimization problems", *Computers & Structures*, vol.*110*, pp.151-166, 2012.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on December 13,2021 at 05:44:00 UTC from IEEE Xplore. Restrictions apply.